

1 Principe de la règle paramétrique

On se place dans le cadre de données de dimension 2 avec 3 classes. Nous avons à notre disposition une matrice `test` qui contient l'ensemble d'apprentissage constitué de 3 classes de 50 individus chacune $\in R^2$ (un "professeur" nous a indiqué que les 50 premiers individus sont issus de la classe 1, les 50 suivants de la classe 2, les 50 derniers de la classe 3) et un vecteur `x` qui contient un individu inconnu à classer.

- Proposer sous forme d'organigramme une discrimination de l'individu inconnu à partir de la règle paramétrique avec l'hypothèse Gaussienne.

2 Règle paramétrique vs Kppv

Vous allez créer 2 programmes permettant d'étudier les algorithmes de discrimination étudiés durant le cours. Tout d'abord le programme :

```
[clas]=decision_kppv(test,classe_origine,k,eye(2,2),x);
```

Ce programme applique la méthode de discrimination de kppv sur un ensemble d'individu élément de R^2 . (La métrique utilisée est $M = Id$ ce qui explique le paramètre `eye(2,2)`).

Les paramètres :

- La variable `test` est une matrice qui doit contenir les différents individus de l'ensemble d'apprentissage rangés par colonne. Le nombre de ligne est 2 et le nombre de colonne est m .
- La variable `classe_origine` est un vecteur qui indique la classification de l'ensemble d'apprentissage. `classe_origine(i)` indique le numéro de la classe de l'individu `test(:,i)`.
- La variable `k` indique le nombre de voisins utilisés dans l'algorithme.
- La variable `x` est une matrice qui doit contenir les différents individus à classer rangés par colonne. Le nombre de ligne est 2 et le nombre de colonne est n .

Le résultat :

- La variable `clas` est un vecteur qui indique le résultat de l'algorithme de la discrimination. `clas(i)` indique le numéro de la classe de l'individu `x(:,i)`.
- Réaliser la fonction.

Ensuite vous devez réaliser le programme suivant :

```
[clas]=decision_bayes(m,sigma,p,x);
```

Ce programme applique la méthode de discrimination paramétrique Gaussien sur un ensemble d'individus éléments de R^2 (2 mesures pour chaque individu). La méthode suppose une distribution Gaussienne.

Les paramètres :

- La variable `x` est une matrice qui doit contenir les différents individus à classer rangés par colonne. Le nombre de ligne est 2 et le nombre de colonne est n .
- La variable `m` indique les vecteurs moyennes de chacune des classes rangés par colonne.
- La variable `sigma` indique les matrices de variance/covariance de chacune des classes. La variable est de taille $2 \times 2 \times 3$. L'utilisation de cette variable est :
`sigma=zeros(2,2,3); sigma(:,:,1)=sig1;` avec `sig1` la matrice de variance/cov de la classe 1
- La variable `p` indique les probabilités $P(w_j)$ rangées dans un vecteur. On considère par la suite que les 3 classes sont équiprobables.

Le résultat :

- La variable `clas` est un vecteur qui indique le résultat de l'algorithme de la discrimination. `clas(i)` indique le numéro de la classe de l'individu `x(:,i)`.
- Réaliser la fonction.

2.1 Comparaison standard

- Charger des données `td3_d1.mat`.

Ce fichier comprend trois variables :

- La matrice `test` qui contient l'ensemble d'apprentissage constitué de 3 classes de 50 individus chacune $\in R^2$ suivant 3 distributions gaussiennes. Un "professeur" nous a indiqué que les 50 premiers individus sont issus de la classe 1, les 50 suivants de la classe 2, les 50 derniers de la classe 3.
- La matrice `x` qui contient un ensemble de 300 individus à classer.
- Le vecteur `clasapp` contient le résultat parfait que l'on doit obtenir par cette discrimination.
- Appliquer la méthode de discrimination paramétrique et Kppv (pour $k = 1, 3, 5, 7, 13, 15$). Pour chacun des tests calculer le taux d'erreur. Vous pouvez visualiser le résultat des classifications par la fonction `affiche_classe`.
- Commenter les résultats.

2.2 Absence de professeur

- Charger des données `td3_d2.mat`).

Ce fichier comprend trois variables :

- La matrice `test` qui contient l'ensemble d'apprentissage constitué de 3 classes de 150 individus au total suivant 3 distributions gaussiennes.
- Le vecteur `orig` contient la classification de l'ensemble d'apprentissage par l'algorithme de coalescence.
- La matrice `x` qui contient un ensemble de 300 individus à classer.
- Le vecteur `clasapp` contient le résultat parfait que l'on doit obtenir par cette discrimination.
- Appliquer la méthode de discrimination paramétrique et *Kppv* (pour $k = 1, 3, 5, 7, 13, 15$). Pour chacun des tests calculer le taux d'erreur.
- Commenter les résultats.

2.3 Influence de la taille de l'ensemble d'apprentissage : taille réduite

- Charger des données `td3_d3a.mat`).

Ce fichier comprend trois variables :

- La matrice `test` qui contient l'ensemble d'apprentissage constitué de 3 classes de 20 individus chacune en dimension 2 suivant 3 distributions gaussiennes. Un "professeur" nous a indiqué que les 20 premiers individus sont issus de la classe 1, les 20 suivants de la classe 2, les 20 derniers de la classe 3.
- La matrice `x` qui contient un ensemble de 300 individus à classer.
- Le vecteur `clasapp` contient le résultat parfait que l'on doit obtenir par cette discrimination.
- Appliquer la méthode de discrimination paramétrique et *Kppv* (pour $k = 1, 3, 5, 7, 13, 15$). Pour chacun des tests calculer le taux d'erreur.
- Commenter les résultats.

2.4 Influence de la taille de l'ensemble d'apprentissage : taille importante

- Charger des données `td3_d3b.mat`).

Ce fichier comprend trois variables :

- La matrice `test` qui contient l'ensemble d'apprentissage constitué de 3 classes de 150 individus chacune en dimension 2 suivant 3 distributions gaussiennes. Un "professeur" nous a indiqué que les 150 premiers individus sont issus de la classe 1, les 150 suivants de la classe 2, les 150 derniers de la classe 3.
- La matrice `x` qui contient un ensemble de 300 individus à classer.
- Le vecteur `clasapp` contient le résultat parfait que l'on doit obtenir par cette discrimination.
- Appliquer la méthode de discrimination paramétrique et *Kppv* (pour $k = 1, 3, 5, 7, 13, 15$). Pour chacun des tests calculer le taux d'erreur.
- Commenter les résultats.

2.5 Distribution inconnue

- Charger des données `td3_d4.mat`).

Ce fichier comprend trois variables :

- La matrice `test` qui contient l'ensemble d'apprentissage constitué de 3 classes de 70 individus chacune en dimension 2 suivant 3 distributions inconnues. Un "professeur" nous a indiqué que les 70 premiers individus sont issus de la classe 1, les 70 suivants de la classe 2, les 70 derniers de la classe 3.
- La matrice `x` qui contient un ensemble de 300 individus à classer.
- Le vecteur `clasapp` contient le résultat parfait que l'on doit obtenir par cette discrimination.

Les données étudiées dans ce paragraphe ne sont pas associées à une distribution Gaussienne.

- Appliquer la méthode de discrimination paramétrique et *Kppv* (pour $k = 1, 3, 5, 7, 13, 15$). Pour chacun des tests calculer le taux d'erreur.
- Commenter les résultats.