

Rapport projet Discrimination

Dalila HANOUTI - Zakaridia DIAWARA

Le projet a été réalisé en utilisant le logiciel **octave**

I. Principe de la règle paramétrique

Étape d'initialisation

données en entrée : ensemble d'apprentissage et classification de chaque individu

données en sortie : vecteur moyen et matrice de variance – covariance pour chaque classe

Instruction pour l'initialisation des données

Pour chaque classe c

Calculer le vecteur moyen de c .

Calculer la matrice de variance – covariance de c .

Classification des individus

données en entrée :

Ensemble contenant vecteur moyen et matrice de variance – covariance chaque classe

Probabilité de choix de chaque classe.

Ensemble des individus à classer.

données en sortie : vecteur contenant la classe de chaque individu.

Application de la décision de Bayes

Pour chaque individu à classer

Évaluer pour chaque classe la règle de décision en utilisant la distribution gaussienne.

Affecter l'individu à la classe pour laquelle cette valeur est maximale.

II. Règle paramétrique Vs Kppv

Les codes sources sont situés dans le répertoire **Source**.

L'algorithme des kppv a été codé dans le fichier `decision_kppv.m`. Pour tester cet algorithme, on lance la fonction `kppv(filename, k)`, où `filename` est le nom du fichier contenant les données (`td3_d1.mat`, `td3_d2.mat` etc) et `k` est le nombre de voisins à utiliser.

L'algorithme de la discrimination paramétrique a été codé dans le fichier `decision_bayes.m`. Pour tester cet algorithme, on lance la fonction `discrimination_parametrique(filename)`, où `filename` est le nom du fichier contenant les données (`td3_d1.mat`, `td3_d2.mat` etc).

Pour chacune des fonctions exécuter, un taux d'erreur de classification est renvoyé ainsi qu'un affichage graphique du résultat de la classification.

1. Comparaison standard

- **Kppv**

Nombre de voisin	1	3	5	7	13	15
Taux d'erreur	18	13	13	8	6	6

On obtient en moyenne **10 erreurs**.

On remarque que le taux d'erreur diminue en fonction du nombre de voisin. Plus le nombre de voisin augmente, moins il y a d'erreur.

Pour obtenir une meilleure estimation, on choisit donc un nombre de voisin grand.

- **Discrimination paramétrique**

En appliquant l'algorithme de discrimination paramétrique, on obtient **8 erreurs**.

En utilisant un nombre de voisin assez grand pour l'algorithme de kppv, on constate que les deux algorithmes génèrent sensiblement le même taux d'erreur.

2. Absence de professeur

- **Kppv**

Nombre de voisin	1	3	5	7	13	15
Taux d'erreur	20	18	15	15	16	17

On obtient en moyenne **16 erreurs**.

- **Discrimination paramétrique**

En appliquant l'algorithme de discrimination paramétrique, on obtient **12 erreurs**.

Ces taux élevés d'erreurs sont dû aux erreurs de classification générés par l'application de l'algorithme de coalescence sur l'ensemble d'apprentissage.

En effet, les algorithmes des kppv et de la discrimination paramétrique s'appuient sur l'ensemble d'apprentissage pour classer un individu. Un taux d'erreur élevé de classification de l'ensemble d'apprentissage entraînerait une forte erreur dans la classification des individus.

3. Influence de la taille de l'ensemble d'apprentissage : taille réduite

- **Kppv**

Nombre de voisin	1	3	5	7	13	15
Taux d'erreur	11	9	7	8	11	11

On obtient en moyenne **10 erreurs**.

Sur un ensemble d'apprentissage réduit, on note un taux d'erreur élevé généré par l'application l'algorithme de kppv. Ceci s'explique par une faible taille de l'ensemble d'apprentissage. En effet, l'algorithme des kppv est à appliquer sur les ensembles d'apprentissage de très grand taille.

- **Discrimination paramétrique**

En appliquant l'algorithme de discrimination paramétrique, on obtient **5 erreurs**.

Nous pouvons en déduire que sur un ensemble d'apprentissage réduit, l'algorithme de discrimination paramétrique est meilleur en terme du taux d'erreur que l'algorithme des kppv.

4. Influence de la taille de l'ensemble d'apprentissage : taille importante

- **Kppv**

Nombre de voisin	1	3	5	7	13	15
Taux d'erreur	6	5	4	3	5	4

Nous obtenons en moyenne **5 erreurs** par l'algorithme des kppv

Nous remarquons l'application de l'algorithme des kppv sur un très grand ensemble d'apprentissage génère moins d'erreurs que sur un faible ensemble d'apprentissage.

- **Discrimination paramétrique**

En appliquant l'algorithme de discrimination paramétrique, on obtient **4 erreurs**.

Nous pouvons conclure que sur un ensemble grand ensemble d'apprentissage, le taux d'erreur généré par des deux algorithmes est quasiment le même. Donc les deux algorithmes ont les mêmes performances sur cet test.

5. Distribution inconnue

- **Kppv**

Nombre de voisin	1	3	5	7	13	15
Taux d'erreur	5	5	5	3	9	6

Nous obtenons en moyenne **6 erreurs** par l'algorithme des kppv.

Le fait de ne pas utiliser de distribution gaussienne sur les données n'a pas d'influence sur l'algorithme des kppv car dans cet algorithme, on a fait implicitement une estimation de la fonction densité à partir de l'ensemble d'apprentissage. Ce qui explique qu'on obtient moins erreur dans l'application des kppv dans cette expérience

- **Discrimination paramétrique**

En appliquant l'algorithme de discrimination paramétrique, on obtient **6 erreurs**.

L'algorithme de la discrimination paramétrique utilise une distribution gaussienne comme fonction densité. En tant donné que dans cet test on utilise pas de distribution gaussienne sur les données, ceci devrait augmenter nos taux d'erreurs mais nous observons un taux d'erreur raisonnable.