SORBONNE UNIVERSITÉS

Examen d'apprentissage artificiel
Année universitaire 2019-2020

Nom:	Formation:
Prénom:	

Exercice 1 (4pts) - Réduction de dimension

On dispose d'un ensemble de données de dimension 2 appartenant à deux classes représentées par des symboles différents sur la Figure . On souhaite représenter ces données par une seule dimension.

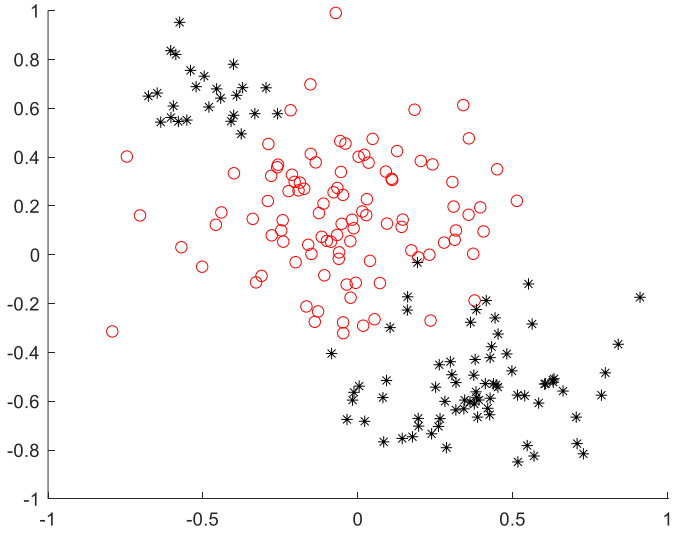


Figure - Exemples des deux classes

On effectue dans un premier temps une analyse en composantes principales et pour cela, on doit estimer la matrice de covariance du nuage de points. Parmi ces matrices, laquelle correspond aux données de la Figure ? Pourquoi ?

$\Sigma 1 = \begin{bmatrix} 0.13 & 0.11 \\ 0.11 & 0.21 \end{bmatrix}$ $\Sigma 2 = \begin{bmatrix} 0.13 & 0.01 \\ 0.01 & 0.21 \end{bmatrix}$ $\Sigma 3 = \begin{bmatrix} 0.13 & -0.11 \\ -0.11 & 0.21 \end{bmatrix}$ $\Sigma 4 = \begin{bmatrix} 0.13 & -0.001 \\ -0.001 & 0.21 \end{bmatrix}$

$\Sigma 5 = \begin{bmatrix} 0.13 & -0.11 \\ -0.11 & 3 \end{bmatrix}$ $\Sigma 6 = \begin{bmatrix} 0.13 & 0 \\ 0 & 0.14 \end{bmatrix}$

Les valeurs propres/vecteurs propres de la matrice sont donnés par :
$\lambda_1 = 0.05,\ \ \lambda_2 = 0.30,\quad v_1 = (-0.81 \quad -0.58)^T,\quad v_2 = (-0.58 \quad 0.81)^T$
Représenter sur la Figure l'axe sur lequel les données seront projetées. Expliquer votre raisonnement. Conclusion sur la pertinence de la méthode.

Donner la coordonnée du point $(-0.3 \quad 0.4)^T$ une fois projeté sur l'axe principal (poser le calcul)

On essaie maintenant une autre méthode de réduction de dimension : l'analyse discriminante linéaire. Pour cela, nous avons besoin de connaitre les matrices de covariance intra-classe et inter-classe. On donne ci-dessous les matrices de covariance de chaque classe et la matrice de covariance inter-classe. Déterminer qui est qui. Justifier votre réponse.

$$\Sigma 1 = \begin{bmatrix} 0.18 & -0.19 \\ -0.19 & 0.29 \end{bmatrix} \quad \Sigma 2 = \begin{bmatrix} 0.06 & 0.00 \\ 0.00 & 0.06 \end{bmatrix} \quad \Sigma 3 = \begin{bmatrix} 0.02 & -0.04 \\ -0.04 & 0.08 \end{bmatrix}$$

On donne ci-dessous les valeurs/vecteurs propre de $\Sigma_{intra}^{-1}\Sigma_{inter}$.
$\lambda_1 = 0.00,\ \ \lambda_2 = 0.46,\quad v_1 = (-0.88 \quad -0.45)^T,\quad v_2 = (-0.07 \quad -0.99)^T$
Représenter sur la Figure l'axe sur lequel les données seront projetées. Expliquer votre raisonnement. Conclusion sur la pertinence de la méthode.

Exercice 2 (3pts)– kppv

Considérons un problème de classification à deux classes. La base de référence est représentée figure 1.

