Exercices en apprentissage supervisé et non supervisé

M2 SISE - Université Lyon 2 - 2018/2019

Responsable : Julien Ah-Pine

Exercice 1

Soit un problème de catégorisation binaire dont l'ensemble d'apprentissage est composé d'environ 59 individus représentés dans $\mathbb{R}^2$. Nous représentons ci-derrière les deux classes respectivement par des disques rouges et des losanges noirs. Nous disposons de 29 individus de la classe 1 et de 30 individus de la classe 2. La figure 1 représente les régions de prédiction de la méthode des plus proches voisins avec comme paramètre k_1 tandis que la figure 2 utilise la même méthode d'apprentissage mais avec un paramètre différent k_2.

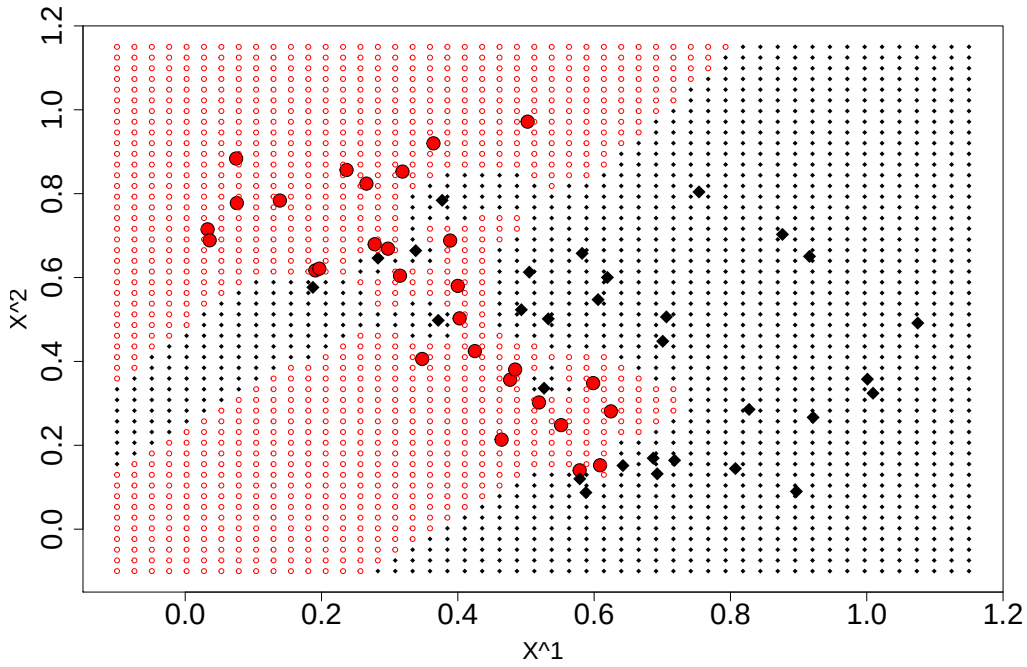


FIGURE 1 – Prédiction avec k_1 plus proches voisins

Questions :

Q1 Déterminez les valeurs de k_1 et de k_2 en supposant que la distance utilisée est la distance euclidienne ? (Indications : k_1 et k_2 sont impaires et compris entre 1 et 7.)

Q2 Déterminez dans chacun des deux cas, le taux d'erreur [1] d'apprentissage.

Q3 Lequel de ces deux modèles a, selon vous, la plus faible variance ?

Q4 Discutez de l'objectif poursuivi en apprentissage automatique en vous appuyant sur les concepts de biais d'apprentissage et de variance.

Q5 Lequel de ces deux modèles choisiriez vous dans le cas pratique étudié ici ?

