Apprentissage Automatique

Feuille d'exercices

Julien Ah-Pine

M2 DM - Université Lyon 2 - 2017/2018

Exercice 1

Soit un problème de régression linéaire simple où Y est la variable à expliquer et X la variable explicative. Nous avons 7 observations et les mesures pour celles-ci sont données ci-dessous :

$$\mathbf{x} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{pmatrix} \quad ; \quad \mathbf{y} = \begin{pmatrix} 6 \\ 4 \\ 2 \\ 1 \\ 3 \\ 6 \\ 10 \end{pmatrix}$$

Q1- Représentez le nuage de points dans le plan (X, Y).

Q2- On se propose d'utiliser une méthode paramétrique et de poser le modèle linéaire suivant : $\forall i : y_i = a_0 + a_1 x_i + \epsilon_i$. Déterminez les estimateurs des moindres carrés ordinaires. Tracez la droite obtenue sur le plan précédent. Calculez les prédictions par le modèle estimé pour chaque observation et déduisez en la moyenne des résidus en valeur absolue.

Q3- On souhaite à présent utiliser une méthode non paramétrique. On utilise pour cela un estimateur à noyau. Le noyau choisi noté K est défini de la façon suivante :

$$K(x, x') = \begin{cases} 1 \text{ si } |x - x'| < 1.5 \\ 0 \text{ sinon} \end{cases}$$

Déterminez les prédictions obtenues par cette méthode pour chaque observation. Tracez les points prédits sur le plan précédent. Déduisez en la moyenne des résidus en valeur absolue.

Q4- En prenant appui sur cet exemple, discutez des différences (points positifs et points négatifs) entre méthodes paramétriques et non paramétriques en apprentissage automatique.

Exercice 2

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.

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.