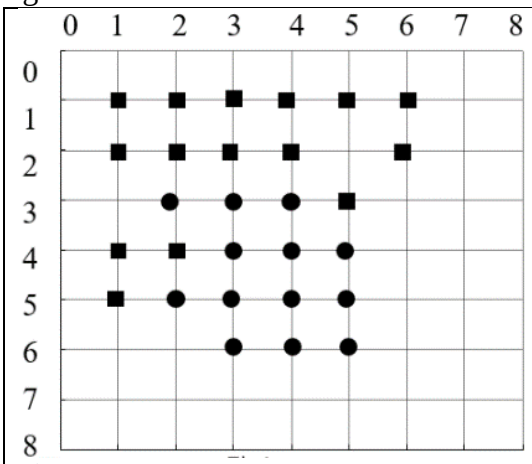


Figure 1

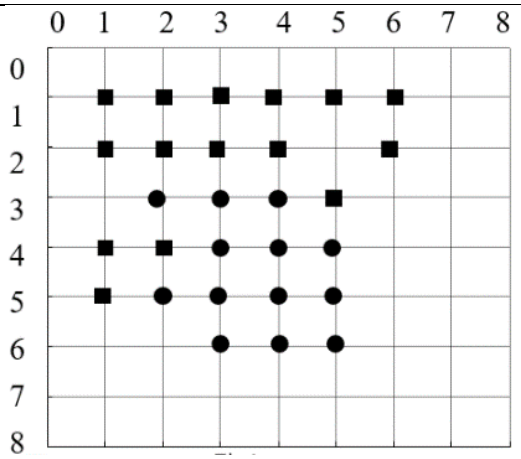


Figure 2

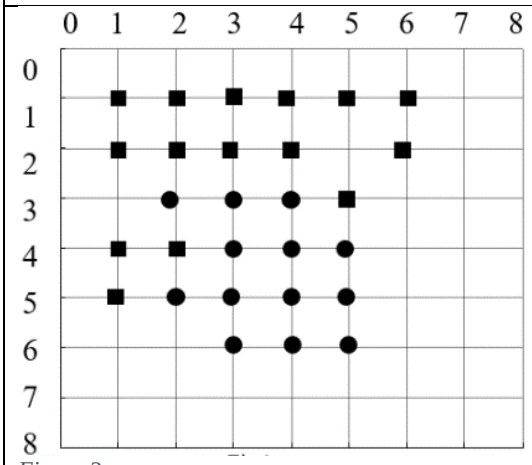


Figure 3

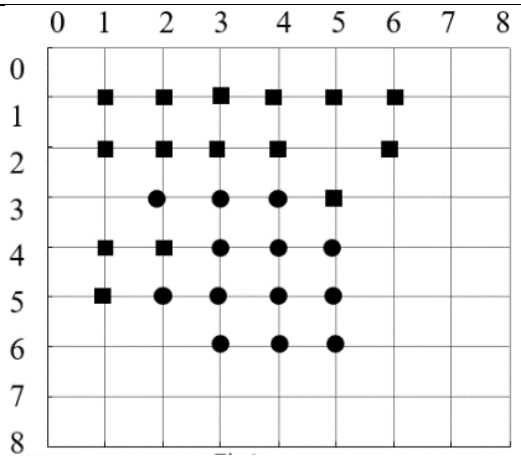


Figure 4

1. Tracer Figure 1 les frontières de décision obtenues avec l'algorithme du 1ppv.
2. On ajoute un point aberrant en (6,6) que l'on gardera jusqu'à la fin de l'exercice. Retracer Figure 2 les nouvelles frontières de décision.
3. Tracer approximativement Figure 3 les frontières de décision obtenues avec k=3. Conclusion ?

4. Comment sont les frontières de décision avec les 28-ppv ?

5. Tracer grossièrement Figure 4 les frontières de décision avec l'algorithme nearest-mean.

Exercice 3 (3 pts) – Distance d'édition

Un codage de forme est basé sur le codage des orientations de contour par un codage de Freeman à 4 états. On dispose de deux modèles de forme:

forme 1 : 01133

forme 2 : 03211

Un nouvel objet apparait et est codé : 013. On souhaite le classer en tant que forme 1 ou forme 2 en utilisant l'algorithme du plus proche voisin et la distance d'édition (on pourra aussi faire du rejet si égalité des distances).

1. Dans un premier temps, afin d'être entièrement invariant en changement d'échelle, on décide d'utiliser des coûts d'insertion et de suppression nuls et des coûts de changement égaux à 1 (nuls si même symbole). Calculer les distances ainsi obtenues et classer l'objet inconnu. Conclusion ?

2. Dans un second temps, on décide de changer les coûts d'insertion/suppression et de les fixer à 0.5 (on gardera les mêmes coûts de changement qu'à la question 1). Pourquoi ? Calculer de nouveau les distances et réaliser la classification.

3. Enfin, on décide de fixer les coûts d'insertion/suppression à 0.5, mais de modifier les coûts de changement en fonction des changements réalisés. Proposer une nouvelle matrice des coûts de changement en justifiant les nouvelles valeurs (inutile de recalculer les distances).

