SY09 - TP03
Discrimination, théorie bayésienne de la décision

Félix Flores - Cristian Garrido

28 mai 2015

1 Classifieur euclidien, K plus proches voisins

Pour chacun des jeux de données, on a estimé les paramètres μ_k et $\sum_k$ des distributions conditionnelles, ainsi que les proportions π_k des classes. On a alors obtenu les résultats suivants :

1. Pour le jeu de données *Synth1-40* :

$$\pi_1 = 0.55, \pi_2 = 0.45 \qquad \mu_1 = \begin{pmatrix} -1.904681 \\ 0.698841 \end{pmatrix}, \mu_2 = \begin{pmatrix} 0.8808594 \\ 0.8587835 \end{pmatrix}$$

$$\sum_1 = \begin{pmatrix} 0.7701334 & 0.2566214 \\ 0.2566214 & 1.0798685 \end{pmatrix}, \sum_2 = \begin{pmatrix} 1.41546550 & -0.01011499 \\ -0.01011499 & 1.22191696 \end{pmatrix}$$

2. Pour le jeu de données *Synth1-100* :

$$\pi_1 = 0.53, \pi_2 = 0.47 \qquad \mu_1 = \begin{pmatrix} -1.808233 \\ 1.057774 \end{pmatrix}, \mu_2 = \begin{pmatrix} 0.9831412 \\ 1.1610775 \end{pmatrix}$$

$$\sum_1 = \begin{pmatrix} 1.50880197 & 0.07097749 \\ 0.07097749 & 0.95833356 \end{pmatrix}, \sum_2 = \begin{pmatrix} 0.84677155 & 0.04545945 \\ 0.04545945 & 0.71631862 \end{pmatrix}$$

3. Pour le jeu de données *Synth1-500* :

$$\pi_1 = 0.48, \pi_2 = 0.52 \qquad \mu_1 = \begin{pmatrix} -1.9101695 \\ 0.9359349 \end{pmatrix}, \mu_2 = \begin{pmatrix} 0.9979001 \\ 0.9843533 \end{pmatrix}$$

$$\sum_1 = \begin{pmatrix} 1.16050554 & 0.03149239 \\ 0.03149239 & 0.87727376 \end{pmatrix}, \sum_2 = \begin{pmatrix} 0.91978725 & -0.03484383 \\ -0.03484383 & 0.95416768 \end{pmatrix}$$

4. Pour le jeu de données *Synth1-1000* :

$$\pi_1 = 0.49, \pi_2 = 0.51 \qquad \mu_1 = \begin{pmatrix} -1.998992 \\ 1.008186 \end{pmatrix}, \mu_2 = \begin{pmatrix} 1.0908900 \\ 0.9837324 \end{pmatrix}$$

$$\sum_1 = \begin{pmatrix} 1.03969441 & -0.07115634 \\ -0.07115634 & 0.97107695 \end{pmatrix}, \sum_2 = \begin{pmatrix} 1.00246869 & -0.02488749 \\ -0.02488749 & 1.02228782 \end{pmatrix}$$

On peut donc bien remarquer en tant que la taille de l'échantillon augmente les paramètres plus s'approchent à les valeurs suivantes :

$$\pi_1 \approx \pi_2 \approx 0.5, \mu_1 \approx \begin{pmatrix} -2 \\ 1 \end{pmatrix}, \mu_2 \approx \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \sum_1 \approx \sum_2 \approx \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Estimation du taux d'erreur

En ayant $E = \frac{1}{m}\sum 1_{z_i \neq z_i}$ alors on peut remarquer que $T_i = \begin{cases} 1 & z_i \neq z_i \\ 0 & sinon \end{cases}$ suit une loi de Bernouilli, donc $T_i \sim \beta(\varepsilon)$ où ε représente le taux d'erreur et

$$P(T_i = 1) = E(E) = E(\frac{1}{m}\sum_i^m T_i) = \frac{1}{m}E(\sum_i^m T_i) = m\sum_i^m E(T_1) = \frac{m\varepsilon}{m} = \varepsilon.$$