1. ou une approximation du taux d'erreur

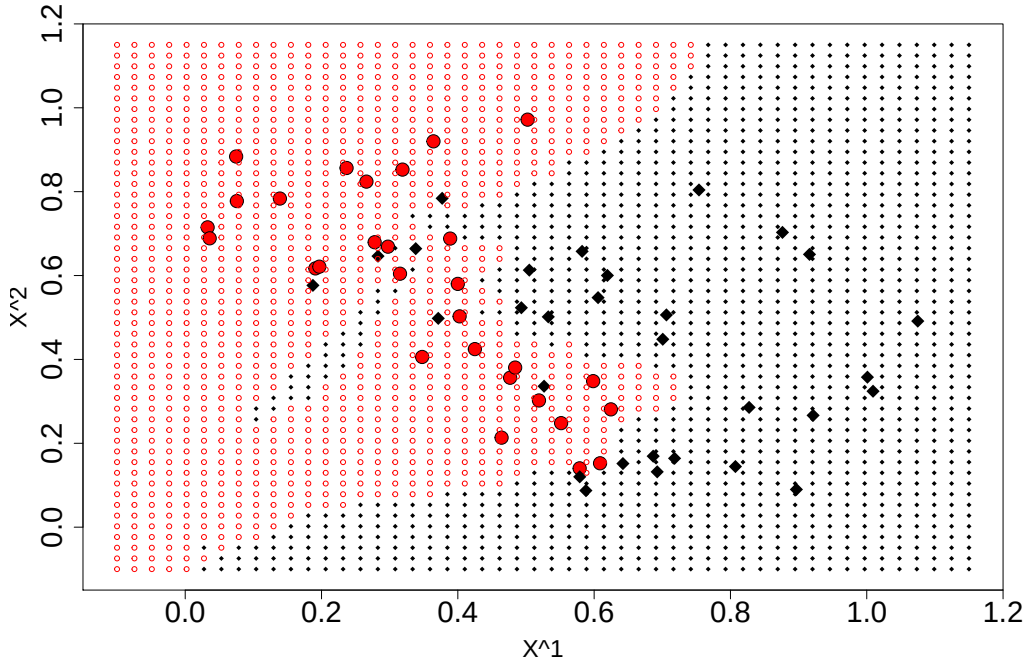


FIGURE 2 – Prédiction avec k_2 plus proches voisins

Exercice 2

Soit les données discrètes suivantes correspondant à 7 observations de quatre variables discrètes X^1, X^2, X^3, Y :

$$\mathbf{X} = \begin{array}{c} \\ \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \\ \mathbf{x}_4 \\ \mathbf{x}_5 \\ \mathbf{x}_6 \\ \mathbf{x}_7 \end{array} \begin{array}{c} \begin{array}{ccc} \mathbf{x}^1 & \mathbf{x}^2 & \mathbf{x}^3 \end{array} \\ \begin{pmatrix} A & \alpha & 1 \\ A & \beta & 1 \\ A & \alpha & 1 \\ B & \alpha & 3 \\ B & \alpha & 1 \\ C & \beta & 2 \\ C & \beta & 2 \end{pmatrix} \end{array} \quad ; \quad \mathbf{y} = \begin{pmatrix} C_1 \\ C_1 \\ C_1 \\ C_1 \\ C_2 \\ C_2 \\ C_2 \end{pmatrix}$$

Questions :

Q1 Appliquez l'algorithme du classifieur bayésien naïf à ce problème de catégorisation binaire. Pour cela vous calculerez les probabilités suivantes : $P(Y = C_1), P(Y = C_2)$ et pour chaque variable X^j et chacune de ses modalités m_k^j les probabilités $P(X^j = m_k^j | C_1)$ et $P(X^j = m_k | C_2)$.

Q2 Soit la distance suivante définie entre deux vecteurs discrets :

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{l=1}^{3} ind(x_{il} \neq x_{jl})$$

où x_{il} est le terme général de la matrice $\mathbf{X}$ et $ind(A) = 1$ si la proposition A est vraie et 0 sinon. Déterminer la prédiction obtenue pour chacune des 3 nouvelles observations suivantes par la méthode des k plus proches voisins avec $k = 3$ (explicitez vos calculs).

$$\begin{array}{c} \\ \mathbf{x}_8 \\ \mathbf{x}_9 \\ \mathbf{x}_{10} \end{array} \begin{array}{c} \begin{array}{ccc} X^1 & X^2 & X^3 \end{array} \\ \begin{pmatrix} A & \alpha & 2 \\ C & \beta & 1 \\ B & \beta & 1 \end{pmatrix} \end{array}$$

Q3 Déterminez la prédiction obtenue par chacune la méthode étudiée en Q1 pour les 3 nouvelles observations de Q3 (explicitez vos calculs). Commentez les résultats obtenus.

