Apprentissage Automatique Numérique

Rapport de TP

Classifieur Knn

M1 ISI

Khalid Al-kassoum Houssam

But :

Le but de ce TP est de créer un classifier Knn capable de classer 3 espèces d'iris dans leur classe respective selon certain critères . Dans ce TP nous disposons respectivement de 4 caractères (longueur pétale et sépale , largeur pétale et sépale), et nous faisons la classification en utilisant deux caractères à chaque fois .

1- Echantillonnage :

Dans cette partie du TP nous divisons l'ensemble des échantillons en 3 groupes (échantillons d'apprentissage , de développement et de test) après avoir pris le soins de mélanger les échantillons . Les échantillons d'apprentissage et de développement nous servent pour fixer les paramètres optimaux k(nombre de voisins pris en compte pour le vote) et D(le type de distance euclidienne ou Manhattan utiliser pour calculer la proximité des éléments) , puis ceux de test nous permettent de tester les performances de notre classifieur .

2- Développement :

Ici nous nous utilisons tout les échantillons d'apprentissage pour classer les échantillons de développement en faisant varier k et en comparent les résultât obtenu .

longueur pétale = 1
longueur sépale = 2
largeur pétale = 3
largeur sépale = 4

Tableau de comparaison des taux de bonne classification(Distance euclidienne)

caractères ——— K	1 et 2	1 et 3	1 et 4	2 et 3	2 et 4	3 et 4
1	0.8	0.93	0.93	0.9	0.9	0.9
5	0.9	0.93	0.93	0.9	0.86	0.9
10	0.8	0.93	0.93	0.9	0.86	0.93
15	0.8	0.93	0.9	0.93	0.86	0.86
20	0.8	0.9	0.9	0.93	0.9	0.9
25	0.83	0.9	0.9	0.93	0.93	0.9
30	0.83	0.86	0.83	0.93	0.93	0.9
35	0.8	0.86	0.83	0.93	0.9	0.9
40	0.83	0.83	0.83	0.9	0.93	0.9

Tableau de comparaison des taux de bonne classification(Distance Manhattan)

caractères ——— K	1 et 2	1 et 3	1 et 4	2 et 3	2 et 4	3 et 4
1	0.7	0.85	0.9	0.9	0.9	0.9
5	0.6	0.85	0.9	0.85	0.95	0.9
10	0.65	0.8	0.9	0.85	0.95	0.9
15	0.75	0.9	0.9	0.85	0.95	0.9
20	0.75	0.85	0.85	0.85	0.95	0.9
25	0.75	0.85	0.85	0.85	0.95	0.9
30	0.8	0.85	0.85	0.85	0.9	0.9
35	0.8	0.85	0.85	0.9	0.9	0.9
40	0.8	0.75	0.8	0.85	0.85	0.9

3- Evaluation(test) :

Après comparaison des différentes variantes du classifieur , le meilleur système est celui utilisant les caractéristiques 2 et 4 (longueur sépale, largeur sépale) et utilisant la méthode Manhattan pour le calcule de la distance , avec un k =15 , on a un taux de bonne classification de 95%

On obtient la matrice de confusion suivante sur les données de test :

Matrice de confusion

CLASSE	0	1	2
0	6	0	0
1	0	5	0
2	0	1	8

bonne classification 19
mauvaise classification 1
taux de bonne classification = 0,95

La seule erreur porte sur un élément de classe 2 et qui a été identifié comme élément de classe 1

4- Comparaison avec Bayes :

Avec le classifieur Bayes on avais un taux de bonne classification de 0,85 sur les données de test en utilisant la meilleur version , alors qu'avec le classifieur Knn on a un taux de bonne classification de 0,95 . Mais on peut noté que la puissance de calcule demander n'est pas meme pour les deux classifieur avec le Bayes on parcourt une seule fois les données d'entrainement et après on a la possibilité de classer un nouvelle element . alors qu'avec le Knn on est obligé de parcourir tout les éléments du corpus d'entrainement a chaque fois que l'on veux classer un nouvelle element , cela peux devenir problématique si l'on a un grand corpus d'apprentissage .