UEOMA504

Traitement statistique de l'information

Vincent Pagé, vpage@univ-ag.fr

I Introduction

Le document que vous vous préparez à lire regroupe les notes des cours que je fournis à l'UAG, je le complète au fur et à mesure, il est donc pour le moment un petit peu aride et très dense en contenu pour un néophyte. J'espère néanmoins qu'il suit une démarche logique de présentation des différentes notions mises en oeuvre.

I.1 Présentation du cours

Au cours des 50 dernières années, la masse d'informations numérisée a connue une progression fulgurante. Ceci s'est accompagné d'un gros effort de recherche dans le sens d'une automatisation du traitement de ces informations. Pour donner quelques exemples d'applications de ces techniques, on pourra citer : La détection de spams dans les emails, la classification automatique de régions agricoles à partir d'images satellites, le diagnostic automatique de certaines maladies à partir de questionnaires et de mesures physiologiques, la reconnaissance de caractères dans des images (OCR), mais aussi la reconnaissance vocale, où la prise de décision financières... Comme vous le voyez, les domaines d'applications ne manquent pas.

Avant de voir comment procéder pour traiter ces informations, il va nous falloir un vocabulaire adapté. Pour la prise en main de ce vocabulaire, le plus simple est de détailler le titre de ce cours : « Traitement Statistique des Informations ». L'analyse de ce titre va ainsi nous occuper pour les trois prochaines sections.

I.1.a) Informations :

De façon très générale, on parle d'informations lorsqu'on dispose d'observations quantitatives ou qualitatives concernant un phénomène (physique, financier, social) qui nous intéresse. Dans ce qui suit, ces observations seront notées des **caractéristiques**.

Exemple : Plaçons nous dans la position d'une banque qui s'intéresse à la solvabilité des entreprises et considérons que l'on dispose de caractéristiques sur les entreprises qui viennent nous consulter telles que leur chiffre d'affaire (CA).

I.1.b) Traitement :

Il s'agit, sur la base de ces caractéristiques, de prendre une **décision** parmi n décision possibles.

Exemple : Lorsqu'une entreprise se présente, nous pouvons faire trois choix : Ne rien lui prêter, lui prêter 50 000 euros ou encore lui prêter 200 000 euros.

Pour prendre notre décision, nous allons commencer par relier chaque décision possible à un état de nature lié au phénomène qui nous intéresse. Cet état de nature doit représenter la décision idéale à prendre. Nous allons également numéroter nos états de nature, et les noterons w_1, w_2, ... w_n,

Exemple : Pour notre problème de prêt, nous pouvons considérer les états de nature suivants : Une entreprise peut être dans l'état w_1 « non solvable », w_2 « solvable pour 50 000 » ou encore w_3 « solvable pour 200 000 ». Les décisions sont alors reliées immédiatement au fait de deviner dans quel état de nature se trouve une entreprise en fonction de son CA.

De façon générale, nous appellerons ces états de natures des **classes**. Notre objectif est de construire un **classifieur** : une machine capable de choisir automatiquement à quelle **classe** affecter un **item** compte tenu de ses **caractéristiques**. Ce choix constituera notre **décision**. Notons que pour deux items présentant une même caractéristique nous prendrons la même décision.

Exemple : Dans l'exemple que nous avons pris, un item est une entreprise qui vient solliciter un prêt. Les caractéristiques faites sur cet item se résument au CA de cette entreprise. Un classifieur est un algorithme qui pour un CA donné, nous dit quelle est la solvabilité d'une entreprise ayant ce CA. Deux entreprises de même CA obtiendront la même réponse de notre part en termes de prêts.

I.1.c) Statistiques :

Les statistiques, ici, vont nous servir de cadre mathématique pour modéliser les caractéristiques, et caractériser notre méthode de prise de décision.

Pour cela, nous considérons qu'une caractéristique est sujette à un aléa dû à la façon d'observer et de choisir l'item. Plus simplement, nous considérerons que chaque caractéristique est le résultat du tirage aléatoire d'une Variable Aléatoire (**VA**) et relève à ce titre de la théorie des probabilité. Cette considération est hautement sujette à caution : Savoir si l'on peut considérer qu'une caractéristique est le résultat d'un tirage aléatoire est un problème qui déborde de loin le cadre de ce cours...

Exemple : Considérons le CA des entreprises (Françaises, Européennes....). On va considérer que le CA d'une

entreprise en particulier (qui vient nous consulter) est le tirage d'une VA. A nous de disposer d'informations pertinentes sur les statistiques du CA des entreprises Françaises ou Européennes...

I.1.d) Objectifs et plan du cours :

L'objectif de ce cours est donc d'utiliser des statistiques pour construire des classifieurs « efficaces » compte tenu de caractéristiques. Plus exactement, c'est pour quantifier l'efficacité des classifieurs qu'il va nous falloir faire des statistiques. Notons que les approches que nous verrons sont extrêmement générales et peuvent s'appliquer dans toutes sortes de cas. Dès qu'une décision doit être prise, on peut imaginer de mettre ces techniques en oeuvre.

Nous verrons en particulier dans ce cours qu'il est parfois possible de définir un classifieur dont on sait qu'il fera en moyenne moins d'erreurs que tous les autres classifieurs possibles. Ce classifieur est le *classifieur bayesien* et est au centre de la *Théorie de la Décision Bayesienne* qui fera l'objet du second chapitre. L'étude de différents cas nous montrera également que le recours aux statistiques peut s'avérer délicat, ce qui nous amènera à trouver des solutions empiriques (des heuristiques) pour réussir à prendre des décisions « sensées » malgré tout. Ceci fera l'objet du chapitre 3. Enfin, nous verrons que sous certaines conditions, nous pourrons améliorer un classifieur en l'entraînant sur des exemples de façon à le rendre plus efficace. Ceci fera l'objet du chapitre 4 sur le *Boosting*. Enfin, nous verrons qu'il est parfois nécessaire de prendre en compte le « contexte » environnant pour classifier les items. Pour mieux définir le problème, je vais changer d'exemple et considérer que je veux déterminer le sexe des individus situés autour d'une table pendant un mariage avec comme seule caractéristique leur taille. Les classifieurs que je pourrais construire en suivant le modèle exprimé précédemment s'intéresserait a chaque individu indépendamment de ses voisins. Dans le cas d'une table de mariage, ceci est dommage, car on sait que les individus sont en général placés en alternance homme/femme. Les façon de prendre en compte ce contexte (au sens « prendre en compte les voisins de l'item a classer ») seront développés dans le chapitre 5.

I.2 Rappels de statistiques

Avant d'aller plus loin, nous allons voir quelques rappels de statistiques qui nous serviront à justifier nos prises de décision dans un cadre mathématique clair, ainsi qu'à définir les notations utilisées.

I.2.a) Propriétés des VA Continues :

Pour simplifier, nos exemples de VA continues se limiteront à des VA à valeurs dans l'ensemble $\mathbb{R}$.

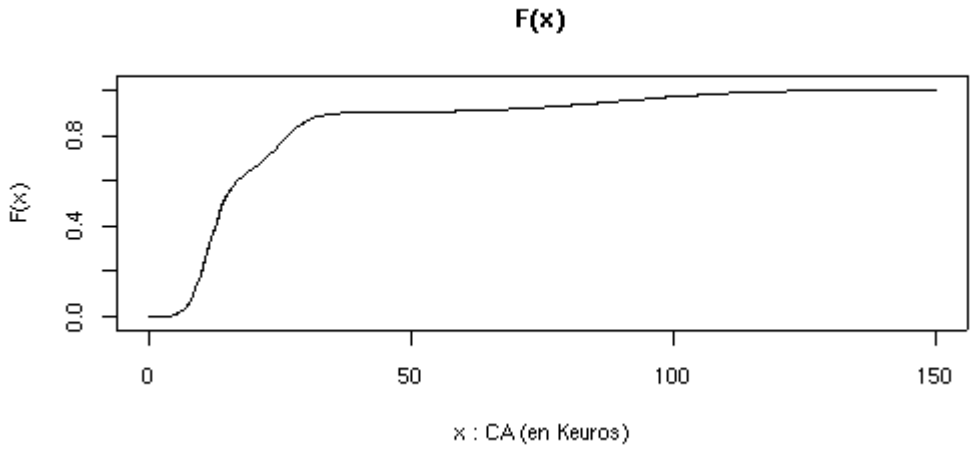
Soit X une VA continue.

On appelle **Fonction de Répartition** de X la fonction $\mathbf{F_X(x)}$ qui donne la probabilité qu'un tirage de X soit inférieur à x. Nous la noterons le plus souvent **F(x)**.

Propriété 1. $\forall x \in \mathbb{R}, F(x) \geq 0$

Propriété 2. $\forall (x, y) \in \mathbb{R}^{2,} on a : x \geq y \Rightarrow F(x) \geq F(y)$

Propriété 3. $\lim_{x \to -\infty} F(x) = 0$ **et** $\lim_{x \to +\infty} F(x) = 1$



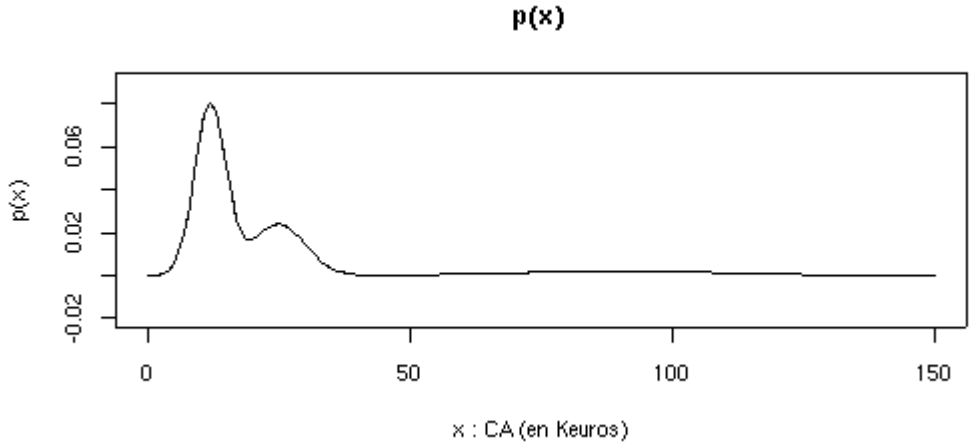
Propriété 4. Pour une VA X continue, la probabilité qu'un tirage de X prenne la valeur x est nulle
Démonstration :
La probabilité qu'un tirage de X prenne une valeur comprise entre x-e et x+e vaut :

$$p(\epsilon)=F(x+\epsilon)-F(x-\epsilon)$$

F étant continue, on a bien $F(0)=\lim_{\epsilon\to 0} p(\epsilon)=F(x)-F(x)=0$

Pourtant, on a l'intuition que certaines valeurs sont plus « probables » que d'autres. Dans l'exemple de la figure précédente lorsque F(x) augmente beaucoup, cela signifie que beaucoup d'entreprises ont un chiffre d'affaire de x !

C'est l'« augmentation » de F(x) qui nous donne cette notion de « probabilité intuitive » d'un événement. plus rigoureusement, il s'agit de la dérivée de F(x), que l'on appelle **Densité de Probabilité** de X. Nous la noterons dans ce qui suit p(X=x) ou encore **p(x)** ce qui est propre a faire bondir les plus mathématiciens d'entre nous mais constitue une notation fort pratique !



Propriété 5. $\forall x\in R, p(x)\geq$ •

Propriété 6. $\lim_{x\to\pm\infty} p(x)=$ •

Propriété 7. $\int_{\mathbb{R}} p(x)dx=1$

Une densité de probabilité définit complètement la VA qui lui est associée. Si on considère que nos caractéristiques sont des VA de densité de probabilité connue, il faudra s'assurer que nos densités de probabilité correspondent à la

réalité ! Quoiqu'il en soit, après ces quelques propriétés des VA continues, voyons ce qu'il en est des VA discrètes...

I.2.b) Propriétés des VA discrètes

Une VA Y est dite discrète si l'ensemble des valeurs qu'elle peut prendre est dénombrable. Le plus souvent, cet ensemble sera même fini. Y est alors dite à valeurs dans $\Omega = \{y_1, y_2, ... y_n\}$.

Si l'ensemble des valeurs est ordonné (ex : tirage d'un dé à 6 faces numérotées), on peut également parler de Fonction de Répartition. Celle-ci est alors « en escalier ». Dans le cas contraire (ex : tirage aléatoire du sexe d'un individu : « homme » ou « femme »), il n'y a pas de sens a parler de fonction de répartition (rigoureusement, la VA est alors qualifiée de VA qualitative nominale).

Dans les deux cas, on pourra en revanche s'intéresser à la **Probabilité de réalisation** d'un événement en particulier que l'on notera **P(X=x)** ou encore **P(x)**. Tout au long de ce cours, j'essayerais de conserver la notation p(x) pour les densités de probabilité (VA continues) et P(x) pour la probabilité de réalisation d'un événement (VA discrète). Si vous voyez une erreur, signalez le à l'auteur !

Pour les Probabilité de réalisation, on a l'équivalent des propriétés 5 et 7 :

Propriété 8. $\forall i \in \{1...n\}, P(y_i) \geq 0$

Propriété 9. $\sum_{i \in \{1...n\}} P(y_i) = 1$

Pour notre exemple, on s'intéressera à $P(w_i)$, probabilité qu'une entreprise soit « non solvable » (w_0) ,« solvable pour 50 000 euros » (w_1) ou encore « solvable pour 200 000 euros » (w_2).