Exercice 3

Soit les 3 observations suivantes :

$\mathbf{x}^1$	1	0	0
$\mathbf{x}^2$	$1/\sqrt{3}$	$2/\sqrt{3}$	0
$\mathbf{y}$	1.5	0.5	1

Soit le modèle de régression multiple sans constante suivant :

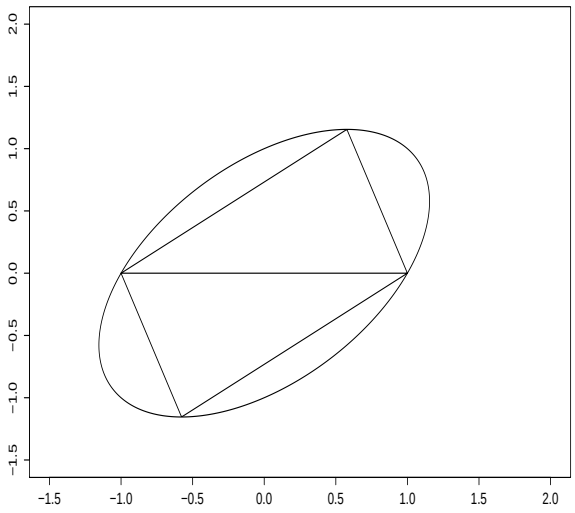
$$\mathbf{y} = \beta_1 \mathbf{x}^1 + \beta_2 \mathbf{x}^2 + \epsilon$$

Les régressions Ridge, Lasso seront effectuées sur les données non centrées et non réduites. L'objectif de cet exercice est d'appréhender la géométrie de ces méthodes.

Q1 Que vaut p, la dimension de l'espace de représentation des données ? Représenter dans $\mathbb{R}^p$ l'ensemble $\mathbb{B}_1$ des $\beta \in \mathbb{R}^p$ vérifiant la contrainte $\|\beta\|^2 = 1$ (norme l_2) et l'ensemble $\mathbb{B}_2$ des $\beta \in \mathbb{R}^p$ vérifiant la contrainte $\|\beta\|_{l_1} = 1$ (norme l_1).

Q2 Calculez l'estimation des MCO de β. Déterminez $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}_{mco}$.

Q3 La matrice des données $\mathbf{X} = \begin{pmatrix} \mathbf{x}^1 & \mathbf{x}^2 \end{pmatrix}$ peut être vue comme une application liénaire de $\mathbb{R}^2$ dans $\mathbb{R}^3$. On remarquera cependant que la troisième composante de l'image est toujours nulle si bien que la dimension de $Im(\mathbf{X})$ vaut 2. Soient $\mathbb{C}_1$ et $\mathbb{C}_2$ les ensembles images de $\mathbb{B}_1$ et $\mathbb{B}_2$ lorsqu'on leur applique $\mathbf{X}$. $\mathbb{C}_1$ et $\mathbb{C}_2$ sont définies par : $\mathbb{C}_1 = \{\mathbf{z} \in Im(\mathbf{X}), \exists \beta \in \mathbb{B}_1 : \mathbf{z} = \mathbf{X}\beta\}$ et $\mathbb{C}_2 = \{\mathbf{z} \in Im(\mathbf{X}), \exists \beta \in \mathbb{B}_2 : \mathbf{z} = \mathbf{X}\beta\}$. Soit la représentation suivante :



Identifiez $\mathbb{C}_1$ et $\mathbb{C}_2$ puis indiquez $\hat{\mathbf{y}}$ sur ce plan.

Q4 Représentez également sur ce plan $\mathbf{X}\hat{\beta}_{ridge}$ et $\mathbf{X}\hat{\beta}_{lasso}$ sachant que les estimations $\hat{\beta}_{ridge}$ et $\hat{\beta}_{lasso}$ sont les solutions de la minimisation de $\|\mathbf{y} - \mathbf{X}\beta\|^2$ sous la contrainte respective que $\beta \in \mathbb{B}_2$ et $\beta \in \mathbb{B}_1$.

Exercice 4

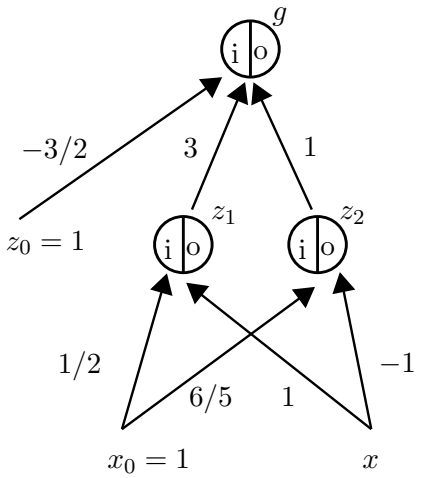
Nous avons un problème de catégorisation binaire et nos observations sont dans un espace réel unidimensionnel. Nos 8 observations sont décrites par la variable X et elles appartiennent aux classes

données dans la variable Y ci-dessous :

$$\mathbf{x} = \begin{pmatrix} -2 \\ -1.5 \\ -1 \\ -0.5 \\ 0 \\ 0.5 \\ 1 \\ 1.5 \end{pmatrix} \quad ; \quad \mathbf{y} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