Exercice 4 (5 pts) – Classification bayésienne

On considère un problème de classification à K classes y_k où chaque classe est caractérisée par sa probabilité P(y_k) et une densité de probabilité conditionnelle :

$$p(\boldsymbol{x}/y_k) = a * \exp(-(\boldsymbol{x} - \boldsymbol{\mu_k})^2)$$

où les exemples $\boldsymbol{x} \in \mathbb{R}^n$.

1. Déterminer la valeur de la constante *a* pour que $p(\boldsymbol{x}/y_k)$ soit bien une densité de probabilité (on supposera connu que la loi normale est une densité de probabilité)

2. On dispose, pour chaque classe de N exemples $\boldsymbol{x_i}$ de dimension *n*. Déterminer l'expression de $\boldsymbol{\mu_k}$ en utilisant le maximum de vraisemblance (détailler les calculs).

3. On souhaite classer les données avec la théorie bayésienne de la décision. Le coût d'une bonne décision est 0 et d'une mauvaise décision α. Donner l'expression des risques conditionnels $R(y_k/\boldsymbol{x})$.

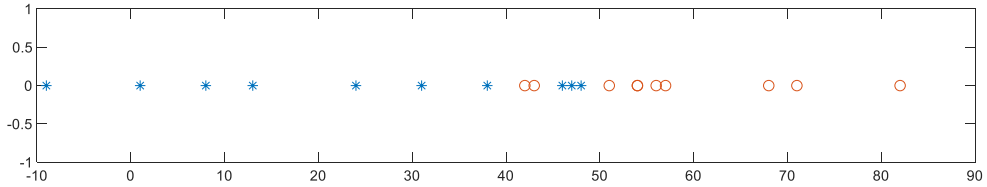
4. Montrer que le risque minimum est obtenu en décidant y_k si $P(y_k/\boldsymbol{x}) > P(y_j/\boldsymbol{x})\ \forall j \neq k$.

5. Expliciter les fonctions de décision dans le cas où K=2 et $P(y_1) = 1/3$.

Exercice 5 (5 pts)– Classification bayésienne

On souhaite déterminer si un patient est malade ou non à partir d'un facteur Rx déterminé à l'aide d'une prise de sang. Les valeurs de Rx, pour les populations saines et malades, sont données dans le tableau ci-dessous et sont représentées la Figure.

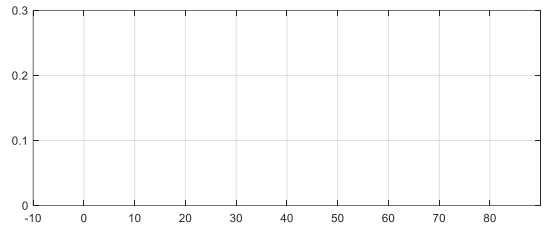
Sain	31	-9	1	24	13	46	47	8	38	48
Malade	56	82	71	54	68	54	42	57	51	43

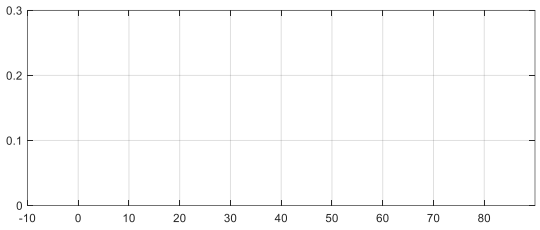


1. Classer un patient avec un $Rx = 45$ avec la règle du plus proche voisin puis des 3 plus proches voisins en détaillant votre raisonnement.

2. On souhaite maintenant utiliser un arbre de décision pour réaliser la classification. Sans faire de calcul, donner l'arbre de décision sur ce jeu de données permettant de classer correctement chaque exemple d'apprentissage. Justifier votre démarche puis réaliser la classification du patient avec $Rx = 45$.

3. On souhaite modéliser $p(Rx/Sain)$ et $p(Rx/Malade)$ à l'aide d'histogrammes où l'origine est fixée à 0 et le pas de discrétisation est 10. Représenter ces deux densités de probabilité sur les Figures ci-dessous (attention à la normalisation). Que valent-elles en $Rx = 45$? Réaliser la classification de ce patient en sachant que $P(malade) = 0.05$ en utilisant la règle de classification du maximum *a posteriori*.





4. On modélise maintenant $P(Rx/Sain)$ et $P(Rx/Malade)$ avec la méthode des noyaux en utilisant le noyau $K(x)$. Déterminer la valeur de ces deux densités de probabilités pour $Rx = 45$ en détaillant vos calculs (on prendra attention au terme de normalisation pour que les densités de probabilité somment à 1). Réaliser la classification avec $P(malade) = 0.05$ et la règle de classification du maximum *a posteriori*.