
Master 1 : TP3 ACID

Rendu WEIL Gabriel et TRICOIRE Simon

./ Commentaire sur le code :

Exécution

De base, les 4 méthodes de classification vont s'exécuter (MLE, MAP, RISK, NDIM), et afficher sur la console matlab les pourcentages d'erreurs de classement pour les bars et les saumons en fonction de chaque méthode de classification (fonction : *displayResultClassification()* et *displayResultClassificationNDim()*).

Les trois premières méthodes de classification sont exécutées par la fonction : *exo1_to_exo3()* et la dernière méthode est exécutée par la fonction : *exo4()*.

Affichage courbe

Par défaut l'affichage des courbes dans la fonction *exo1_to_exo3()* est désactivée.

Pour afficher les histogrammes, il suffit d'enlever le commentaire devant la fonction : *displayCurveSet(VTBar, VTSaumon);*.

Note : dans la fonction *exo4()* l'affichage des histogrammes n'est pas présent, nous n'avons pas réussi à trouver comment les afficher lorsqu'ils étaient de dimensions 2.

Pour afficher les courbes d'erreurs des différentes méthodes de classification, là encore il suffit d'aller voir à la fin de ces deux fonctions et d'enlever le commentaire devant la fonction : *displayCurveErrors(nblter, _);* qui nous intéresse.

./ Résultats :

```
Nombre Itération: 100
=== Résultat MLE ===
  Pourcentage Erreur Bar: 22.677500
  Pourcentage Erreur Saumon: 7.791111
=== Résultat MAP ===
  Pourcentage Erreur Bar: 29.027500
  Pourcentage Erreur Saumon: 3.561111
=== Résultat RISK ===
  Pourcentage Erreur Bar: 34.802500
  Pourcentage Erreur Saumon: 1.718889
```

```
Nombre Itération: 100
=== Résultat N Dimension ===
  Pourcentage Erreur Bar: 5.757500
  Pourcentage Erreur Saumon: 0.727778
```

./ Réponses aux questions :

Exercice 2 :

Q1) Quelles sont ici les informations utilisées pour classifier un échantillon (paramètres du classifieur)

La moyenne et l'écart-type des ensembles entraînés (Bar/Saumon) sont les deux informations utilisées pour classifier un échantillon.

Exercice 3 :

Q1) Évaluer le classifieur se basant sur le Maximum A Posteriori (MAP) :

Dans notre cas, le MAP nous permet de calculer la probabilité que notre échantillon, en se basant sur sa taille, appartienne soit à la classe Bar, soit à la classe Saumon. On calcule alors les probabilités pour les deux classes, et on prend la classe pour laquelle la probabilité est la plus haute.

Pour 100 itérations, ce classifieur nous donne un pourcentage d'erreur de classement sur les bars d'environ 29% et 4% pour les saumons.

Ainsi on peut constater que classifieur se trompe beaucoup moins sur les saumons que sur les bars. Ce rapport se retrouvait aussi, lorsque l'on a utilisé la classification se basant sur le Maximum Likelihood (MLE) lors du premier exercice. En effet on avait alors un pourcentage d'erreur de classement sur les bars d'environ 23% et 8% sur les saumons.

On constate que les deux classifications ne donnent pas les mêmes résultats, la classification MAP est plus performante à bien classer les saumons mais classe moins bien les bars. Le seuil de décision a ainsi été agrandi en faveur de la classe des saumons, un échantillon sera plus souvent considéré comme un saumon.

Q2) Évaluer le classifieur se basant sur le MLE mais prenant en compte une fonction de coût

Dans notre cas, notre fonction de coût nous précise qu'il est plus coûteux et donc dommageable de classer notre échantillon en tant que bar alors qu'il s'agit d'un saumon, plutôt que de classer notre échantillon en tant que saumon alors qu'il s'agit d'un bar. Ainsi notre seuil de décision sera plus important pour la classe saumon, ce qui fait qu'on considérera plus souvent notre échantillon comme un saumon que comme un bar.

Pour 100 itérations, ce classifieur nous donne un pourcentage d'erreur de classement sur les bars d'environ 35% et 2% sur les saumons. On constate que le résultat en pratique correspond bien à la logique énoncée plus haut en théorie.

Ainsi comparativement à la classification MAP on remarque que celle intégrant la fonction de coût est un peu plus performante à classer les saumons cependant elle classe vraiment moins bien les bars : le pourcentage d'erreurs sur les bars est d'environ 5% supérieur.

Exercice 4:

Finalement, la dernière classification se basant sur deux caractéristiques, la longueur et la brillance, est la classification la plus performante.

En effet avec environ, 6% d'erreurs de classement sur les bars et seulement 1% sur les saumons, elle classe mieux les échantillons que toutes les autres méthodes de classification.

On remarque qu'elle a drastiquement diminué le nombre d'erreurs sur le classement des bars (environ 25% de moins que les autres classifications) cependant le nombre d'erreurs sur les saumons n'a lui que faiblement diminué (environ 1%).