I.2.c) Couples de Variables aléatoires

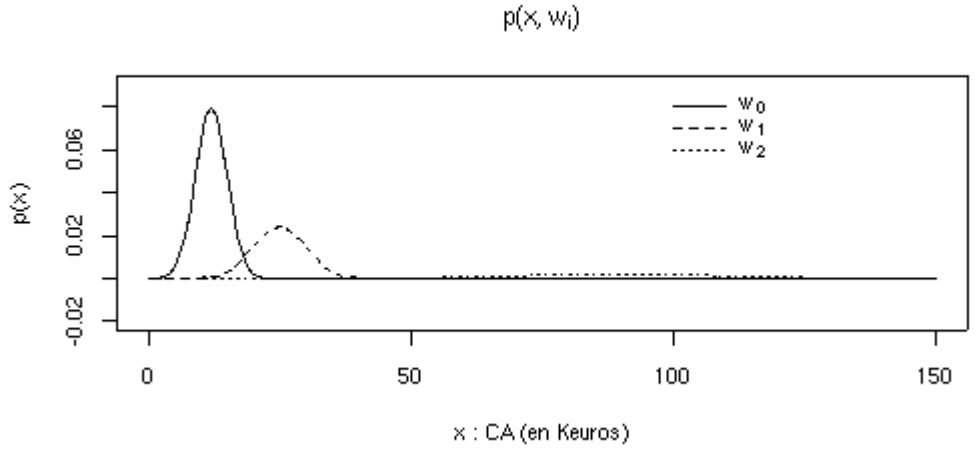
Nous en avons fini avec les VA isolées. Pour nos problèmes de classification, il va en revanche nous falloir considérer la probabilité d'une classe compte tenu des caractéristiques dont on dispose. (Pour notre problème de prêt bancaire, il va nous falloir considérer les probabilités qu'une entreprise soit solvable ou non, compte tenu de son CA). Ceci nous amène a considérer des **couples de VA**.

Pour simplifier, nous allons nous restreindre à des couples constitués d'une VA réelle et continue (pour les caractéristiques) et d'une VA discrètes (pour les classes) . Notons X la VA associée aux caractéristiques. Elle prends ses valeurs dans $\mathbb{R}$ ou un sous ensemble de $\mathbb{R}$. Notons W la VA associée à la classe, pouvant prendre ses valeurs dans $\{w_1, w_2, ... w_n\}$.

Pour notre exemple, X est le CA des entreprises (à valeurs dans $\mathbb{R}^+$). W est la solvabilité des entreprises. Les différents w_i sont « non solvable », « solvable pour 50 000 euros » et « solvable pour 200 000 euros ».

On appelle **densité de probabilité conjointe** de x et de w_i la quantité : $p(X = x \cap W = w_i)$ notée également $p(x, w_i)$

Pour notre exemple, $p(x, w_1)$ est la densité de probabilité qu'une entreprise nous consultant ait un chiffre d'affaires de x et que cette entreprise soit « non solvable ».



Notons que cette densité de probabilité conjointe observe également les propriétés 5 et 7 pour un événement constitué du couple (x, w_i) . La propriété 7 (intégrale égale à 1) nous donne alors la propriété suivante :

Propriété 10. $\int_{\mathbb{R}} [\sum_{i\in\{1...n\}} p(x, w_i)]dx = \sum_{i\in\{1...n\}} [\int_{\mathbb{R}} p(x, w_i)dx] = 1$

Notons pour être complet un cas extrême de densité de probabilité conjointe obtenu lorsque les VA sont indépendantes :

Propriété 11. Si X et W sont indépendantes, on a $\forall x, \forall i: \quad p(x, w_i) = p(x)P(w_i)$

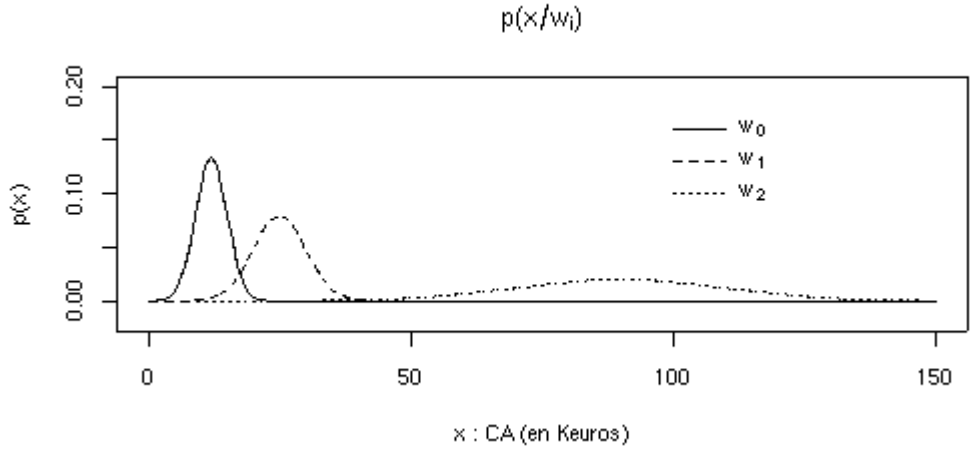
On peut obtenir les lois de X et W a partir de leur densité de probabilité conjointe à l'aide des formules suivantes. Les lois de X et de W déduites de leur densité de probabilités conjointes sont appelées lois marginales.

Propriété 12. lois marginales : $p(x) = \sum_{i\in\{1...n\}} p(x, w_i)$ **et** $P(w_i) = \int_{\mathbb{R}} p(x, w_i)dx$

On peut remarquer que la donnée de la densité de probabilité conjointe définit complètement nos VA (le couple mais aussi chacune des deux VA par l'intermédiaire des lois marginales). La plupart du temps, il va nous falloir calculer la densité de probabilité conjointe qui est rarement accessible directement. Pour cela, nous allons introduire la notion de probabilités conditionnelles.

On appelle **densité de probabilité conditionnelle de x sachant w_i** la quantité : $p(X = x/W = w_i)$ notée également $p(x/w_i)$. Ceci représente la densité de probabilité d'une caractéristique sachant à quelle classe appartient l'item auquel elle se rapporte. C'est le plus souvent cette densité qui est accessible facilement à la modélisation.

Pour notre exemple, $p(x/w_1)$ *est la densité de probabilité qu'une entreprise nous consultant ait un chiffre d'affaires de x sachant quelle est « non solvable ». Pour la modélisation, il nous faut disposer de tous les p(x/wi). Ainsi nous supposerons qu'une entreprise non solvable a un CA qui suit une loi normale de moyenne 10000 de variance 1000, qu'une entreprise « solvable a 50 000 euros » a un CA qui suit une loi normale de moyenne 30 000 de variance 5 000 et enfin qu'une entreprise « solvable a 200 000 euros » a un CA qui suit une une loi normale de moyenne 90 000 de variance 10 000.*

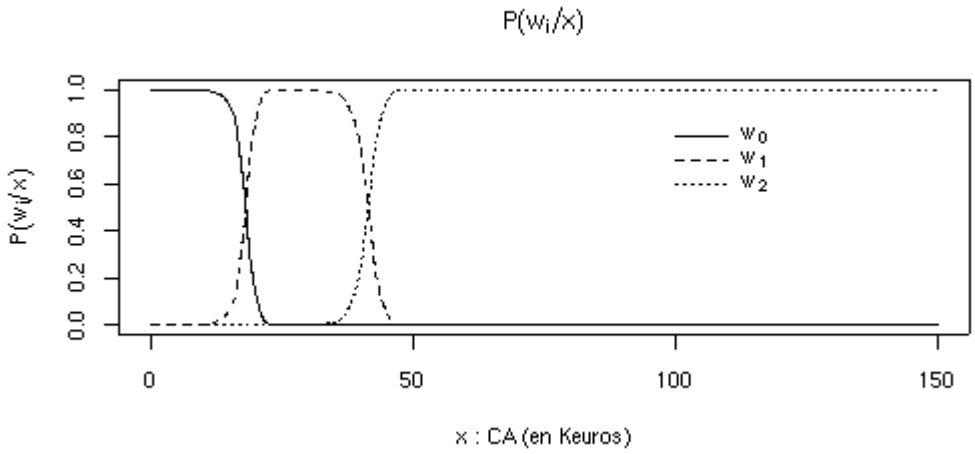


Propriété 13. Lois de Bayes : $\forall x, \forall i: \quad p(x/w_i) = \frac{p(x, w_i)}{P(w_i)}$

De fait, si nous connaissons la densité de probabilité conjointe de X et W, la propriété 12 nous indique comment calculer $p(w_i)$, donc $p(x/w_i)$.

On appelle de même **densité de probabilité conditionnelle de w_i sachant x** la quantité : $p(W=w_i/X=x)$ notée également $p(w_i/x)$. Ceci représente la densité de probabilité de la classe d'un item compte tenu de sa caractéristique.

Pour notre exemple, $p(w_1/x)$ est la densité de probabilité qu'une entreprise nous consultant soit « non solvable » sachant que son chiffre d'affaire est x.



Propriété 14. Lois de Bayes (le retour) : $\forall x, \forall i: \quad p(w_i/x)=\frac{p(x,w_i)}{p(x)}$

L'application de la loi de Bayes et de la propriétés sur les lois marginales nous permet de plus d'obtenir les propriétés suivantes :

Propriété 15. $\forall i \quad \int_{\mathbb{R}} p(x/w_i)dx=1$

Propriété 16. $\forall x \quad \sum_{i\in\{1...n\}} p(w_i/x)=1$

Pour notre exemple, la propriété 16 signale que pour une entreprise nous consultant dont on connaît le chiffre d'affaire x, $p(w_1/x)$ est la probabilité que cette entreprise soit « non solvable » compte tenu de ce CA.

Ceci conclue la section de rappels de statistiques que je vous conseille de bien maîtriser, la suite n'étant qu'une utilisation de ce formalisme. La section suivante est l'illustration de ce qui précède dans le cas de notre exemple.

I.2.d) Application pratique

Il s'agit ici de vous montrer dans un cas pratique comment on peut obtenir les différentes densité de probabilités présentées ci dessus. Reprenons l'exemple de la banque qui peut accorder ou non des prêts. Comme on l'a dit, le plus facile est souvent de commencer par $p(x/w_i)$.

Il semble relativement raisonnable de considérer que toutes les entreprises dans un état de solvabilité donné aient un CA qui soit réparti de façon « harmonieuse » autour d'une certaine moyenne. Nous pouvons donc faire l'hypothèse que les densités de probabilités conditionnelles $p(x/w_i)$ sont normales, de moyenne m_i , et d'écart type σ_i .

Nous pouvons alors écrire $p(x/w_i)=\frac{1}{\sqrt{2\pi\sigma_i^2}}e^{\frac{-(x-m_i)^2}{2\sigma_i^2}}$

Notre problème ne sera complètement déterminé que si nous connaissons $p(x,w_i)$. Pour cela, il nous faut également disposer des $P(w_i)$, qui représentent la probabilité de chaque classe.

Nous pouvons alors écrire $p(x, w_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\frac{-(x-m_i)^2}{2\sigma_i^2}} P(w_i)$

Si nous devons calculer $P(w_i/x)$, nous pouvons utiliser les lois de Bayes $P(w_i/x) = p(w_i, x)/p(x)$, mais nous ne connaissons pas encore p(x). Nous pouvons calculer celle ci à partir des propriétés sur les lois marginales.

$$p(x) = \sum_{i \in \{1...n\}} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\frac{-(x-m_i)^2}{2\sigma_i^2}} P(w_i)$$

On obtient ainsi $$P(w_i/x) = \frac{\frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\frac{-(x-m_i)^2}{2\sigma_i^2}} P(w_i)}{\sum_{j \in \{1...n\}} \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{\frac{-(x-m_j)^2}{2\sigma_j^2}} P(w_j)}$$

Cette expression est certes peu avenante, mais présente l'énorme avantage de pouvoir être calculée pour toute valeur de x au prix du calcul de n+1 exponentielles, ce qui nous intéressera plus tard lorsque nous nous intéresserons à la complexité algorithmique des classifieurs.

Pour l'exemple que nous avons pris, les différents paramètres ayant permis de construire les courbes présentées sont les suivants :

- m_1 = 12000 (euros) , σ_1 = 3000, $p(w_1)$ = 0.6
- m_2 = 25000 (euros) , σ_2 = 5000, $p(w_2)$ = 0.3
- m_3 = 90000 (euros) , σ_3 = 20000, $p(w_3)$ = 0.1

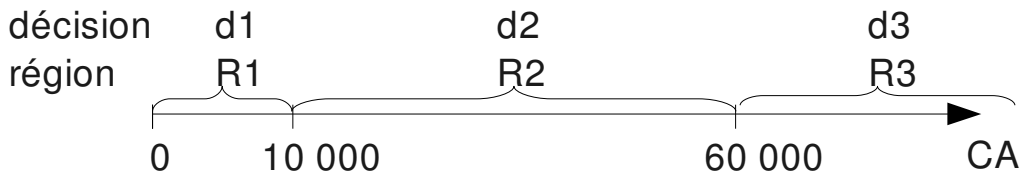
Retrouver ces différents paramètres à partir des courbes est un exercice assez intéressant.

I.3 Généralités sur les classifieurs

I.3.a) Régions de décisions

On l'a dit plus haut, un classifieur est un algorithme permettant pour chaque caractéristique possible x, de définir quelle décision w_i nous prendrons. Cette décision n'étant basée que sur la caractéristique et déterministe compte tenu de celle ci, cela revient a découper l'espace des valeurs possibles en **régions de décisions**. Chaque région correspond à une des n décisions possibles. Si la caractéristique d'un item tombe dans la région R_i , nous prendrons la décision d_i considérant que sa classe est w_i (à tort ou à raison)

Dans notre exemple nous pouvons décider arbitrairement que les entreprises ayant un CA inférieur à 10 000 euros sont « non solvables », que les entreprises dont le CA est compris entre 10 000 et 60 000 euros sont « solvables pour 50 000 euros » et enfin que les entreprises dont le CA est supérieur à 60 000 euros est « solvable pour 200 000 euros ». Dans ce cas nos trois régions sont les suivantes :



Le classifieur effectue une partition de l'espace des caractéristiques. Dans l'exemple ci-dessus, les régions sont connexes, ce qui n'est absolument pas une obligation.

