Université des Antilles et de la Guyane Vincent Pagé

UEOMA504 : Traitement Statistique de l'information
Master - Deuxième année

Travaux Pratiques

Feuille 1

Préliminaires :

Il s'agit ici de reprendre les exemples du cours, afin de visualiser chacune des notions que vous avez pu y voir. On s'intéresse donc aux individus âgés de plus de 20 ans dans travaillant dans le domaine de l'informatique, l'objectif étant de les classer par sexe en fonction de leur taille.

Voici les informations dont on dispose :

- Parmi ces informaticiens, une proportion p_1 est de sexe féminin, p_0 de sexe masculin.
- l'hypothèse de sexe masculin sera notée H_0 , celle de sexe féminin sera notée H_1 . Pour un individu donné, la mesure de sa taille sera notée x

Questions :

A quoi correspondent les grandeurs suivantes ? Comment sont elles liées ?

- $p(x/H_0)$, $p(x/H_1)$, $p(H_0,x)$, $p(H_1,x)$, $p(x)$, $p(H_0/x)$, $p(H_1/x)$

Si nous utilisons la taille des individus pour classifier les individus selon leur sexe, cela revient à affecter à chaque taille possible un label (0 pour les hommes, 1 pour les femmes). On découpe donc l'ensemble des tailles en 2 régions (R_0 , R_1). Un individu dont la taille appartient a R_i se voit affecter le label i . Le choix de ces régions définit ainsi notre classifieur.

Quel que soit le classifieur, il est vraisemblable qu'il commette des erreurs pour chaque x donné. En moyenne, ce classifieur fera deux types d'erreurs. Notons $P(\epsilon_0)$ et $P(\epsilon_1)$ les probabilités d'erreurs qui consistent respectivement à classer un individu de sexe féminin comme un homme et inversement.

Questions :

Comment exprimez vous ces probabilités d'erreur en fonction de R_0 , R_1 et des différentes densités de probabilités ? Comment exprimez vous la probabilité d'erreur totale ?

Pour définir un classifieur de bonne qualité, une première idée consiste à minimiser la probabilité d'erreur totale. Le classifieur ainsi construit sera noté C_b (Classifieur de Bayes).

Questions :

- Comment construisez vous C_b ?

Applications :

Nous allons utiliser le logiciel de statistiques nommé « R ». Ce logiciel peut utiliser des « scripts » (des programmes) pré-écrits pour vous aider a réaliser ce TP. Ces scripts sont disponibles sur internet a la même adresse que celle ou vous avez trouvé ces énoncés.

Dans tout ce qui suit, on va supposer les choses suivantes :

- La courbe de répartition de taille des hommes informaticiens est supposée gaussienne, de moyenne moyenne m_0 , d'écart type σ_0 .
- La courbe de répartition de taille des femmes informaticiennes est supposée gaussienne, de moyenne m_1 , d'écart type σ_1 .

Premier exemple : on suppose $m_0=1.78$ et $\sigma_0=0.1$, $m_1=1.68$ et $\sigma_1=0.08$, $p_1=0.4$

1. Sur un premier graphique, tracer les courbes $p(x/H_0)$, $p(x/H_1)$
2. Sur un second graphique, tracez les courbes $p(H_0,x)$, $p(H_1,x)$, $p(x)$
3. Sur un troisieme graphique, tracez les courbes $p(H_0/x)$, $p(H_1/x)$
4. Définissez le classifieur bayésien adapté a ce problème.
5. Considérons un classifieur fonctionnant de la façon suivante : Si x est inférieur a un seuil s donné, l'individu est déclaré « femme » sinon, il est déclaré « homme ». Tracer la courbe de probabilité d'erreur totale en fonction du seuil. Pour quel seuil l'erreur totale est elle minimale ?

Refaite les manips pour $p_1=0.5$. Voyez vous une simplification permettant de resoudre le problème plus rapidement ?

Refaite les manips pour $p_1=0.5$, $\sigma_0=\sigma_1=0.1$. Voyez vous une simplification permettant de resoudre le problème encore plus rapidement ?

Aide pour le logiciel R :

En plus du polycopié qui vous a été fourni, voici quelques indications sans doute utiles :

- L'aide de R s'appelle avec la fonction help() ou help (truc) pour avoir l'aide sur « truc »
- Ne tapez rien directement dans R ! tapez plutôt toutes vos instructions dans un fichier texte que vous appellerez dans R avec l'instruction source("monfichier");
- Pour calculer la valeur de $p(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-(x-m)^2}{\sigma^2}}$, utilisez la fonction dnorm(x,m, sigma)
- Pour calculer $\int_{-\infty}^{s}\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-(x-m)^2}{\sigma^2}}dx$, utilisez pnorm(s, m,sigma)
- Pour tracer des courbes, inspirez vous du script plot_gauss.R fourni à l'adresse http://calamar.univ-ag/uag/ufrsen/coursenligne/vpage/new_site/