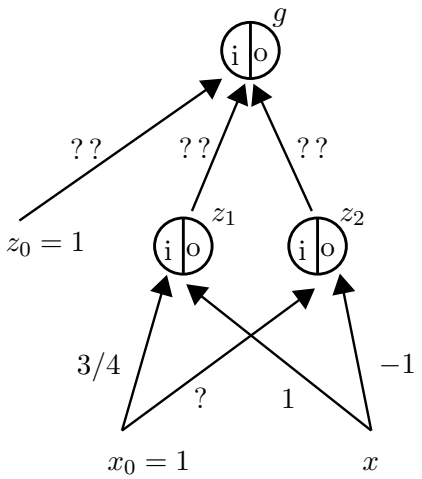
Q1- Représentez à l'aide d'une droite X ces observations. Est-ce que ces deux classes sont linéairement séparables dans cet espace unidimensionnel ?

Nous souhaitons utiliser un réseau de neurones multicouche afin de catégoriser nos observations. Soit alors le réseau de neurones multicouche défini par le schéma ci-dessous où les fonctions d'activation de la 1ère et de la 2ème couche sont des fonctions d'Heaviside.



Q2- Représentez dans le plan (Z^1, Z^2) les observations à l'issue de la projection obtenue à la sortie de la 1ère couche. Représentez également la frontière de décision du réseau de neuronnes. Donnez la matrice de confusion issue de ce modèle et calculez le taux d'erreur ?

Le modèle ci-dessus commettant plusieurs erreurs, nous souhaitons proposer un nouveau réseau de neuronnes avec les paramètres partiellement données dans le schéma ci-dessous où nous considérons que les fonctions d'activation des deux couches sont toujours des fonctions d'Heaviside.



Q3- Déterminez là où est indiqué un " ?" au niveau de la 1ère couche, un coefficient synaptique (il peut en avoir plusieurs) permettant d'obtenir une représentation des deux groupes linéairement séparables dans le plan (Z^1, Z^2).

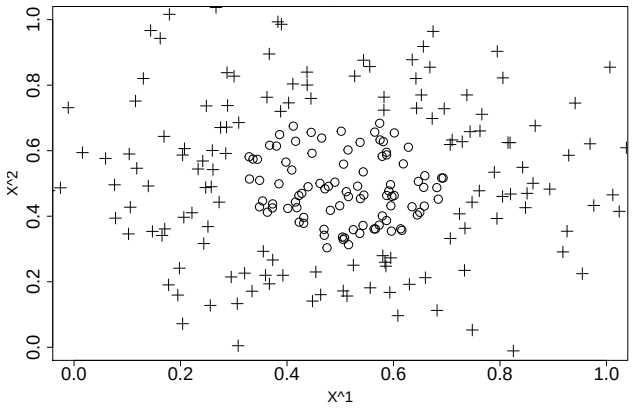
Q4- Précisez les coefficients synaptiques (il peut en avoir plusieurs) de la 2ème couche indiqués par un " ? ?", permettant d'obtenir un taux d'erreur nul.

Exercice 5

Questions de cours :

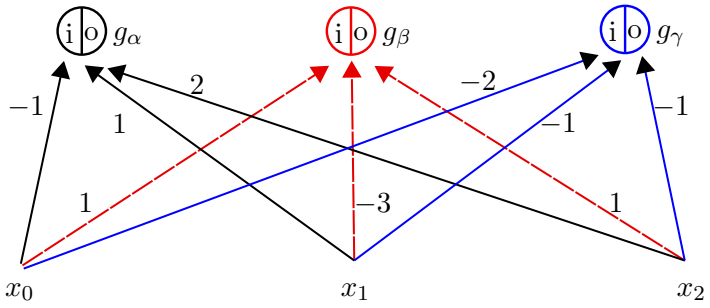
Q1- En quelques lignes, rappelez le principe des méthodes non paramétriques vues en cours et expliquez pourquoi celles-ci ne peuvent pas traiter de façon efficace les données en grande dimension.

Q2- Considérez le problème de catégorisation binaire dans la figure suivante où les données sont décrites dans l'espace engendré par X^1 et X^2. Proposez alors deux méthodes vues en cours pour traiter ce cas. Vous donnerez des indications concernant les paramètres des méthodes proposées et vous expliquerez en quoi vos recommandations permettraient de traiter efficacement ce cas.