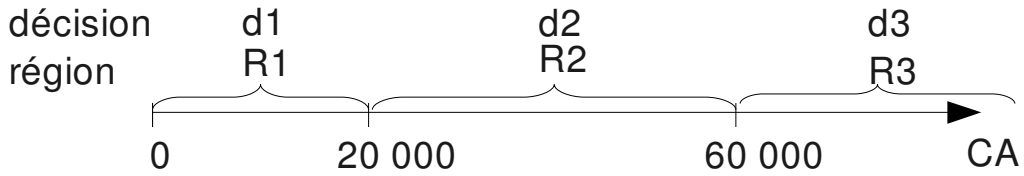
I.3.b) Fonctions discriminantes

La définition de régions étant parfois un peu contraignante, un autre formalisme pour la prise de décision consiste à définir ce que l'on nomme un ensemble de **fonctions discriminantes**. que l'on applique sur les caractéristiques. Il faut alors une fonction par décision possible $g_i(x), \quad i \in \{1...n\}$. Dans ce cas, la prise de décision est faite de la façon suivante : Lorsqu'on présente un item au classifieur, celui-ci calcule la valeur des différents $g_i(x)$. Le classifieur prendra la décision d_j si $\exists j \quad tq \quad g_j(x) > g_i(x), \quad \forall i \neq j$.

De fait, si on ajoute une règle supplémentaire pour gérer les cas d'égalités (ex: en cas d'égalité, le plus petit indice est systématiquement choisi), l'usage de fonctions discriminantes définit de façon indirecte des régions de décision puisque cela définit implicitement une partition de l'espace des caractéristiques. Les frontières entre deux régions R_i ,et R_j correspondent à des x tels que $g_i(x) = g_j(x)$

Par exemple : Ces fonctions seront ici basées sur une distance à un point de chaque classe. Si l'on prend CA_1 = 5 000,CA_2 = 35 000 et CA3 = 85 000, et que l'on se donne $g_i(x) = -|C_i - x|$, la prise de décision pour un item de caractéristique x consiste à calculer la distance de x à chacun des trois C_i puis à choisir la classe associée au plus proche de x.

Un rapide calcul montre que les régions de décisions sont alors



Dans la pratique, la plupart du temps on procède souvent à l'aide de fonctions discriminantes plutôt qu'en définissant explicitement les régions de décisions. Il nous faut donc trouver de « bonnes » fonctions discriminantes ce qui pose le problème de l'évaluation de la qualité d'un classifieur.

Par ailleurs, il est intéressant de constater la propriété suivante :

Propriété 17. Deux classifieurs ayant les fonctions discriminantes $g_i(x)$ et $g'_i(x)$ sont équivalents si l'ordre des $g_i(x)$ et des $g'_i(x)$ est le même pour tout x.

En particulier, cela se produit pour $g'_i(x) = F(g_i(x))$ avec F une fonction croissante.

Cette propriété nous permettra bien souvent de simplifier l'expression des fonctions discriminantes.

I.4 Évaluation de l'efficacité d'un classifieur

I.4.a) Probabilité d'erreur moyenne

Essayons de poser le problème de façon rigoureuse : la probabilité d'erreur moyenne sur l'ensemble des x possibles est obtenue par l'équation suivante :

Propriété 18. $P(erreur) = \int_{\mathbb{R}} p(erreur, x)\, dx$

où p(erreur,x) est la probabilité de : tirer une certaine caractéristique x ET de commettre une erreur. Le calcul de cette dernière grandeur n'est pas immédiat et il nous faut effectuer la démarche suivante :

Pour une caractéristique x donnée, notre classifieur prend une décision d_i en affectant à tous les items de caractéristique x la classe w_i. Le classifieur commet une erreur pour cette caractéristique lorsque des items des autres classes présentent cette caractéristique x. La probabilité qu'un item de caractéristique x donnée fasse partie de la classe w_i est par définition $P(w_j / x)$. On obtient ainsi la propriété suivante :

Propriété 19. Si nous avons pris la décision d_i pour une caractéristique x précise, la probabilité d'erreur pour ce x particulier s'écrit : $P(erreur / x) = \sum_{j \neq i} P(w_j / x)$

On peut alors calculer pour ce x $p(erreur, x) = P(erreur/x)\, p(x)$ (sous réserve que l'on puisse calculer p(x) ce qui est vrai si l'on dispose des densités de probabilités conjointes de X et de W).

Pour effectuer l'intégrale sur $\mathbb{R}$, le plus simple consiste à regrouper les x en régions. Celles-ci formant une partition de $\mathbb{R}$ on peut écrire $\int_{\mathbb{R}} p(erreur, x)\, dx = \sum_i [\int_{R_i} p(erreur, x)\, dx]$. A partir de ce résultat et des propriétés 18 et 19, le calcul de la probabilité d'erreur moyenne est immédiat et donne :

Propriété 20. La probabilité d'erreur moyenne s'écrit : $P(erreur) = \sum_i [\int_{R_i} \sum_{j \neq i} P(w_j/x)\, p(x)\, dx]$

En conclusion, pour tout classifieur défini par ses régions de décisions), si l'on connaît les densités de probabilités conjointes de X et de W, on peut calculer sa probabilité d'erreur moyenne. On peut donc imaginer de choisir le meilleur classifieur parmi tous ceux dont on dispose comme étant celui qui en moyenne, se trompe le moins souvent. Notons également qu'une application directe de la propriété 16 au sein de la propriété 20 nous donne la dernière propriété de cette section :

Propriété 21. $P(erreur) = \sum_i [\int_{R_i} (1 - P(w_i/x))\, p(x)\, dx]$

Remarque : Dans la pratique, ce calcul est parfois extrêmement délicat à mener lorsque l'intégrale ne peut pas se calculer de façon analytique. Il faut alors alors recourir à des méthodes numériques d'intégrations qui peuvent être très coûteuses en temps de calcul rendant cette estimation inapplicable (typiquement lorsque l'espace des caractéristique est trop grand).

Ceci, néanmoins, va nous permettre de construire un classifieur qui minimise la probabilité d'erreur moyenne, quelle qu'elle soit sans grandes difficultés. Le classifieur ainsi construit sera le classifieur bayesien que nous verrons dans le chapitre suivant.

II La théorie de la décision bayesienne

II.1 Le classifieur bayesien

II.1.a) Classification à erreur moyenne minimale

Notre objectif est donc de trouver un classifieur qui soit optimal du point de vue de la probabilité d'erreur moyenne. Le classifieur étant défini par ses régions, il faut choisir les « bonnes » régions de décision, « bonnes » étant à prendre au sens de « qui minimise la probabilité d'erreur moyenne ».

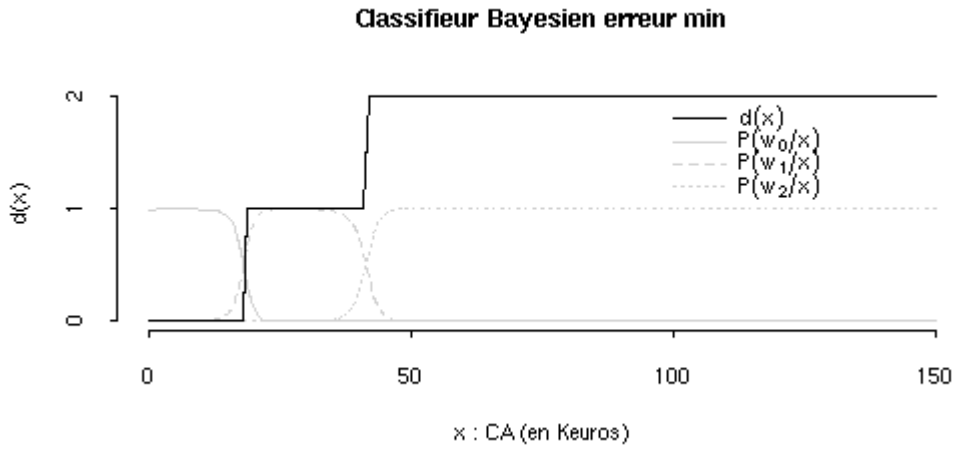
Si l'on reprend le résultat obtenu par la propriété 20, cela revient à choisir les régions de façon à ce qu'en chaque x, la décision d_i retenue minimise $(1 - P(w_i/x))\, p(x)$. Ceci est facilement obtenu en choisissant d_i tel que $P(w_i/x)$ soit le plus grand de nos n choix possibles.

Ce résultat est somme toute conforme à l'intuition : pour un x donné, on choisira la classe la plus probable compte tenu de ce x, ou plus formellement, celle pour laquelle $P(w_i/x)$ est le plus grand.

Notons de plus que ce classifieur nous offre spontanément ses fonctions discriminantes :

$$\forall i \in \{1...n\} \quad g_i(x) = P(w_i/x)$$

Pour l'exemple que nous avons pris, le classifieur bayésien obtenu est le suivant :



En utilisant la propriété 17, on montre aisément que les classifieurs bayesiens à erreur minimale admettent également les fonctions discriminantes suivantes :

- $g_i(x) = \log(P(w_i/x))$
- $g_i(x) = p(w_i, x)$
- $g_i(x) = \log(p(w_i, x))$

Cette dernière dernière expression nous permettra de simplifier l'expression des fonctions discriminantes dans le cas de lois appartenant à la famille exponentielle.

Pour résumer tout ce qui s'est dit jusqu'ici : Face à un problème dont nous connaissons complètement les lois de probabilités si l'objectif est de minimiser l'erreur moyenne, pour chaque caractéristique possible il suffit de choisir la classe la plus probable !

L'ensemble d'équations que nous avons pu utiliser ne nous aura servi qu'à formaliser ceci (et nous donne surtout les moyens de calculer ces probabilités).

Notons qu'il est cependant parfois moins important de faire peu d'erreurs que de faire des erreurs « graves ». La notion de « gravité » d'une erreur doit alors être définie extérieurement. ce que nous allons faire dans la section suivante.

II.1.b) Classification à risque moyen minimal

Il s'agit ici de considérer comme critère d'évaluation des classifieurs non plus leur probabilité d'erreur, mais un critère qui nous permettra de distinguer les différentes erreurs possibles, en leur donnant des poids différents. Ceci constitue ce que l'on nomme le **Risque Bayesien**. Posons donc les bases de ce qui va nous permettre de faire quelques calculs :

Pour chaque décision prise, nous allons considérer qu'il s'ensuit une **perte** (un gain étant une perte négative). De fait, cette perte dépend de la classe réelle de l'item classifié. On peut ainsi définir la perte associée à la décision d_i lorsque l'on sait connaît l'état de nature (w_j) de l'item présenté sous la forme $\lambda(d_i/w_j)$ que nous noterons λ_{ij} .

Pour notre exemple, nous pourrons prendre comme exemple de valeurs de pertes les valeurs suivantes : $\lambda(0/0)=0$ *(une entreprise non solvable à laquelle nous ne faisons pas de prêt ne nous coûte rien.* $\lambda(1/0)=50000$ *(une entreprise non solvable à laquelle nous prêtons 50000 euros va nous coûter a peu près 50000 euros !).* $\lambda(0/1)=3000$ *(une entreprise solvable pour 50 000 euros à laquelle nous ne prêtons rien va nous couter 3000 euros (les intérêts du prêt)). En continuant ainsi, nous pouvons définir tous les coûts qui figurent ci-dessous (ceux-ci constituent un exemple plus ou moins raisonnable, mais nous pouvons en fait choisir n'importe quoi) :*

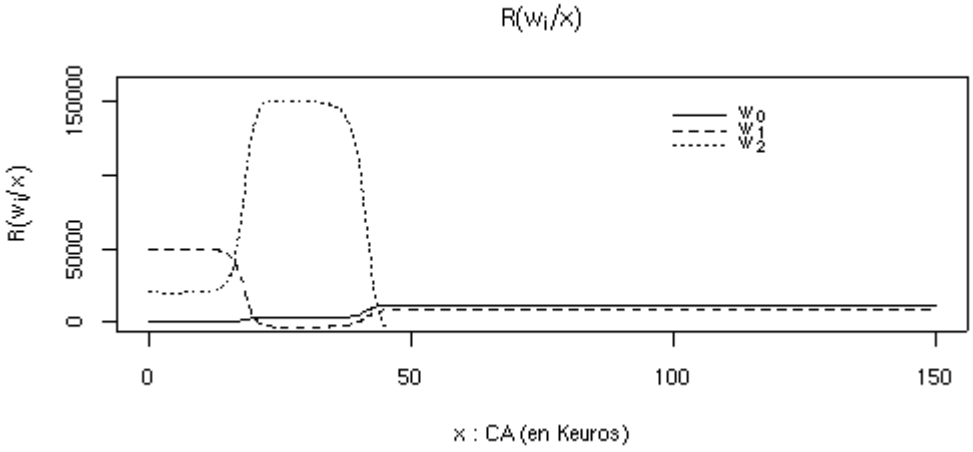
	w_j	w_j	w_j
d_0	*0*	*3000*	*12000*

	w_j	w_j	w_j
d_1	50000	-3000	9000
d_2	200000	15000	-12000

Il est alors facile de voir que

Propriété 22. Le risque moyen lorsqu'on prend une décision i pour un x donné s'écrit $R(d_i/x)=\sum_j \lambda_{ij} P(w_j/x)$

