Exercice 01

Classifieur euclidien, K plus proches voisins

Jeux de données Synth1

**Question 01 Estimer les paramètres des distributions conditionnelles**

D’abord, nous avons un jeu de données (X est une matrice n par p, n est nombre d’individus et p est le nombre de variable, p= 2 dans ce cas). Et puis, nous avons un vecteur , qui indique la classe de chaque individu. Dans nos données, il y a deux groupes. Nous supposons que le nombre d’individus de chaque groupe est

Nous savons à priori que les jeux des données sont générés par un modèle gaussien mélangé (2 groupes et 2 dimensions). Nous supposons que les proportions sont , et puis pour chaque groupe, son espérance est et sa matrice de variance est .

D’après l’estimation du maximum vraisemblance, nous avons des estimateurs suivants :

D’ après cela, nous calculons des estimateurs pour chaque jeu de données.

**Synth1-40**

**Synth1-100**

**Synth1-500**

**Synth1-1000**

D’après les résultat des estimations, nous trouvons que

Comme , nous pouvons imaginer que un modèle de l’analyse discriminante linéaire suit bien à ces jeux de données.

**Question 02 Estimer le taux d’erreur avec le classifieur distance euclidienne**

**Question A.**

Selon l’énoncé,

Nous supposons que , qui suit évidemment la loi de Bernoulli. Si nous supposons que la probabilité d’erreur d’un point ici est p, nous savons que , où est le taux d’erreur.

Ensuite, nous supposons que , comme les variable suivons la loi de Bernoulli, , donc X suit la loi binomiale, nous avons

Comme E suit également la loi binomiale. Nous supposons que son espérance est , sa variance est . D’après la loi binomiale, ,

Quand m est grand, selon le théorème de la limite centrale, E suit approximativement la loi normale. Nous avons

Maintenant, nous avons un ensemble d’échantillon de E, c’est .

Comme , l’intervalle de confiance sur est la même de celle sur , le problème devient calculer l’intervalle de confiance sur l’espérance d’une échantillon qui suit la loi normale avec la variance inconnue.

Nous supposons que la moyenne d’échantillon est , et sa variance est , nous avons donc , qui suit la loi de Student.

Nous pouvons obtenir une intervalle de confiance bilatérale de

**Question B**

Pour chaque jeu de données, nous avons calculé l’estimation de taux d’erreur et l’intervalle de confiance sur l’ensemble d’apprentissage et l’ensemble de teste.

Pour le niveau de confiance, nous prenons que

**Synth1-40**

L’ensemble d’apprentissage

L’ensemble de teste

**Synth1-100**

L’ensemble d’apprentissage

L’ensemble de teste

0.179419

**Synth1-500**

L’ensemble d’apprentissage

L’ensemble de teste

0.118862

**Synth1-1000**

L’ensemble d’apprentissage

L’ensemble de teste

D’abord, quand le jeu de données est plus grand, l’estimation du taux d’erreur fait plus faible.

Quand le nombre de données est 1000, nous obtenons une taux d’erreur de 10% environ pour l’ensemble d’apprentissage et 11% pour

ceux de teste.

Généralement, l’estimation du taux d’erreur d’apprentissage sont la même comme la taux d’erreur de teste pour le classifieur distance euclidienne.

**Question 03 Déterminer le nombre d’optimal du classifieur le plus proche voisin.**

Nous prenons le jeu de données de 40 comme un exemple, et nous prenons le nombre des voisins k de 1 jusqu’à 39.

Nous faisons le test 10 fois, et tous les 10 fois, le meilleur k vaut 1.

Le résultat est évident, car nous prenons l’ensemble d’apprentissage comme l’ensemble de validation. Si on choisit 1 voisin le plus proche, c’est forcément le points soi-même. Du coup, la classification fait toujours correct et le taux d’erreur vaut 0. Par conséquent, le plus optimal k vaut 1.

**Question 04** **Estimer le taux d’erreur avec le classifieur k le plus proche voisin**

Comme l’exercice précédent, pour chaque jeu de données, nous avons calculé l’estimation de taux d’erreur et l’intervalle de confiance sur l’ensemble d’apprentissage et l’ensemble de teste.

Pour le niveau de confiance, nous prenons que

**Synth1-40**

L’ensemble d’apprentissage

0.08

L’ensemble de teste

0.13

**Synth1-100**

L’ensemble d’apprentissage

0.081

L’ensemble de teste

**Synth1-500**

L’ensemble d’apprentissage

0.0694

L’ensemble de teste

0.0844

**Synth1-1000**

L’ensemble d’apprentissage

0.0557

L’ensemble de teste

D’abord, nous observons également que quand le jeu de données est plus grand, l’estimation du taux d’erreur pour l’ensemble d’apprentissage et celle de l’ensemble de teste sont devenue plus faible.

En comparant avec celles obtenue par le classifieur de distance euclidienne, nous constatons que :

Comme le classifieur le plus proche voisin sont beaucoup compliqué que le classifieur distance euclidienne, il obtenu un résultat plus fiable.

Ensuite, nous observons également que

Pour le classifieur de distance euclidien, les résultat des deux ensemble sont presque la même. Cependant, pour le classifieur le plus proche voisin, le résultat de l’ensemble d’apprentissage sont plus fiable que celui de l’ensemble de teste.

La raison est que, pour le classifieur distance euclidien, nous calculons les distances des points avec les centres de gravité. Du coup, cela ne différencie pas trop pour l’ensemble d’apprentissage et l’ensemble de teste.

Par contre, dans le classifieur le plus proche voisin, nous recherchons les distances pour déterminer les k plus proches voisin. Du coup, pour l’ensemble d’apprentissage, il compare avec les points de lui-même, cela fait moins d’erreur. C’est pourquoi il est plus fiable que l’ensemble de teste.

Jeux de données Synth2

Ensuite, nous étudions le jeu de données Synth2.

**Question 01 Estimer les paramètres des distributions conditionnelles**

Comme l’exercice précédent, nous utilisons l’estimation du maximum vraisemblance à estimer des paramètres. Nous avons des résultats comme ci-dessous :

Nous pouvons estimer que :

**Question 02 Estimer les résultats avec les deux classifieurs**

*Classifieur de distance euclidienne*

Nous avons obtenu le résultat comme ci-dessous :

L’ensemble d’apprentissage

L’ensemble de teste

*Classifieur de plus proche voisin*

L’ensemble d’apprentissage

L’ensemble de teste

Avec ce jeu de données, nous obtenons les estimations des taux d’erreur beaucoup plus fiable.