Puis on peut déduire que mE suit aussi une loi binomiale. Maintenant on peut approcher par une loi normale en supposant que m est assez grand. on a alors $mE \sim \mathcal{N}(m\varepsilon, m\varepsilon(1-\varepsilon))$. et grâce à cela on a $E \sim \mathcal{N}(\varepsilon, m^{-1}\varepsilon(1-\varepsilon))$ Pour connaître la variance, mais on saut que on a une loi de student E : $\frac{\bar{E}-\mu}{S^*/\sqrt{N}} \sim \tau_{N-1}$. On assume les taux d'erreur E_j sont indépendants. Alors On peut calculer l'intervalle de confiance pour l'espérance de E, en trouvant μ, $P(\frac{\bar{E}-\mu}{S^*/\sqrt{N}}) < 1-\alpha$ Finalement on obtient l'intervalle de confiance bilatérale suivante :

$$IC = [\bar{E} - t_{N-1;1-\alpha/2}\frac{S^*}{\sqrt{N}}, \bar{E} + t_{N-1;1-\alpha/2}\frac{S^*}{\sqrt{N}}]$$

En suite on a calculé les intervalles de confiance pour chaque jeu de donnés. En utilisant un taux d'erreur obtenu lors de vingt exécutions et avec un niveau de confiance de $1-\alpha = 0,95$:

1. Pour le jeu de données *Synth1-40* :
   — Ensemble d'apprentissage : $\varepsilon = 0.124074$ et $IC = [0.123785, 0.124363]$
   — Ensemble de test : $\varepsilon = 0.111539$ et $IC = [0.111250, 0.111827]$
2. Pour le jeu de données *Synth1-100* :
   — Ensemble d'apprentissage : $\varepsilon = 0.143939$ et $IC = [0.143753, 0.144126]$
   — Ensemble de test : $\varepsilon = 0.179419$ et $IC = [0.179225, 0.179598]$
3. Pour le jeu de données *Synth1-500* :
   — Ensemble d'apprentissage : $\varepsilon = 0.122372$ et $IC = [0.122337, 0.122407]$
   — Ensemble de test : $\varepsilon = 0.118862$ et $IC = [0.118827, 0.118897]$
4. Pour le jeu de données *Synth1-1000* :
   — Ensemble d'apprentissage : $\varepsilon = 0.104880$ et $IC = [0.104870, 0.104889]$
   — Ensemble de test : $\varepsilon = 0.1121257$ et $IC = [0.112116, 0.112135]$

On peut observer qu'en tant que la taille de l'échantillon augmente, l'estimation du taux d'erreur deviens au même temps plus faible. Particulièrement, dans le cas quand on a 100 individus, on a obtenu une taux d'erreur proche au dix et once percent pour l'ensemble d'apprentissage et l'ensemble de test respectivement. On peut remarquer que le taux d'erreur de l'ensemble de apprentissage et du ensemble de test est toujours approché.

Nombre optimal de voisins

Pour obtenir le nombre optimal de voisins, on a utilisé le jeu de données *Synth1-1000*, on a effectué une séparation aléatoire de l'ensemble de données en un ensemble d'apprentissage et ensemble de test ; et on a comme nombre de classes candidates les valeurs 1, 2, 3, 4 et 5. Puis on a fait le test 100 fois et on a obtenu 48 fois le nombre $K = 2$, 20 fois $K = 5$, 14 fois $K = 3$, 12 fois $K = 4$ et 6 fois $K = 1$. Alors on a une forte tendance au nombre de classes $K = 2$. Cette instabilité du résultat est du à l'aléatoire. Car en dépendant les centres choisis pour l'algorithme *kmeans* la classe à qui appartient ce centre peut varier et ceci fait changer aussi la classe à qui les points de validations appartient. Par conséquent, on peut bien choisir comme nombre de classe optimal $K = 2$.

