

UEOMA504 : Traitement Statistique de l'information

Master - Deuxième année

Travaux Pratiques

Feuille 3

Préliminaires :

On s'intéresse cette fois-ci à la gestion de paramètres de nuisance. Nous nous concentrerons sur la méthode du Maximum de Vraisemblance et sur celle du MAP qui consistent à estimer ces paramètres.

Considérons que l'on s'intéresse à la classification du TP1 (classification par sexe basée sur la taille uniquement). On sait que la répartition de taille suit une loi gaussienne, de moyenne inconnue et de variance connue (disons 10 cm pour les hommes et 9 pour les femmes). Dans la première partie de ce TP, on étudiera l'estimation de la moyenne. Dans la seconde partie, on appliquera ces méthodes pour bâtir un classifieur bayésien approché pour regarder l'influence de l'estimation sur la probabilité d'erreur moyenne du classifieur.

Estimation de paramètres :

Donnons nous un ensemble d'échantillons déjà classifiés. Nous nous concentrerons sur les échantillons de la classe « hommes ».

- Pour générer ces échantillons, nous utiliserons la fonction `rnorm` avec comme paramètre : moyenne 175 (cm) et écart type 10 (cm). Ces informations seront bien évidemment oubliées par la suite pour l'estimation...
- Le nombre d'échantillons sera noté N .
- Nous nous intéressons à l'ensemble de tailles possibles $[100, 250]$, échantillonné tous les cm.

Estimation de paramètres au sens du MV :

Donner l'expression de la log-vraisemblance pour une valeur possible du paramètre « moyenne ».

CAS 1 : $N=3$

- Tracer la courbe de la vraisemblance en fonction de chaque valeur possible de moyenne.
- Donner la valeur de l'estimateur correspondant au maximum relevé sur cette courbe.
- Donner une expression analytique de l'estimateur et sa valeur.
- Tracer sur un même graphe la véritable densité de probabilité et la densité de probabilité estimée.

CAS 2, 3, 4 : respectivement $N = 20$ et $N = 100$ et $N = 1000$

Mêmes questions. Que constatez-vous empiriquement sur vos tests ?

Estimation de paramètres au sens du MAP :CAS 1 : $N=3$

On prendra comme prior pour la moyenne à estimer l'information suivante : la densité de probabilité de la moyenne cherchée suit une loi gaussienne de moyenne 175 (c'est la véritable valeur du paramètre) et d'écart type 15. Ceci représente notre connaissance a priori sur la valeur à rechercher

- Quelle quantité le paramètre cherché doit il maximiser ?
- Tracer la courbe de la probabilité du paramètre en fonction de chaque valeur possible de moyenne.
- Donner la valeur de l'estimateur correspondant au maximum relevé sur cette courbe.
- Tracer sur un même graphe la véritable densité de probabilité, la densité de probabilité estimée au sens du MV et la densité de probabilité au sens du MAP.

CAS 2,3,4 : respectivement $N = 20$ et $N= 100$ $N= 1000$

Mêmes questions. Que constatez vous empiriquement sur vos tests ?

CAS 5 : Prior minimaliste : On considère que le paramètre cherché est uniformément réparti sur tout l'ensemble des valeurs possibles. Que constatez vous ?

CAS 6 : Prior inadapté : on prendra comme prior l'information suivante : la densité de probabilité de la moyenne cherchée suit une loi gaussienne de moyenne 195 (valeur erronée) et d'écart type 15

Classification supervisée (densités de probabilités connues sous forme paramétrique) :

Nous ne considérons ici que l'estimation au sens du MV. Construire sur la base de N exemples pour chaque classe un classifieur bayésien lorsque les densité de probabilités conditionnelles sont supposées gaussiennes de moyenne inconnue et de variance connue (10)

Comparer ses performances avec celle du classifieur bayésien du TP1 sur une base d'exemples tirée au hasard.