La quantité $R(d_i/x)$ est appelée le **Risque Conditionnel.**



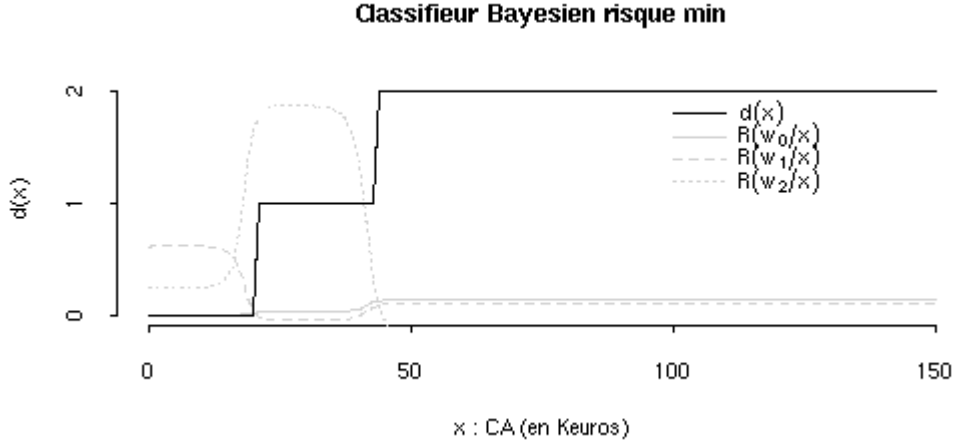
En regroupant les x dans les différentes régions de décisions de notre classifieur, on obtient facilement le **risque total moyen** : $R=\sum_i \int_{R_i} R(d_i/x)\,p(x)\,dx$. En utilisant l'expression de la propriété 21, on obtient la propriété suivante :

Propriété 23. le Risque Total moyen s'écrit $R=\sum_i \int_{R_i} \sum_j [\lambda_{ij} P(w_j/x)]\,p(x)\,dx$

Comme dans le cas de la probabilité d'erreur moyenne, on peut construire un classifieur qui minimise cette quantité. Il devra en chaque x prendre la décision qui minimise la quantité $\sum_j \lambda_{ij} P(w_j/x)$. On obtient ainsi le **classifieur bayesien à risque moyen minimal**. Aucun classifieur ne pourra avoir un risque total moyen inférieur à celui du classifieur bayesien ainsi construit. Notre règle de construction du classifieur nous donne instantanément ses fonctions discriminantes (notez le signe – dans la formule qui suit puisque la règle des classifieurs à fonctions discriminantes consiste a prendre la décision qui maximise $g_i(x)$:

$$\forall i \in \{1...n\} \quad g_i(x) = -\sum_j \lambda_{ij} P(w_i/x)$$

Pour l'exemple que nous avons pris, le classifieur bayésien obtenu est le suivant :



Notons enfin que le classifieur bayesien à erreur moyenne est équivalent à un classifieur bayesien à risque minimal particulier. On l'obtient pour une distributions des λ_{ij} telle que $\forall i \quad \lambda_{ii}=0$ et $\forall i \neq j \quad \lambda_{ij}=1$.

II.1.c) Approche de Neymann Pearson

Dans le cas très particulier d'un choix entre deux hypothèses (problème à deux classes) pour lequel les erreurs n'ont pas la même importance, on peut également appliquer l'approche de Neymann Pearson qui formellement n'est pas différente de la précédente mais fait partie de l'arsenal des techniques classiques de prise de décision.

II.2 Généralisations des classifieurs

II.2.a) Caractéristiques vectorielles

Jusqu'ici, les caractéristique que nous avons définies étaient scalaires et à valeurs dans $\mathbb{R}$. Ce modèle ne convient pas si l'on dispose de plusieurs caractéristique différentes sur le même item. Dans ce cas, il va nous falloir reprendre ce qui précède en procédant à quelques modifications mineures, qui font l'objet de cette section.

Imaginons que nous disposions de m caractéristique pour chaque item. Pour simplifier, chacune de ces caractéristique sera supposée réelle. Notons chacune de ses caractéristiques x_k, $k \in \{1...m\}$. Ces caractéristiques peuvent être regroupées dans un vecteur x de dimension m. x est maintenant un vecteur de caractéristique et $x \in \mathbb{R}^m$. Formellement, la plupart des équations que nous avons écrites précédemment restent valides, en prenant soin de transformer les intégrales sur $\mathbb{R}$ en intégrales sur $\mathbb{R}^m$. La définition des densités de probabilité, de probabilité d'erreur, de risque, mais aussi la méthode de construction des classifieurs bayésiens restent inchangées. Nous voilà donc capable de classifier automatiquement des items caractérisés par un nombre de caractéristiques quelconques !

Néanmoins, tous ces calculs nécessitent de connaître les $p(w_i, x)$ ce qui dans le cas vectoriel peut s'écrire $p(w_i, x_1, x_2, .., x_m)$

Notons d'ors et déjà que la définition de ces différentes probabilités sous forme analytique est extrêmement délicate. Ci dessous, nous présentons deux cas où l'on peut utiliser ces techniques :

- Si les différentes composantes d'une caractéristique sont indépendantes
- Si pour chaque classe, x suit une loi normale de dimension m.

Composantes indépendantes

Dans ce cas, on peut écrire immédiatement :

$$p(w_i, x) = \prod_{k \in \{1...m\}} p(w_i, x_k)$$

Le calcul des $p(w_i, x_k)$ et leur utilisation est alors identique à tout ce que nous avons vu auparavant. Notez qu'il arrive que l'on fasse cette hypothèse d'indépendance alors que l'on sait qu'elle est erronée. On la fait lorsque l'on ne peut pas faire autrement, ce qui donne alors un **classifieur bayesien naif** (idiot Bayes classifier).

Cas de la loi normale multidimensionnelle (multivariée)

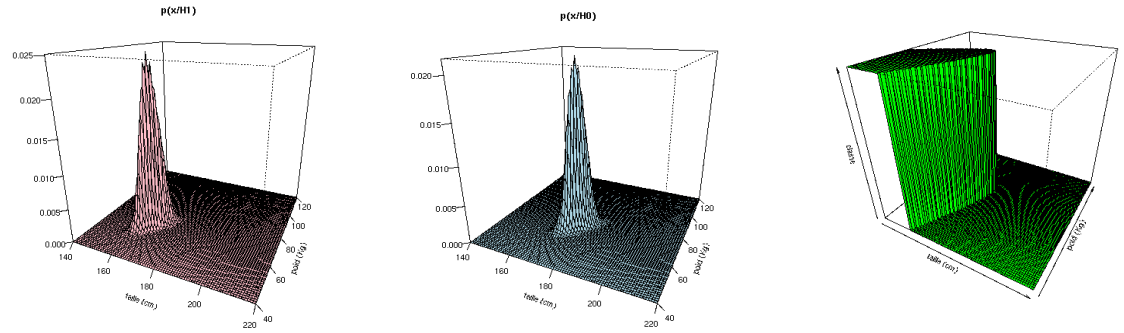
Dans le cas général de loi normale en m dimension, les différents paramètres ne sont pas indépendants.

Dans ce cas, les paramètres de la loi sont en nombre beaucoup plus grand : on compte pour chaque classe un **vecteur moyen** μ_i de dimension m mais aussi une **matrice de covariance** Σ_i de dimension m par m. Ceci représente au total n.m(m+1) paramètres à connaître ! Sous réserve que cela soit le cas, on peut alors écrire :

$$p(x/w_i)=\frac{1}{(2\pi)^{m/2}det(\Sigma)^{1/2}}e^{-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)}$$

Le lecteur intéressé par de plus amples informations sur les propriétés de la loi normale multivariée est prié de consulter les ouvrages de mathématiques s'y rapportant.

L'exemple qui suit présente des courbes $p(x/w_i)$ *ainsi que le classifieur bayesien associé. Dans cet exemple, on veut classifier des individus par sexe en fonction de leur taille et de leur poids. Ces deux caractéristiques sont liées et supposées ici obéir pour chaque sexe à une loi normale bivariée. La courbe de gauche représente la densité de probabilité d'une caractéristique (poid, taille) pour les femmes. Celle du milieu se rapporte aux hommes. Celle de droite illustre le comportement du classifieur minimisant l'erreur moyenne (Une valeur de 1 en z indique que le classifieur choisit la classe « femme », une valeur de 0 indique le choix de la classe « homme »). Pour la construction de ce classifieur nous avons besoin de connaitre les probabilités de chaque classe ici fixées à une proportion 0.6 d'hommes contre 0.4 de femmes.*



II.2.b) caractéristiques discrètes

Si certaines de nos caractéristiques sont maintenant discrètes, les VA associées n'ont pas à proprement parler de densité de probabilité. Leur caractérisation se fait au moyen de la probabilité que chaque VA prenne une valeur en particulier. Cependant, encore une fois, les différentes équations que nous avons pu voir restent valables sous réserve de remplacer les densités de probabilités par des probabilités la où cela est nécessaire, et en remplaçant les intégrales par des sommes discrètes dans les calculs incluant ces VA.

II.2.c) Utilisation du contexte :

Il est important de constater que toutes nos décisions considèrent que les items présentés à notre classifieur sont indépendants statistiquement. Prenons l'exemple précédent, de classification homme/femme. Nous prenons, pour chaque individu passant la porte de ce cours des caractéristiques (poid et taille) sur la base desquelles nous prenons une décisions, en considérant que l'ordre d'arrivée des individus est sans importance. Or il apparaît dans la pratique que les individus se déplacent souvent en groupes de même sexe. Il serait donc intéressant que nos classifieurs prennent en compte l'ordre d'arrivée pour améliorer nos classifications. Dans ce cas, il ne faut plus considérer une classification individu par individu, mais par groupe. Il nous faut disposer des probabilités non plus simplement des classes « homme »/ « femmes », mais les probabilités d'arrivées de chaque combinaison possible (disons « homme » puis « hommes » puis « femme »……). Si l'on dispose de ces probabilités (très rare dans la pratique), il est possible de mener les calculs de façon assez similaire a ce que nous avons fait. Comme précisé dans l'introduction, ceci sera évoqué dans le chapitre 5. Nous conseillons au lecteur placé face à ce type de cas de chercher dans la littérature des articles se rapportant à la « Compound Bayes Theory ».

II.2.d) conclusion sur les généralisations :

Dans ce qui précède, nous avons vu (parfois entre-aperçu) qu'il est possible de définir un classifieur optimisant un critère donné.

Il convient de noter qu'il est nécessaire de disposer des formes analytiques des densités de probabilités. Nous avons pris l'exemple de la loi normale pour plusieurs raisons : Cette loi se rencontre fréquemment dans les cas réalistes (disons dans 50% des cas pour fixer les idées) du fait de la loi des grands nombres mais aussi puisque les calculs s'en trouvent

simplifiés. De façon générale, le même calcul peut être fait de façon simple pour toutes les lois de la famille exponentielle (normale, poisson, gamma...). Ces lois présentent l'inconvénient majeur d'être mono-modales : elles présentent un unique pic centralisé sur leur moyenne. Ceci représente une limitation majeure pour certains cas pratiques (disons 30% également pour fixer les idées). Pour passer outre ce type de problèmes, nous verrons plus loin que l'on peut avoir recours à des mixtures de lois mais pouvons déjà noter que de façon très générale, la force de l'approche bayésienne est aussi sa faiblesse : il faut disposer de lois dont la densité de probabilité a une forme analytique connue.

Dans la pratique, ceci n'est pas toujours le cas. La suite de ce cours va donc s'intéresser aux cas ou cette densité de probabilité n'est pas complètement connue. Ceci va nous amener à distinguer deux cas :

- Lorsque nous connaissons la forme analytique des densités de probabilité mais pas ses paramètres (exemple : une loi gaussienne de moyenne et variances inconnue)
- Lorsque nous ne connaissons pas la forme analytique des densités de probabilité.

De fait, dans ce type de cas, nous pouvons bien souvent obtenir des exemples d'items déjà classifiés. Nous aimerions que notre classifieur puisse apprendre à partir de ces exemples les informations qui nous manquent. Dans ce cas, notre problème est appelé un problème d'**apprentissage supervisé**. On parle d'apprentissage supervisé dès lors que le classifieur sera construit sur la bases d'exemples dont on connaît les véritables classes. (des items déjà classifiés peuvent également servir a vérifier la qualité du classifieur pour une estimation de sa probabilité d'erreur mais dans ce cas il ne s'agit pas spécifiquement d'apprentissage supervisé).

Cet apprentissage supervisé va faire l'objet des sections suivantes.

II.3 Apprentissage supervisé et estimation de paramètres

Comme nous l'avons vu, il s'agit ici d'apprendre un certain nombre de choses à partir d'exemples donnés. Commençons par envisager le cas ou la forme des densités de probabilité est connue.

Si la forme des densité de probabilité est connue sous forme paramétrique, il arrive fréquemment que les paramètres de ces lois soient inconnus. Prenons encore une fois l'exemple de la loi normale. Celle-ci dépend de deux paramètres : moyenne et variance (ou moyenne et écart-type ou encore ses deux premiers moments). Nous nous plaçons ici dans le cas ou nous savons que les densité de probabilités sont gaussiennes mais par exemple, de moyenne connue et de variance inconnue. De fait, nous pouvons imaginer tout un ensemble de cas dans lesquels chaque paramètre est connu ou inconnu. Toutes les démarches précédentes sont techniquement inapplicables, puisqu'on ne peut pas calculer directement les p(x/wi). Ce sont les paramètres inconnus qui nous en empêchent. Ces paramètres sont, dans notre jargon des **paramètres de nuisance** : nous ne nous y intéressons pas pour eux mêmes, mais leur indétermination nous limite. On dispose alors, dans la théorie classique de différentes méthodes qui consistent a gérer ces paramètres. Deux de ces méthodes consistent a estimer ces paramètres, la troisième a intégrer ce paramètre sur l'ensemble de ses valeurs possibles, mais encore une fois, l'objectif est simple : faire disparaître ces paramètres en tant qu'inconnues de nos équations.