Maintent, comme dans l'exercise précedent. on a calcule pou chaque jeu de donnés, l'estimation des taux d'erreurs et les intervalles de confiance pou l'ensamble d'apprentissage et l'ensemble de test avec un niveau de confiance de $1-\alpha = 0,95$:

1. Pour le jeu de données *Synth1-40* :
   — Ensemble d'apprentissage : $\varepsilon = 0.082$ et $IC = [0.07826, 0.081774]$

— Ensemble de test : $\varepsilon = 0.132$ et $IC = [0.128227, 0.131776]$

2. Pour le jeu de données *Synth1-100* :
   — Ensemble d'apprentissage : $\varepsilon = 0.081$ et $IC = [0.080616, 0.0813836]$
   — Ensemble de test : $\varepsilon = 0.11$ et $IC = [0.109616, 0.110384]$
3. Pour le jeu de données *Synth1-500* :
   — Ensemble d'apprentissage : $\varepsilon = 0.0694$ et $IC = [0.069330, 0.069460]$
   — Ensemble de test : $\varepsilon = 0.0844$ et $IC = [0.084339, 0.084461]$
4. Pour le jeu de données *Synth1-1000* :
   — Ensemble d'apprentissage : $\varepsilon = 0.0557$ et $IC = [0.055672, 0.055728]$
   — Ensemble de test : $\varepsilon = 0.0736$ et $IC = [0.073572, 0.073628]$

On peut encore remarquer, comme dans l'exercice précédent, que en tant que la taille de l'échantillon est plus grand la différence entre le taux d'erreur pour l'ensemble d'apprentissage et celle de l'ensemble est à chaque fois plus faible.

Jeu de données Synth2-1000

Comme l'exercice précédent, on utilise l'estimation du maximum vraisemblance pour estimer les paramètres μ_k et $\sum_k$, ainsi que les proportions π_k des classes. On a alors obtenu les résultats suivants :

$$\pi_1 = 0.488, \pi_2 = 0.512 \qquad \mu_1 = \begin{pmatrix} -4.055942 \\ 1.011496 \end{pmatrix}, \mu_2 = \begin{pmatrix} 4.028957 \\ 1.067821 \end{pmatrix}$$

$$\sum_1 = \begin{pmatrix} 1.014560 & 0.018451 \\ 0.018451 & 0.939754 \end{pmatrix}, \sum_2 = \begin{pmatrix} 4.953398 & 0.107909 \\ 0.107909 & 5.022125 \end{pmatrix}$$

En approximant ces valeurs on peut les estimer comme :

$$\pi_1 \approx \pi_2 \approx 0.5, \mu_1 \approx \begin{pmatrix} -4 \\ 1 \end{pmatrix}, \mu_2 \approx \begin{pmatrix} 4 \\ 1 \end{pmatrix}, \sum_1 \approx \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \sum_2 \approx \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}$$

.

Maintenant on calcule l'estimation des taux d'erreur avec les deux classificateurs.

1. Avec classificateur de distance euclidienne :
   — Ensemble d'apprentissage : $\varepsilon = 0.0063063$ et $IC = [0.006305, 0.006308]$
   — Ensemble de test : $\varepsilon = 0.006587$ et $IC = [0.006585, 0.006588]$
2. Avec classificateur de plus proche voisin :
   — Ensemble d'apprentissage : $\varepsilon = 0.0047$ et $IC = [0.004697, 0.004703]$
   — Ensemble de test : $\varepsilon = 0.0066$ et $IC = [0.006597, 0.006603]$

Après ce calcule, on peut noter que les estimations obtenus sont beaucoup plis fiables. Cela principalement du à la grande quantité de donnés et surtout parce que les variances entre les classes son beaucoup plus distantes par rapport aux jeux de donnés précédant. Ce fait alors avoir des centres plus éloignés et donc par le deux classificateurs un tirage plus forte par les classes.