Exercice 6

Soit un problème de catégorisation où $Y \in \{\alpha, \beta, \gamma\}$ est la variable à expliquer et où X^1 et X^2 sont des variables continues. A partir d'un ensemble d'apprentissage on estime les paramètres d'un réseau de neuronnes. Celui-ci est donné dans la figure suivante :



Questions :

Q1- De quel type de réseau de neuronnes s'agit-il ?

Q2- Les fonctions d'activations sont des softmax. Déterminez pour les deux vecteurs $\mathbf{y}, \mathbf{z}$ suivants, les scores obtenus pour les neurones $g_\alpha, g_\beta, g_\gamma$ ainsi que les résultats des prédictions :

$$\mathbf{y} = (1, 2, 1) \text{ et } \mathbf{z} = (1, -1, -2)$$

Q3- Déterminez la fonction objectif que cherche à minimiser ce réseau de neuronnes. Puis, montrez en quoi ce-dernier est équivalent à une autre méthode de catégorisation vue en cours (détaillez votre démonstration en introduisant les différents concepts pertinents à la méthode proposée).

Exercice 7 -les questions sont ici indépendantes-

Q1- Les trois paramètres (a_0, a_1, a_2) d'un modèle de régression linéaire multiple $Y = a_0 + a_1X^1 + a_2X^2$ sont estimés selon plusieurs modèles distincts. Voici ci-dessous les estimations obtenues par trois modèles paramétriques :

$$\begin{aligned} \mathbf{a}_{m1} &= (1.5, 0, 1.24) \\ \mathbf{a}_{m2} &= (2.1, 0.75, 1.12) \\ \mathbf{a}_{m3} &= (5.2, 2.1, 4.5) \end{aligned}$$

Pour chacun des résultats, quelle est, selon vous, la méthode utilisée parmi celles vues en cours (justifiez vos réponses) ?

Q2- On considère un modèle de régression logistique pour un problème de catégorisation à trois classes A, B, C. Nous avons alors les probabilités conditionnelles suivantes :

$$\begin{aligned} \log \frac{P(A|\mathbf{x})}{P(C|\mathbf{x})} &= a_0 + \mathbf{a}^\top \mathbf{x} \\ \log \frac{P(B|\mathbf{x})}{P(C|\mathbf{x})} &= b_0 + \mathbf{b}^\top \mathbf{x} \end{aligned}$$

Donnez une expression de $P(A|\mathbf{x}), P(B|\mathbf{x}), P(C|\mathbf{x})$ en fonction de $a_0, \mathbf{a}, b_0, \mathbf{b}$ et $\mathbf{x}$ (détaillez vos développements).

Q3- Commentez l'équation suivante en interprétant les termes ou ensemble de termes qui la composent, et en indiquant son utilité en fouille de données :

$$\sum_{i=1}^{n} \phi(\mathbf{x}_i)\phi(\mathbf{x}_i)^\top \mathbf{u}_1 = \lambda_1 \mathbf{u}_1$$

où $\mathbf{x}_i \in \mathbb{R}^p, \forall i = 1, \ldots, n$; $\phi : \mathbb{R}^p \mathbb{F}$; $\lambda_1 \in \mathbb{R}$.

Exercice 8

Soit un problème de catégorisation binaire où $Y \in \{0, 1\}$ est la variable à expliquer et où X^1 et X^2 sont des variables continues. A partir d'un ensemble d'apprentissage on estime les paramètres des fonctions de score [2] des méthodes suivantes : régression logistique et SVM avec un noyau linéaire. Pour chaque modèle on considère une variable pour l'ordonnée à l'origine ("intercept") X^0.

L'estimation des paramètres de la régression logistique est la suivante :

$\hat{a}_{lr,0}$	-429.38
$\hat{a}_{lr,1}$	-64.87
$\hat{a}_{lr,2}$	171.04

L'estimation des paramètres du SVM est donnée dans le tableau ci-après :

$\hat{a}_{svm,0}$	6.12
$\hat{a}_{svm,1}$	1.03
$\hat{a}_{svm,2}$	-2.56

Pour **chacun des deux modèles** estimés :

Q1- Rappelez la définition de la fonction de score et de la fonction de décision [3].

Q2- Calculez la valeur prédite de la fonction de score (à trois décimales près) et celle de la fonction de décision, pour la nouvelle donnée suivante :

$$\mathbf{x} = (6.3, 4.9)$$

2. Celle qui permet de quantifier par un réel l'appartenance à une ou l'autre classe.
3. Celle qui repose sur la fonction de score et qui permet de décider si un élément appartient à l'une ou l'autre classe