II.3.a) Estimation au sens du Maximum de Vraisemblance (MV)

Comme cela a été dit, nous disposons pour chaque classe, d'exemples de caractéristiques dont les vraies classes sont connues. Ces exemples vont, entre autre nous servir a estimer les paramètres de nuisance qui nous manquent. Pour procéder a cette estimation, posons notre notation de façon claire :

Nous disposons d'un ensemble d'exemples ξ . Notons c le nombre de classes de notre problème. La classe de chacun des exemples étant connue, l'ensemble ξ peut être partitionné en c sous-ensembles ξ_i représentatifs de la classe i. plus rigoureusement, les exemples de ξ_i ont été tirés indépendamment en respectant la loi $p(x/w_i)$. Cette loi est connue a un vecteur de paramètres θ_i prêt. Nous noterons θ l'ensemble des paramètres : $\theta=\{\theta_1,\theta_2,...\theta_c\}$.

Pour notre exemple, on peut maintenant supposer que les entreprises dans un état de solvabilité donné ont un CA qui suit une loi normale, de moyenne et de variance inconnue. Le vecteur de paramètre se résume pour chaque classe au couple moyenne/variance du CA. On a alors $\theta_i=[m_i,\sigma_i]$.

Pour simplifier, commençont par considérer sans trop de perte de généralité que les n échantillons de ξ_i

n'apportent aucune information sur les paramètres θ_j pour $i \neq j$. On peut alors s'intéresser à chacune des classes indépendamment, en ne considérant que les échantillons dont on dispose sur cette classe.

Pour résoudre ce problème, l'estimation au sens du maximum de vraisemblance consiste a choisir comme valeur des paramètres celle qui rend notre ensemble d'exemples le plus probable. Voyons donc comment procéder : Pour l'ensemble d'échantillons ξ_i , les paramètres inconnus sont contenus dans le vecteur θ_i . L'estimation au sens du maximum de vraisemblance va consister à choisir θ_i tel que $p(\xi_i/\theta_i)$ soit maximal.

Les exemples étant tirés indépendamment, et observés a travers leurs caractéristiques, on peut écrire

$$p(\xi_i/\theta_i)=\prod_{k=1}^{n} p(x_k/\theta_i)$$

Si l'on voit cette quantité comme une fonction de θ_i , cette grandeur est appelée vraisemblance de θ_i en fonction de l'ensemble d'exemples. Encore une fois, l'estimation de θ_i au sens du maximum de vraisemblance consiste a choisir comme valeur de θ_i celle qui maximise cette quantité.

Dès que l'on utilise un estimateur pour un paramètre, il est courant de le distinguer de la véritable valeur du paramètre. Dans ce cours, nous suivrons la notation suivante : l'estimateur de θ_i sera noté $\hat{\theta}_i$.

On peut donc écrire ce qui précède sous forme mathématique :

$$\hat{\theta}_i=argmax_{\theta_i}\{\prod_{k=1}^{n} p(x_k/\theta_i)\}$$

Il s'agit alors de déterminer ce maximum. Pour cela, on peut remarquer qu'en un point maximum, la dérivée de la fonction s'annule. On peut donc être sur que la dérivée de la vraisemblance s'annule au point cherché ce qui va limiter la recherche. Par ailleurs, le maximum d'une fonction reste inchangé si l'on applique une transformation strictement croissante a cette fonction. Pour toutes les lois de la famille, on s'intéressera ainsi au logarithme de la vraisemblance, qui est appelé la **log-vraisemblance du paramètre**.

Pour notre exemple, intéressons nous aux sociétés non solvables. On dispose de n exemples de ces sociétés, et l'on sait que le CA de ces sociétés est une gaussienne de moyenne et variance inconnue. On recherche $\hat{\theta}_i=argmax_{\theta_i}\{\prod_{k=1}^{n} p(x_k/\theta_i)\}$ *. Pour simplifier la notation, notons* $J(\theta_i)=\prod_{k=1}^{n} p(x_k/\theta_i)$ *. En considérant la log-vraisemblance, on a aisément :* $\log(J(\theta_i))=logJ(\theta_i)=\sum_{k=1}^{n} \log(p(x_k/\theta_i))$.

Les lois étant gaussiennes, on peut écrire : $logJ(\theta_i)=\sum_{k=1}^{n} \log(\frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\frac{-(x_k-m_i)^2}{2\sigma_i^2}})$ *qui se simplifie en*

$$logJ(\theta_i)=\sum_{k=1}^{n} [\log(\frac{1}{\sqrt{2\pi\sigma_i^2}})+\frac{-(x_k-m_i)^2}{2\sigma_i^2}]$$ *ou encore*

$$logJ(\theta_i)=n\log(\frac{1}{\sqrt{2\pi\sigma_i^2}})-\frac{1}{2\sigma_i^2}\sum_{k=1}^{n}[(x_k-m_i)^2]$$

Cherchons à estimer la moyenne : on dérive l'expression ci-dessus par rapport à m_i *:*

$$\frac{\partial logJ(\theta_i)}{\partial m_i}=\frac{1}{2\sigma_i^2}\sum_{k=1}^{n} 2(x_k-m_i)$$

Cette quantité s'annule pour une seule valeur qui est notre estimateur :

$$\hat{m}_i = \frac{1}{n}\sum_{k=1}^{n} x_k$$

Ce résultat est très satisfaisant : Dans le cas d'une loi gaussienne, l'estimateur au sens du maximum de vraisemblance est simplement la moyenne des échantillons. Ce raisonnement peut être refait pour toutes les lois de la famille exponentielle. Les mêmes calculs peuvent être également fait pour la variance d'une gaussienne. Je recommande au lecteur de mener ces calculs. On obtient alors l'estimateur de variance suivant :

$$\hat{\sigma}_i = \frac{1}{n}\sum_{k=1}^{n} (x_k - \hat{m}_i)^2$$

Ceci, certes un peu fastidieux, nous donne une justification théorique à l'utilisation des estimateurs classiques (a dire vrai, c'est une des raisons pour lesquelles ces estimateurs sont si classiques !)

L'estimateur du maximum de vraisemblance nous permet ensuite de nous ramener aux cas de lois connues évoqué précédemment. De fait, n'importe quel estimateur permet de s'y ramener. Les performances du classifieur ainsi construit seront alors liées à deux choses indépendantes : difficulté intrinsèque du problème (si tout était connu) et qualité de l'estimation des paramètres. En fonction des cas pratiques, il peut être intéressant de choisir empiriquement d'autres estimateurs (pour estimer la moyenne, par exemple, la médiane peut être parfois beaucoup plus intéressante) mais ceci déborde du cadre de ce cours.

Les autres solutions classiques de gestions des paramètres de nuisances s'appuient sur une considération assez surprenante : elles consistent à considérer ces paramètres (les paramètres inconnus) comme des VA. Plus rigoureusement, elles consistent à considérer que le véritable paramètre est une réalisation d'une VA dont on connaît certaines propriétés. Ces connaissances vont être de cette façon injectées dans les équations pour pour améliorer la qualité de l'estimation ou encore pour supprimer ces paramètres, ce que nous verrons plus loin.

II.3.b) Estimation au sens du Maximum A Posteriori (MAP)

Si l'on considère le paramètre inconnu (ou le vecteur de paramètre) comme une VA, on peut utiliser les outils de probabilités. Intuitivement, plutôt que de choisir les arguments qui maximisent la probabilité des échantillons connaissant ces paramètres, on voudrait pouvoir choisir le paramètre maximisant la probabilité du paramètre connaissant les échantillons. C'est ce que propose l'approche du *maximum a posteriori*.

Rigoureusement, on choisira $\hat{\theta}_i = argmax_{\theta_i}\{\prod_{k=1}^{n} p(\theta_i / x_k)\}$

Pour calculer cette quantité, nous devons pouvoir calculer les $p(\theta_i / x_k)$. Pour cela, nous pouvons utiliser les lois de Bayes :

$$p(\theta_i / x_k) = p(\theta_i, x_k) / p(x_k) = p(x_k / \theta_i)\, p(\theta_i) / p(x_k)$$

La quantité $p(x_k / \theta_i)$ est celle que nous avons manipulé précédemment (elle est donc calculable). $p(\theta_i)$ représente la densité de probabilité d'une valeur particulière de θ_i . C'est à travers cette grandeur que l'on injecte nos connaissances a priori (on peut savoir que le paramètre a une valeur comprise entre deux bornes par exemple, ce qui est toujours mieux que rien !). On peut ensuite faire disparaître la grandeur $p(x_k)$ est disparaît des calculs car elle est indépendante de θ_i (donc sans intérêt pour la maximisation).

Nous pouvons ainsi dire que l'estimation au sens du maximum a posteriori est donnée par :

$$\hat{\theta}_i = argmax_{\theta_i}\{\prod_{k=1}^{n} p(\theta_i / x_k)\, p(\theta_i)\}$$

Ceci revient en fait à pondérer l'importance des observations dont on dispose en fonction des informations dont on dispose sur la probabilité du paramètre cherché. Notons que si les θ_i sont équiprobables, estimateur du MV et du MAP sont équivalents.

II.3.c) *Approche bayesienne de la gestion de paramètres de nuisance*

Ici encore, on considèrera le paramètre de nuisance comme une VA sans l'estimer. De fait, il est assez pénible de ré-écrire tout ceci. Dans une première version de ce cours, cette approche ne sera pas abordée. Disons que cela consiste a intégrer le paramètre sur toutes ses valeurs possibles...

II.4 Conclusion sur la théorie de la décision bayesienne

Nous avons vu au cours de ce chapitre que sous réserve de connaître le nombre de classes, leur probabilité et la forme analytique des densités de probabilité des caractéristiques considérées comme des VA, nous pouvions définir un classifieur optimal du point de vue de l'erreur moyenne. Nous avons également généralisé ceci aux cas ou l'objectif n'est pas de minimiser cette erreur moyenne, mais un coût variable en fonction du type d'erreur en introduisant les fonctions de risque. Ceci correspond souvent dans la pratique à un cas d'école, les densités de probabilités ne sont pas si facilement accessibles. De fait, ce classifieur nous indiquerait la difficulté intrinsèque du problème. Aucun classifieur ne pouvant avoir de meilleurs résultats (en moyenne), d'autres techniques auront nécessairement des résultats moins bons.

Si les densités de probabilités ne sont pas connues de façon « complète », il arrive fréquemment que l'on dispose d'une forme paramétrique de ces densités (par exemple, les VA suivent des lois de poisson, dont les paramètres sont différents en fonction de la classe de l'item présenté, ou encore les VA suit une loi de poisson pour une classe et une loi gaussienne pour une autre classe.....). Ces paramètres des lois, pour lesquels nous n'avons pas d'intérêt particulier sont appelés paramètres de nuisance. Lorsque nous disposons d'exemples d'apprentissage, nous avons vu trois méthodes pour la gestion de ces paramètres à l'aide des exemples. Deux de ces méthodes (MV et MAP) permettent de les estimer alors que la troisième (Approche Bayesienne) les intègre. Conformément a ce que nous avons dit ci dessus, toutes ces méthodes sont sous-optimales par rapport au cas ou ces paramètres sont connus. Techniquement, le MV est le plus facile à appliquer, et le MAP ainsi que l'Approche Bayesienne permettent d'injecter simplement des connaissances a priori sur ces paramètres si ces connaissances sont disponibles.

On couvre ainsi toute une gamme de cas. Néanmoins, ils arrive également (le plus souvent a dire vrai) que l'on dispose d'exemples mais d'aucune information sur la forme des densité de probabilité. Un autre cas pratique concerne les VA dont on observe que la densité de probabilité est multimodale. Il n'y a alors pas de loi parmi les plus classiques permettant de définir une formule analytique pour ces densités de probabilité (on peut néanmoins trouver des solutions, par exemple, recourir à des mixtures de gaussiennes). Considérons donc que nous ne disposons pas de ces densités. Le traitement de cette classe de problème va faire l'objet du chapitre suivant.

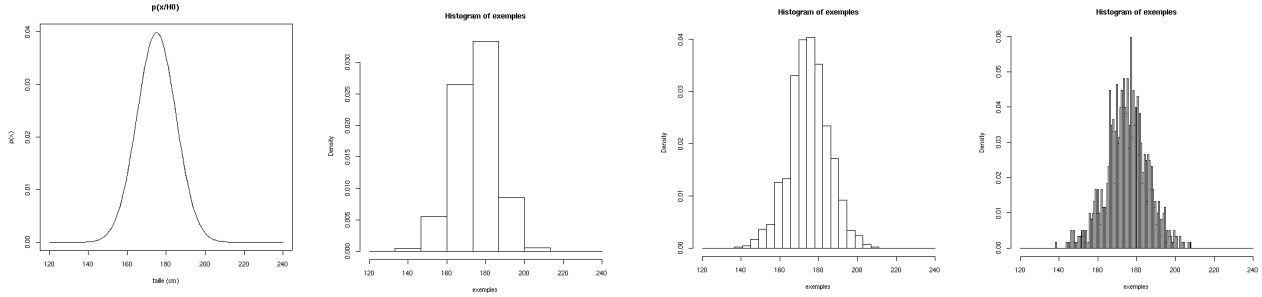
III Apprentissage supervisé et estimation de densités de probabilité

Dans cette partie, nous ne supposons plus disposer des densités de probabilité des échantillons. Ceci pour deux raisons possibles : Soit on ne les connaît pas, soit leur forme ne correspond pas a des lois classiques. Ces dernières sont en général multimodales. Or des hypothèses erronées sur une classe peut amener à des résultats catastrophiques. Dans l'exemple suivant, on dispose d'échantillons répartis suivant l'exemple ci-dessous (on suppose que l'espace des observations est R) et où on fait l'hypothèse que la loi suivie par la classe est normale.

Pour faire face à ces problèmes, nous allons donc nous employer à estimer les densités de probabilité pour pouvoir les utiliser comme précédemment. Pour cela, nous disposons d'exemples dont les véritables classes sont connues (il s'agit bien d'apprentissage supervisé). Ce domaine n'étant plus d'une grande nouveauté, il viendra sans doute à l'esprit du lecteur (et avec raison) que ce que l'on recherche est assez proche d'un histogramme de la répartition des caractéristiques.

III.1 Concepts de l'estimation de densité.

Dans le cas le plus simple, calculer un histogramme consiste à diviser l'espace des caractéristiques en « boites ». On compte ensuite le nombre d'échantillons dont on dispose qui tombent dans chaque boite. En divisant chacun des nombres obtenus par le nombre d'échantillons total, on obtient alors une estimation de la densité de probabilité des échantillons. Cette densité de probabilité correspond à une version discrétisée de la densité de probabilité originelle (on dispose d'un chiffre par boite, considéré valable pour tout le volume de la boite).



On peut voir dans les exemples qui précèdent que la taille des boites doit être liée au nombre d'échantillons dont on dispose. Idéalement, le nombre d'échantillons tends vers l'infini et la taille des boite tend vers zéro. On peut montrer que sous certaines conditions (liant volume des boites et nombre d'échantillons), la fonction décrite par l'histogramme tends vers la densité de probabilité de la VA lorsque le nombre d'échantillons tends vers l'infini.

Posons maintenant le problème de façon plus rigoureuse (et apportons quelques modifications...) : Notre objectif est d'estimer une densité de probabilité p(x) pour toutes les valeurs de x possibles à partir de n échantillons. Nous souhaiterions de plus éviter l'approximation par histogramme qui s'intéresse exclusivement aux valeurs de p(x) au centre de boites définies de façon arbitraire. Pour cela, intéressons nous à la densité de probabilité p(x) au point x0. Nous allons toujours considérer une boite, mais le centre de boite est x0 (contrairement au cas des histogrammes).

Soit R(x0) une région centrée sur x0. Notons P la probabilité qu'un échantillon tiré au hasard tombe dans R(x0). On a :

$$P = \int_{R(x_0)} p(x)\,dx$$

Notons Pk la probabilité que k échantillons sur les n tombent dans R(x0). Pk suit une loi binomiale et l'on a :

$$P_k = C_k^n P^k (1-P)^{(n-k)}$$

On a bien $E[k] = nP$

Par ailleurs, si R(x0) est suffisamment petite, on peut considérer que P est constante sur toute la région (de volume V) :

$$P = \int_{R(x_0)} p(x)\,dx \quad \approx \quad p(x0) \int_{R(x_0)} dx \quad = \quad p(x_0)V$$

$$\hat{p}(x_0) = P/V$$

La dernière relation nous donne p(x0) en fonction de P et V. La précédente nous donne P en fonction de E et n. on a donc aisément :

Propriété 24. $\hat{p}(x_.) = \dfrac{k/n}{V}$

Lorsque le nombre d'échantillons disponible (n) tends vers l'infini, nous désirons une résolution de plus en plus précise. il faut donc que V tende vers 0. Par ailleurs, pour un volume V donné, si P est différent de zéro, il faut que k tende vers l'infini. Enfin, pour que p(x0) ne soit pas nul, il faut que la quantité k/n tende vers zéro.

Il faut donc tout d'abord relier V et k à n. Notons ces quantités Vn et kn. Ce qui précède peut maintenant s'écrire :

$$(١)\ \lim_{n\to+\infty} \quad V_n = ٠$$
$$(٢)\ \lim_{n\to+\infty} \quad k_n = +\infty$$
$$(٣)\ \lim_{n\to+\infty} \quad k_n/n = ٠$$

Notre problème étant posé, on peut déjà voir deux possibilités :

– Choisir un volume (taille adaptée à n) et « compter » les échantillons présents dans ce volume. C'est la méthode des fenêtres de Parzen.

– Choisir un nombre kn (adapté a n) et trouver le volume englobant ces kn voisins. C'est la méthode des k plus proches voisins.

III.2 Fenêtres de Parzen

Dans le cas des fenêtres de Parzen, une attention particulière est mise sur la définition des régions $R_n(x_0)$. Commençons par considérer un hypercube de coté h_n centré sur x_0 .

On a facilement $V_n=(h_n)^d$

Pour aller plus loin, définissons la fonction d'appartenance à un hypercube centré sur 0 de largeur 1 :

$\phi(u)=١ \quad si|u|<١/٢$
$\phi(u)=٠ \quad sinon$

Un point x appartient à un hypercube centré sur x0 et de largeur hn si et seulement si : $\phi(\frac{(x-x_0)}{h_n})=1$

Si l'on note {t1, t2, ...,tn} l'ensemble des échantillons dont on dispose, le nombre d'échantillons dans l'hypercube qui nous intéresse est alors donné par :

$$k=\sum_1^n \phi(\frac{(t_i-x_0)}{h_n})=1$$

On peut alors revenir à la densité de probabilité exprimée par la propriété 24 :

$$\hat{p}(x_0)=\frac{k_n/n}{V_n}=\frac{1}{n}\frac{1}{h_n^d}\sum_1^n \phi(\frac{(t_i-x_0)}{h_n})$$

Ceci revient à faire une convolution entre les positions de nos échantillons et la fonction $\phi(x)$. Celle ci est appelée noyau de l'estimateur.

Lorsque ceci est compris, on peut alors faire varier la forme de ce noyau. L'objectif majeur sera le suivant : un pixel « loin » de x0 devrait moins contribuer à l'estimation de p(x0). Il nous faudra simplement respecter les conditions suivantes pour que ce que l'on estime reste homogène à une densité de probabilité :

$$\forall x\in R^d, \quad \phi(x)\geq 0$$
$$\int_{R^d}\phi(x)\,dx=1$$

Quelques exemples de noyaux couramment utilisés :

fonction	***courbe***
Noyau rectangulaire $\phi(u)=١ \quad si\|u\|<١/٢$ $\phi(u)=٠ \quad sinon$	
Noyau triangulaire $\phi(u)=١-\|u\| \quad si\|u\|<١$ $\phi(u)=٠ \quad sinon$	
Noyau normal	

fonction	*courbe*
$\phi(u)=\frac{1}{\sqrt{2\pi}}e^{u^2/2}$	
Noyau exponentiel $\phi(u)=\frac{1}{2}e^{-\lvert u\rvert}$	

Ici, le noyau joue un rôle de fonction d'appartenance généralisée : un point appartiendra plus ou moins a la région en fonction de son éloignement de x0. L'appartenance totale à la région est exclusivement obtenue en x0 puisque l'on a : $\phi(0)=1$.

Essayons de mieux comprendre le sens de cette estimation : L'estimation de p(x0) par $\hat{p}_n(x_0)$ peut se comprendre comme la « superposition » de n fonctions (les noyaux centrés sur les n échantillons définissant la classe). On a vu différents types de noyaux et on comprend aisément que dans la définition de ces fonctions plus h_n est petit plus le nombre d'échantillons ayant une influence sur $\hat{p}_n(x_0)$ sera faible.

Dans la figure suivant nous présentons différentes valeurs de h_n et sa conséquence sur $\hat{p}_n(x_0)$

On voit donc que le paramètre hn joue un rôle primordial pour une estimation de qualité. Il demeure néanmoins un problème : La valeur de ce paramètre peut être fixée en fonction du nombre d'échantillons (pour que le volume décroisse moins vite que 1/n) . Disons par exemple qu'on se donne une certaine largeur h_1 lorsque l'on dispose de 1 échantillon. Pour n échantillons, on pourra prendre $h_n=h_1/\sqrt{n}$. Ceci étant, on peut prendre une valeur arbitraire de h_1 sans altérer la convergence. Disons donc que pour un nombre donné d'échantillons, il nous faut choisir plus ou moins arbitrairement la largeur h_n .

Par rapport à la méthode classique des histogrammes, nous avons fait un grand pas : Certes, il nous reste un paramètre a définir, correspondant plus ou moins à la largeur des boites de l'histogramme, mais nous avons maintenant une notre densité de probabilité est maintenant une fonction continue (si les noyaux le sont) et nous estimons sa valeur en tout point de l'espace des caractéristiques sans regrouper les points en considérant les densités de probabilités constantes sur une région donnée.

Il est alors intéressant de s'intéresser à la complexité algorithmique de cette estimation ce que je ferais essentiellement à l'oral dans ce cours.

III.3 Méthode des k plus proches voisins

Ici, le principe est extrêmement simple, comme cela a été dit auparavant, la méthode des k plus proches voisins consiste a fixer le nombre de points qui vont compter pour l'estimation de la densité de probabilité en x0. On cherche alors le plus petit volume centré sur x0 contenant ces k voisins. Encore une fois, la densité de probabilité estimée est donnée par la propriété 24. Il reste uniquement a définir le comportement de la grandeur k lorsque n tends vers l'infini.

On peut par exemple prendre k=1 pour n=1 puis $k_n=k_1\sqrt{(n)}$ (en prenant la partie entière de ce résultat pour avoir garder kn entier. Le volume Vn doit alors être défini, et on pourra prendre un hypercube. Pour un point x0 donné, on cherchera le plus petit hypercube contenant les kn points les plus proches de x0.

Ceci nous dispense du problème de définition de largeur des noyaux. Ceci se fait au détriment d'une chose : la probabilité estimée n'est nulle part nulle (le volume Vn n'est jamais infini). Par ailleurs, si les caractéristiques ne sont pas normalisées, considérer un hypercube pose un problème : on ne tient pas compte de la dispersion des données sur chaque caractéristique.

III.4 Résultats expérimentaux

III.5 Conclusion

Nous sommes maintenant, sous réserve de disposer d'exemples pré-classifiés, de construire un classifieur « aussi bayesien que possible » dans la mesure ou les densités de probabilités ne sont pas connues mais estimées. Ce classifieur n'est donc pas optimal. Cependant, il peut constituer une référence intéressante pour des comparaisons entre méthodes. Lorsqu'un nouvel item est présenté à notre classifieur, il faut construire son vecteur de caractéristiques, puis estimer quelle est la probabilité d'appartenance à chaque classe compte tenu de ce vecteur. Pour cela, nous estimons principalement les densités de probabilité du vecteur, les classes étant supposées connues.

Ceci se fait en un temps algorithmique variable en fonction de la complexité de calcul (ou d'estimation) de la densité de probabilité. Par exemple, si les densités sont connues et gaussienne, le temps de calcul est constant pour chaque nouvel item (calcul d'une exponentielle). Si on estime la densité de probabilité par la méthode des fenêtres de Parzen, pour un noyau rectangulaire et n échantillons dans un espace a d dimensions, il faudra n*d tests (le vecteur appartient ou pas a chaque rectangle). Pour le même cas et des noyaux triangulaires, il faudra n*d multiplications. Dans le cas de fenetres normales, il nous faudra n*d calculs d'exponentielles. Enfin, dans le cas d'estimation par la méthode des k plus proches voisins, il faut calculer la distance de chaque échantillon au point x0, puis trier ces distances pour choisir les points les plus proches. Le tri est l'opération la plus coûteuse et se fait en moyenne en n*log(n) opérations. On voit clairement que la charge de calcul n'est pas la même, et peut rapidement devenir rédibitoire !

Des lors que ce calcul devient trop coûteux, il arrive que l'on recherche non plus une solution optimale, ni même efficace, mais une solution calculable dans des temps raisonnables. Dans ce cas, une solution consiste a prendre un classifieur dont la méthode de calcul soit figée et de complexité correcte. On parle alors de classifieurs a forme imposée, ce qui va faire l'objet du chapitre suivant.

IV Classificateurs a forme imposée

Dans ce chapitre, nous nous plaçons encore dans le cadre de la classification supervisée. Il s'agit cette fois de définir la forme du classifieur (ie : on paramètre la forme des régions de décisions, ou encore la forme des fonctions discriminantes). L'objectif, le plus souvent, est de limiter la complexité algorithmique du classifieur. Nous verrons au cours des sections suivantes le cas de **classifieurs linéaires**, ainsi que la méthode du plus proche centre de classe ou encore la **méthode des k plus proches voisins**,....

Précisons un peu les notations : Nous disposons d'un nombre N d'échantillons pré-classés (nous sommes dans le cas de l'apprentissage supervisé) en M classes.

Notons $X=\{x_k, k\in\{1..\mathrm{N}\}\}$ l'ensemble des échantillons. A chacun de ces échantillons correspond un label $l_k=L(x_k) \quad avec \quad l_k\in\{1..\mathrm{M}\}$.

Chaque échantillon est un point dans l'espace des caractéristiques. L'ensemble des échantillons d'une classe constitue un nuage de points que nous appelerons Ci.

$$C_i=\{x_k, k\in\{1...\mathrm{N}\} \quad tq \quad L(x_k)=i\}$$

IV.1 Classifieurs basés distance.

L'idée centrale de tous les classifieurs basés distance est de se donner une métrique de distance entre un point et un nuage. Le classifieur construit utilise comme fonction discriminante :

$$g_i(x)=-d(x, C_i)$$

Dit plus simplement : pour classifier un point, on calcule la distance de ce point a chacun des nuages et on affecte au point le label du nuage le plus proche.

Dans cet ensemble de méthode, nous disposons d'une grande liberté : la définition de la distance utilisée.

Il s'agit d'une distance entre un point et un nuage. Nous allons nous ramener a la définition d'une distance entre deux points. On peut par exemple (et non exhaustivement) définir la distance entre un point et un nuage comme la distance entre le point et :

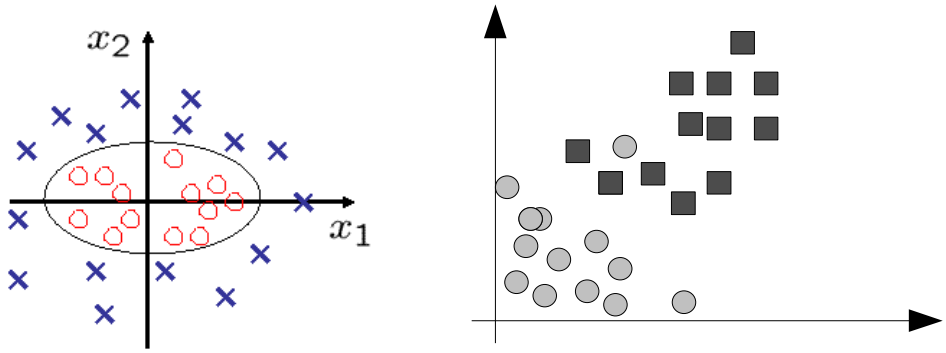
- le barycentre du nuage (centre de classe)
- le point le plus proche du nuage (plus proche voisin)

Reste a définir la distance entre deux points, par exemple (et pas plus exhaustivement que précédemment) :

- La distance euclidienne (L2)
- La distance City Block ou Manhattan (L1)
- La distance ?? (L infini)

Ceci nous donne alors une quantité de classifieurs possible. Il n'y a, a priori, aucune raison de préférer l'un ou l'autre de ces classifieurs en termes de performances (ils se distinguent éventuellement en termes de temps de calculs). Dans la pratique, on choisira telle ou telle méthode en fonction des cas.

Voici quelques exemples pour vous aider a effectuer vos choix. Prenez soin de voir quelles régions de décisions vont émerger suite a vos choix.



Enfin, notez l'importance d'un changement d'échelle sur les caractéristiques.

(A Faire)

Dans le deuxième exemple de la figure précédente, on peut imaginer que le point rond isolé soit une erreur ou encore qu'il soit peu probable. Dans le cas d'une approche au sens du plus proche voisin, ce point va avoir une importance sans doute démesurée sur la région qui l'entoure. Une idée consisterait alors a ne pas considérer que le plus proche voisin, mais de faire voter les k plus proches voisins....

On peut alors avoir différentes options :

- Chaque voisin parmi les k plus proches a une voix (et vote pour sa classe)
- Chaque voisin parmi les k plus proches a un poids lors du vote inversement proportionnel à sa distance au point.

A ce stade, en faisant un petit effort, on peut montrer une similarité très forte entre ces notions de classifieurs basé distance et un classifieur "bayesien" pour lequel les densité de probabilité des classes ont été estimées a l'aide de méthodes telles que les fenêtres de Parzen ou celle des k plus proches voisins.

IV.2 Classifieurs linéaires et SVM

Concentrons nous sur le cas a deux classes qui se généralise aisément. Nous disposerons donc de deux fonctions discriminantes $g_i(x), i \in \{0,1\}$. Attention : Pour simplifier les notations plus loin, les labels possibles pour x seront

respectivement -1 et 1 (et non 0 et 1 comme précédemment)

On parle de classifieur linéaire si les fonctions discriminantes sont des combinaisons linéaires des caractéristiques :

$$g_i(x) = -\sum_{j=1}^{d} (w_i[j]x[j]) + b_i$$

Que l'on peut exprimer de la façon suivante : pour chaque classe, on définit un vecteur $w_i = [w_i[1], ..., w_i[d]]$ correspondant aux poids relatifs de chaque composante de x pour la classe i. Le terme b_i est un terme supplémentaire permettant de positionner la valeur de chaque fonction discriminante a l'origine.

On peut écrire pour simplifier les notations

$$g_i(x) = \langle w_i, x \rangle + b_i$$

Ceci donne comme fonction de classification :

$f(x) = 1 \quad ssi \quad g_1(x) > g_0(x)$ soit $f(x) = 1 \quad ssi \quad \langle w_1 - w_0, x \rangle + b_1 - b_0 > 0$

La surface séparant les deux classes est donc donnée par l'équation suivante : $\langle w, x \rangle + w_0 = 0$

Cette équation correspond a un hyperplan de dimension d-1.

Notre problème se résume donc a trouver l'hyperplan qui sépare au mieux nos classes. On cherche donc le vecteur w et la valeur b qui réalisent cette séparation.

Plus précisément, nous allons chercher l'hyperplan permettant de séparer au mieux les exemples dont on dispose.

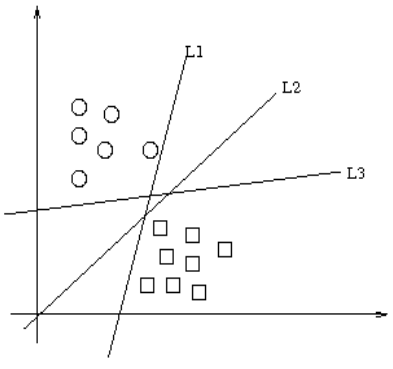
Si nos exemples sont linéairement séparables, il existe w et b tels que :

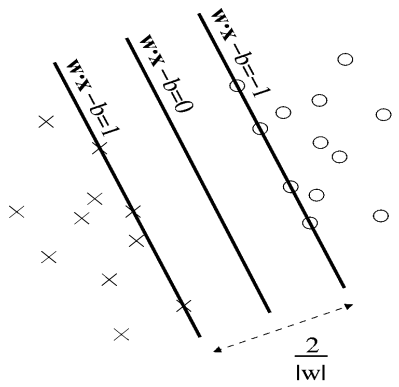
$$\langle x_i, w \rangle + w_0 > 0 \quad si \quad y_i = 1$$
$$\langle x_i, w \rangle + w_0 < 0 \quad si \quad y_i = -1$$

Soit encore

$$y_i \langle x_i, w \rangle + w_0 \; > \; 0 \quad , \quad \forall i \in \{1, ..., M\}$$

Un problème apparaît alors de façon évidente : il n'y a pas qu'une seule solution. La figure suivante illustre ce phénomène. Il faut alors un critère correspondant a « Le meilleur hyperplan possible ». L'idée de base des SVM est de choisir la droite la plus distante de ses points les plus proches.





On a alors évidemment deux hyperplans parallèles (H0 et H1) qui viennent toucher certains échantillons de la classe

correspondante. La séparatrice est l'hyperplan situé à égale distance de ces deux plans. Les échantillons touchés par H0 et H1 sont appelés les vecteurs supports du classifieur puisque supprimer tous les autres points d'apprentissage ne changerait rien au classifieur ainsi trouvé. La distance entre l'hyperplan cherché et H0 (ou H1) est appelée la marge du classifieur (que l'on notera δ)

Voyons comment s'exprime la recherche de ces hyperplans ainsi que la marge obtenue :

De fait, on veut $y_i\langle x_i, w\rangle + w_0 > \delta \quad , \quad \forall i \in \{1, \ldots, M\}$ avec delta maximal.

On peut ré-écrire ceci sous la forme :

$$y_i\langle x_i, w/\delta\rangle + w_0/\delta > 1 \quad , \quad \forall i \in \{1, \ldots, M\}$$

Notre problème est donc maintenant de trouver les $w \quad et \quad w_0$ qui maximisent δ sous contrainte que chacune de ces inégalités soient vraies.

En posant $w' = w/\delta \quad et \quad w_0' = w_0/\delta$, notre problème devient :

minimiser $|w|'$ sous les contraintes : $y_i\langle x_i, w'\rangle + w_0' - 1 > 0 \quad , \quad \forall i \in \{1, \ldots, M\}$

Dans la pratique, n'importe quelle technique d'optimisation convient. On peut par exemple, si l'espace de recherche n'est pas trop grand (c'est rare), tester toutes les valeurs possibles pour $w \quad et \quad w_0$.

Dans la solution classique des SVM, on va déplacer le problème pour se ramener a un problème d'optimisation convexe.

Ceci sera fait en utilisant les multiplicateurs de Lagrange : En se donnant un multiplicateur α_i par contrainte, on écrit le lagrangien :

$$L_P(w', w_0, \alpha) = |w'| - \sum_{i=1}^{M} \alpha_i (y_i \langle x_i, w' \rangle + w_0 - 1)$$

Qu'il faut minimiser sous contraintes que chaque α_i soit positif ou nul et que la derivée de L_p par rapport à chaque α_i soit nulle.

Pour des raisons qu'objectivement j'aurais du mal a vous exposer (et que l'on appelle conditions de Kuhn-Tucker), ceci revient de fait a maximiser L_p sous contrainte que le gradient de L_p par rapport à w' et b' s'annule (et toujours $\alpha_i \geq 0$)

On obtient ainsi les conditions suivantes :

$$L_P = |w'| - \sum_{i=1}^{M} \alpha_i (y_i \langle x_i, w' \rangle - b' - 1) \text{ et } \sum_{i=1}^{M} \alpha_i y_i = 0$$

Ce que l'on peut ré-injecter dans l'équation ci-dessus :

$$L_D = \sum_{i=1}^{M} \alpha_i - 1/2 \sum_{i=1}^{M} \sum_{j=1}^{M} \alpha_i \alpha_j (y_i y_j \langle x_i, x_j \rangle)$$

Ce problème étant convexe, on dispose de méthodes rapides pour résoudre ce problème, qui revient a trouver les α_i adaptés à notre problème. Ceci est fait de façon numérique car on ne dispose pas de résultat analytique dans le cas général.

Lorsque cette optimisation est terminée, on dispose donc des valeurs de chaque α_i. Chaque α_i correspond à un échantillon d'apprentissage. La plupart des α_i sont nuls. Les α_i non nuls correspondent aux vecteurs de supports (les points placés en bordure de marge).

On montre également que la valeur de la fonction pour tout x est donnée par :

$$f(x) = \sum_{i=1}^{M} \alpha_i y_i \langle x_i, x \rangle$$

Dans laquelle on compte tous les α_i. Dans la pratique, la plupart de ces valeurs sont nulles, sauf celles des vecteurs de support, ce qui rend la classification très rapide à calculer : il suffit de faire les produits scalaires entre le vecteur x a classifier et les vecteurs de support, et les sommer (avec le poids adapté).

Il y a ensuite deux généralisations possibles des SVM :

- Lorsque les échantillons d'apprentissage ne sont pas linéairement séparables. Dans ce cas, la méthode d'optimisation ne convergera pas, car il n'existe aucun hyperplan permettant de respecter les contraintes. Il faut donc modifier le critère de choix de l'hyperplan pour relâcher les contraintes (permettre des erreurs). Ne sera pas évoqué dans ce cours, mais dans l'exposé correspondant fait par l'un de vos camarades.
- Le Kernel trick : Ceci permet dans certains cas d'adapter la technique des SVM lorsque le probleme n'est intrinsèquement pas linéairement séparable. De la même façon, les détails seront développés dans l'exposé correspondant aux SVM fait par l'un de vos camarades.

V Prise en compte du contexte.

Le principal problème de toutes les méthodes que nous avons vu précédemment est qu'elle classifient tous les items de façon indépendante. Elles ne prennent en compte que les caractéristiques de l'item pour prendre une décision. Si ceci suffit dans de nombreux cas, il en existe certains dans lesquelles cette approche est insuffisante. Voici deux exemples :

- Vous devez construire un système de détection de panne sur un système évoluant dans le temps. Nos classifieurs, jusqu'ici, mesurent les caractéristiques à l'instant t, et prennent une décision sur cet instant uniquement. Il est vraisemblable qu'un système capable de prendre en compte ce qui s'est passé auparavant devrait pouvoir faire mieux. Par exemple, si a un instant, vous êtes dans une situation critique, a l'instant d'après, il est beaucoup plus vraisemblable que vous soyez encore dans une situation critique que le contraire.
- Vous devez classifier automatiquement le sexe des individus placés autour d'une table lors d'un repas officiel (disons pour qu'un robot leur apporte une rose ou un ballon de foot...). Votre seule mesure sur ces individus est leur taille. Même en construisant le classifieur bayesien adapté au problème, vous ferez un grand nombre d'erreurs, pour les hommes petits et pour les femmes grandes. En revanche, il doit être possible de prendre en compte l'information suivante : En général (et sauf cas particulier), les convives sont placés en alternance de sexe. Ainsi, une configuration homme / homme est moins probable qu'une configuration homme / femme et la configuration homme/homme/homme est beaucoup moins probable qu'une configuration homme / grande femme / homme...

L'objectif de ce chapitre est de voir un certain nombre de techniques permettant de prendre en compte cette information de contexte. Pour cela, nous verrons des solutions extrêmement pragmatiques pour finir par des solutions plus théoriques (les champs de Markov et champs de Markov cachés)

V.1 Post traitements

Supposons que l'on dispose d'un classifieur bayesien. Une première idée assez intuitive consisterait a observer le résultat de notre première classification (sans prise en compte du contexte) puis a juger de sa pertinence. Pour chaque classement raisonnable, on ne change rien, puis pour les successions « étranges », tenter de modifier la décision. Par exemple, en imposant l'alternance homme / femme. Ceci risque fort de poser quelques problèmes (cette alternance n'est pas certaine) et n'est absolument pas satisfaisante théoriquement...

Une façon plus fine de faire le même genre de chose serait de considérer que, lorsque l'on classifie un individu, les p(Hi) varient en fonction du sexe ses voisins. Par exemple, P(homme) est plus faible que 0.5 lorsque ses voisins ont été détectés comme étant des hommes. On pourrait donc fixer une valeur pour les probabilités pour P(Hi/ sexe estimé voisin de gauche , sexe estimé du voisin de droite) . A l'issue de la première passe, on pourrait détecter les configurations peu probables, et repasser sur ces cas la en remplaçant la valeur de P(Hi) par P(Hi/ sexe estimé voisin de gauche , sexe estimé du voisin de droite). La nouvelle classification est alors adoptée.

Cette dernière solution, assez empirique, a sans doute le mérite de régler plus ou moins certains problèmes. Cependant, elle pose quelques problèmes théoriques et pratiques :

- Comment détecter les successions « étranges » ? Ces successions ne prennent en compte que la décision prise, sans aucun regard sur les caractéristiques ayant conduit a cette décision. Par exemple, si l'on se cantonne a l'exemple du repas, notre méthode se focaliserait sur une succession homme / homme / homme en essayant de changer au besoin le sexe de l'individu du milieu... alors qu'il faudrait sans doute reverifier chacun de ces individus... en prenant en compte leurs propres voisins....
- Quelle étendue a le contexte ? Dans l'exemple que je donne, on ne s'interesse qu'aux deux voisins de gauche et droite. Pourquoi pas les autres ? Ou s'arrêter ?
- Enfin, est il légitime de remplacer P(Hi) par P(Hi/ sexe estimé voisin de gauche , sexe estimé du voisin de droite) alors qu'il faudrait sans doute P(Hi/ sexe vrai du voisin de gauche , sexe vrai du voisin de droite) qui serait plus facile à déterminer (Mme de Rotschild pourrait vous dire que dans 98% des cas, on a une alternance des vrais sexes. Elle serait incapable de vous donner cette valeur pour les sexes estimés....) : La connaissance a priori porte dans la pratique plus souvent sur les valeurs vraies.

En résumé, on peut construire des solutions empiriques qui fonctionneront a peu près sur la base d'un post traitement, mais il sera difficile de construire une version théoriquement fiable du problème. Ceci simplement parce que notre classifieur, au départ, ne prend pas en compte le contexte.

V.2 Injection de caractéristiques liées au contexte dans le classifieur.

Une seconde idée pratique consisterait a simplement ajouter aux caractéristiques de l'item les caractéristiques de son voisinage. Pour en revenir a notre exemple, pour classer un individu, on pourrait choisir comme caractéristiques : sa taille, la taille de son voisin de gauche et la taille de son voisin de droite. Dans ce modèle, on continue a classer les individus d'une façon qui est formellement indépendantes du contexte (on ne tient pas en compte le sexe des voisins), mais ce contexte est tout de même intégré de façon implicite dans le classifieur comme des caractéristiques de l'individu a prendre en compte (je ne suis pas sur d'etre vraiment clair).

Ce type de méthode est très couramment utilisé dans la pratique, car très facile à mettre en oeuvre dans le cadre de l'apprentissage supervisé à forme imposée (K plus proches voisins, classifieurs linéaires, réseaux de neurones, boosting)

Si l'on reprend toujours le même exemple, il faudra disposer d'exemples pré-classifiés....Simplement, nos exemples seront de la forme [1.45, 1.78, 1.65] : homme. (ou 1.45 correspond a la taille du voisin de gauche, 1.78, celle de l'individu a classer et 1.65 celle du voisin de droite). Une fois le classifieur entraîné, on lui fournit un vecteur de paramètres et il agit comme dans les chapitres précédent.

La question qui se pose alors est de savoir s'il est possible de construire un classifieur bayésien sur ce modèle.

Pour cela, il va falloir calculer les p(Hi,x) = P(x/Hi) P(Hi).

Pour en revenir a notre exemple, pour un individu donné, les caractéristiques sont x = [x0,x1,x2] avec x0, la taille du voisin de gauche, x1 la taille de notre individu, x2 la taille du voisin de droite.

P(x/H0) est la probabilité qu'un homme ait une taille x1 et que son voisin de gauche ait une taille x0 et que son voisin de droite ait la taille x2... Chose qui va être relativement difficile à définir dans la pratique (Il nous faudra parler de champs de Markov, ce que nous verrons dans la section suivante).

V.3 Champs de Markov, variantes et applications.

V.3.a) Définitions des Champs de Markov et résultats importants.

Pour commencer, voyons ce qu'est un champ de Markov...De fait, on parle de chaine de Markov pour des processus 1D, et de champs de Markov pour les processus de dimension supérieur. Nous confondrons les deux en nommant tout cela des champs de Markov...

Posons donc quelques définitions (agrémentées d'exemples):

- Soit S un ensemble de sites, ordonné par un indice. (S est l'ensemble des sièges ou s'assoient les invités).
- Soit V un systeme de voisinage. Vs est l'ensemble des voisins du site s au sens de V. (pour nous V est : voisin de gauche et voisin de droite). Note : Le site s n'appartient pas à Vs.
- Enfin, à chaque site sk est associé un descripteur xk (pour nous la taille). L'ensemble des descripteurs sera noté X

On considérera chaque descripteur observé comme la réalisation d'une VA Xk. X est également vu comme la réalisation d'une VA.

X est un champ de Markov ssi pour tout site s, on a la propriété suivante :

$$P(X_s = x_s | x_r, r \neq s) \;=\; P(X_s = x_s | x_r, r \in Vs)$$

Dis autrement, X est un champ de Markov ssi la probabilité d'observer une valeur en un site ne dépend que de la réalisations de ses voisins au sens de V.

Il s'agit clairement d'une hypothèse restrictive qui nous permettra de définir des modèles...

Le principal problème est de définir un champ de Markov qui corresponde à nos données...

- On appelle une clique : soit un singleton de S (un siège), soit un ensemble de sites tous voisins au sens de V (pour nous une clique a forcement la forme {sk,sk+1}. L'ensemble de toutes les cliques possibles sur S est noté C.
- On appelle Potentiel d'une clique c la quantité Uc, qui dépend de la valeur des descripteurs associés a la clique c. Ce potentiel décrit l'interdépendance entre les sites de la clique.

On appellera Potentiel du champ complet la quantité suivante :

$$U = U(x) = \sum_{c \in C} U_c$$

On appelera Potentiel d'un site la somme des potentiels des cliques auquels ce site appartient :

$$U_s = \sum_{c \in C,\, s \in c} Uc$$

Le théorème d'Hammersley-Clifford nous indique que X est un champ de Markov ssi

$$P(X_s = x_s | x_r, r \neq s) = \frac{\exp^{-U_s(X_s = x_s)}}{\sum_{e \in E} \exp^{-U_s(X_s = e)}}$$

Et par ailleurs la probabilité d'une configuration est donnée par :

$$P(X = x) = \frac{\exp^{-U(X = x)}}{\sum_{z \in Omega} \exp^{-U(X = z)}}$$

Notons encore une fois que le dénominateur n'est quasiment jamais calculable. Neanmoins, s'il s'agit d'estimer laquelle de deux configurations est la plus probable, ce terme apparaît comme une constante.

Nous avons un peu progressé puisque maintenant, si l'on définit une fonction de potentiel pour chaque clique, on peut relativement simplement calculer les quantité précédentes.

Voyons un peu la pratique de tout cela :

1. On commence par définir le potentiel d'une clique. Vu qu'une clique est soit un singleton, soit un ensemble de sites voisins au sens de s, on va avoir différentes formes possibles. Pour les cliques de singletons, Uc = f(xs). Pour les cliques comprenant 2 elements, Uc = g(xs,xs'). En fonction du système de voisinage, on peut avoir des cliques de plus grand cardinal, mais elles sont rarement employées...(Dans le cas de nos sieges, il n'y en a pas et nous nous arreterons la). Encore faut encore définir ces fonctions f et g.

A dire vrai, on ne dispose pas d'un nombre incroyable de possibilités. En général, on se contente de quelques modèles célèbres :

Le modèle d'Ising : utilisé pour des variables aléatoires (Xs) pouvant prendre la valeur 1 ou -1.

Selon ce modèle, pour les cliques d'ordre 1 : U(xs) = -Bxs et pour les cliques d'ordre 2 : Uc = -beta xs xs'.

La seconde forme est intéréssante car vu que les VA peuvent prendre les valeurs 1 ou -1, Uc = - beta si xs=xs' et vaut beta sinon.

Si l'on utilise le modèle d'Ising, si beta est positif, les configuration d'energie minimal (les plus probables) sont celles ou toutes le VA ont la meme valeur. Si beta est négatif, les configurations d'energie maximales sont celles ou toutes les VA ont une valeur opposée.

Le modèle de Potts : on ne considère que les potentiels des cliques d'ordre 2 et celle ci a la meme forme que celle du modele d'Ising, mais généralisé a des VA discretes.

Les configuration d'énergie minimales sont celles ou les voisins prennent la même valeur si beta est positif.

Les autres modèles : ce que vous voulez (le modèle gaussien, non évoqué ici est également classique...)

Il s'agit donc essentiellement de modéliser le problème sous la forme de potentiels entre sites voisins, un potentiel négatif étant considéré « bon », un potentiel négatif étant considéré comme « mauvais »... Un potentiel très négatif etant « très bon » et un potentiel « très positif » étant « très mauvais »...

2. Si l'on souhaite estimer la probabilité d'une configuration autour d'un site, il s'agit de calculer la quantité P(xs/xr) grace à la formule due au théorème d'Hammersley-Clifford.
3. Si l'on souhaite générer un champ de Markov (pourquoi pas)... Il existe de nombreux algorithmes. Le plus simple est celui de l'échantilloneur de Gibbs :

Echantilloneur de Gibbs :

On se donne une configuration initiale quelconque pour les observées, puis on itère le processus suivant :

(i) Choix d'un site s (par tirage selon une loi uniforme sur E)
(ii) Calcul des probabilités conditionnelle P(Xs = e / les voisins) pour toute valeurs de e.
(iii) Mise à jour du site s par tirage aléatoire selon cette loi.

Ceci converge un jour ou l'autre vers un champ de Markov. Dans la pratique, on s'arrête au bout d'un certain nombre d'itération (grand) ou lorsque le champ ne bouge plus trop...

Revenons a notre problème de départ : un classifieur bayesien prenant en compte la taille des voisins de gauche et de droite de l'individu a classer. Nous avons sans doute fait un pas en avant, puisque nous avons maintenant les outils permettant de définir les probabilités d'observer une taille connaissant les tailles des voisins. Ceci dit, le problème serait très mal posé car la taille d'un individu autour de la table n'est pas vraiment liée à celle de ces voisins (sauf règle de bienséance que je ne connaîtrais pas).

Pour résoudre ce problème, nous allons maintenant introduire l'outil principal de cette section : Les champs de Markov cachés, ou Hidden Markov Fields (HMF).

V.3.b) Champs de Markov cachés.

De fait, nos observations sont des tailles (un vecteur Y constitué des tailles de chaque individu). Notre objectif est de trouver le vecteur de label a associé aux individux et représentant leur sexe (X). Dans le cadre bayesien, on cherche la réalisation de Y qui maximise la quantité P(X/Y). Conformément aux lois de bayes, on peut écrire P(X/Y) = P(Y/X) * P(X) / P(Y).

P(Y) représente la probabilité d'observer l'ensemble des tailles obtenu. C'est une constante quel que soit X, que nous ne prendrons pas en compte (comme d'habitude, nous maximisons en fait P(X,Y) = P(Y/X) * P(X).

Dans ce cas précis, nous pouvons considérer que si les sexes sont connus, les observations sont indépendantes entre elles. En effet, les seuls liens statistiques existant pour le placement autour de la table concernent les sexes, pas les tailles (...). Dans ce cas, on a :

$$P(Y/X)=\prod_{s} P(y_s/x_s)$$

Les P(ys/xs) sont les quantités que nous manipulons depuis le départ.

En revanche, P(X) lui a clairement été défini comme étant un champ de Markov. Nous n'observons pas le champ de Markov, il s'agit d'une variable explicative qu'il faut découvrir... Il est caché...

On sait donc maintenant que

$$P(X=x)=\frac{\exp^{-U(X=x)}}{\sum_{z\in Omega}\exp^{-U(X=z)}}$$

Comme cela a été signalé, le terme au dénominateur n'a aucun intérêt pour notre problème. Nous allons donc chercher le maximum de :

$$J(X=x)=\exp^{\ln(P(Y=y/X=x))-U(X=x)}$$

ou encore maximiser :

$$J'(X=x)=\sum_{s}\ln(P(y_s/x_s))-\sum_{c\in C}Uc$$

Vu que la plupart des algorithmes d'optimisation cherchent des minimums globaux, on cherchera le plus souvent a minimiser la quantité

$$J''(X=x)=\sum_{s}-\ln(P(y_s/x_s))+\sum_{c\in C}Uc$$

Pour cela, des algorithmes de type recuit simulé feront traditionnellement l'affaire ...

De façon générale, le premier terme sera nommé terme d'attache aux données (cela va consister a classer les grands individus comme des hommes, et les petits individus comme des femmes). Le second terme est un terme de régularisation et a pour objectif de trouver des solutions correspondant à notre modèle (une alternance de sexes).

Reste à définir un potentiel pour chaque clique : Notre variable de sexe a deux états possibles, nous nous tournons donc vers un modèle d'Ising.

- U(xs) = - B xs
- U(xs, xr) = - beta xs xr avec beta négatif pour avoir une configuration qui tende vers une alternance d'états.

Pour fixer les constantes,

- B est relatif a la probabilité d'avoir un certain sexe. On doit pouvoir définir un modèle plus près de la réalité en définissant un différent par sexe et intuitivement, je dirais qu'il faut prendre p(homme) = e-Bh et p(femme) = e-Bh le tout renormalisé pour que ces quantités se somme à 1 (mais je n'ai pas eu le temps de vérifier).
- Beta sera choisi négatif, plus ou moins grand pour régulariser plus ou moins les solutions. Il me semble aussi que l'on peut le régler proprement en le reliant a des données externes comme la proba d'alternance homme / femme, homme/homme, femme/homme, femme/femme.

Dans la pratique, encore une fois, il s'agit le plus souvent de régulariser une classification ... on choisis donc les potentiels, les valeurs des constantes de façon complètement empirique !