1. ou une approximation du taux d'erreur

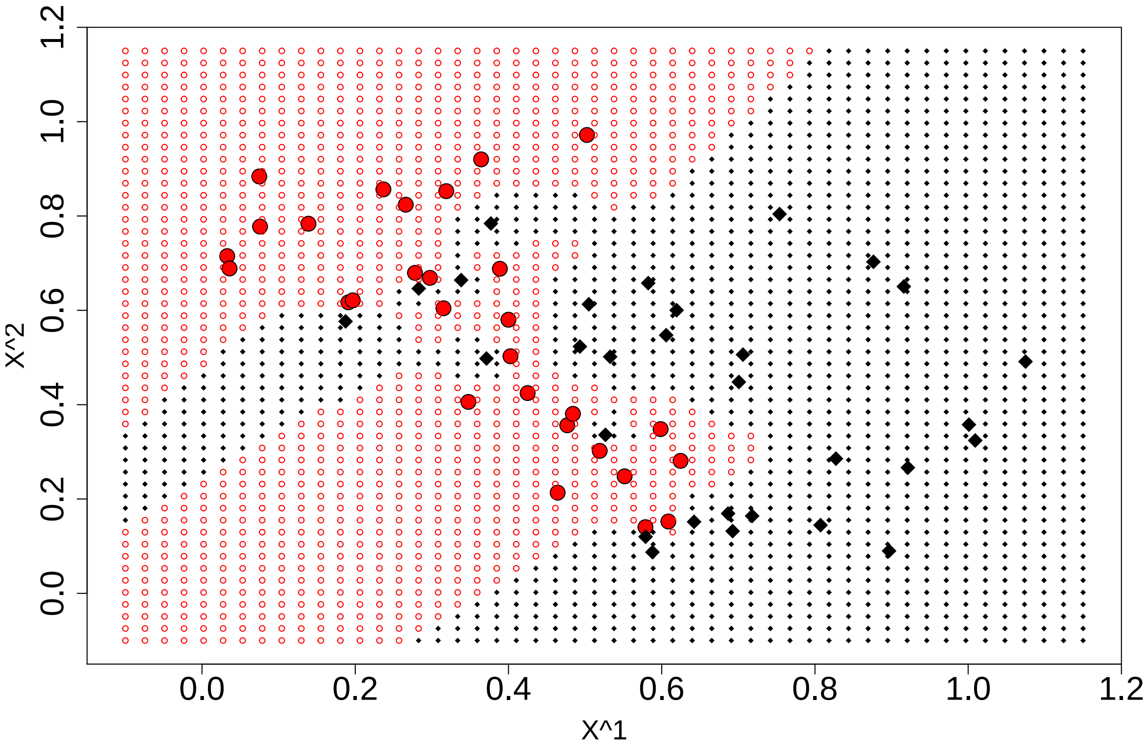


FIGURE 1 – Prédiction avec k_1 plus proches voisins

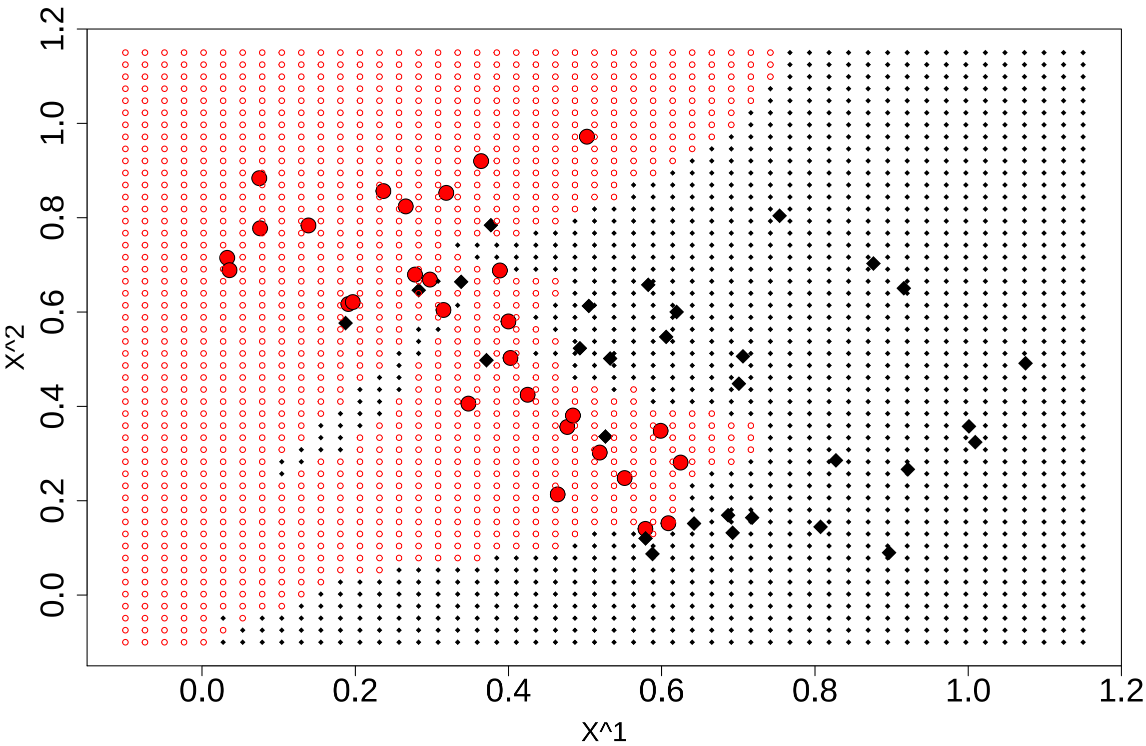


FIGURE 2 – Prédiction avec k_2 plus proches voisins

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.

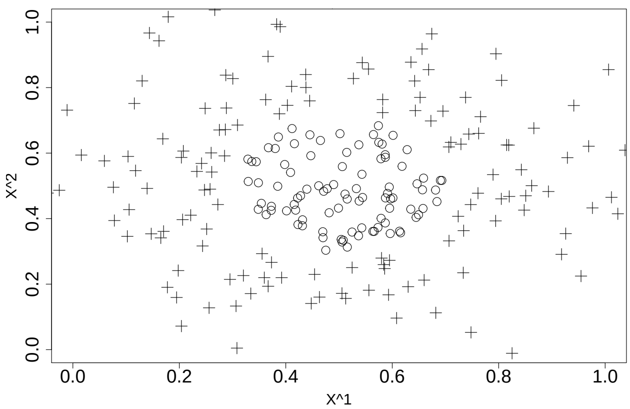
Q5 Lequel de ces deux modèles choisiriez vous dans le cas pratique étudié ici ?

Exercice 3

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 4

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 :

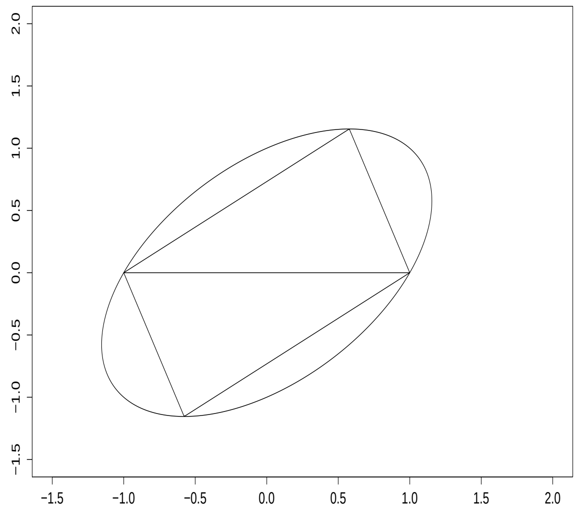
$$\mathbf{y} = \beta_1 \mathbf{x}^1 + \beta_2 \mathbf{x}^2 + \epsilon$$

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 5

Il existe une autre formulation de la méthode SVM dans le cas non linéairement séparable que l'on dénote ν-SVM (prononcé "nu-SVM"). Il s'agit du problème d'optimisation suivant que l'on dénommera par p (pour problème primal) :

$$\begin{array}{ll} & \min_{a_0\in\mathbb{R},\mathbf{a}\in\mathbb{R}^p,\xi\in\mathbb{R}^n,\rho\in\mathbb{R}} \frac{1}{2}\|\mathbf{a}\|^2 + \frac{1}{n}\sum_{i=1}^n \xi_i - \nu\rho \\ slc: & \\ (i) & \forall i : y_i(\mathbf{a}^t\mathbf{x}_i + a_0) \geq \rho - \xi_i \\ (ii) & \forall i : \xi_i \geq 0 \\ (iii) & \rho \geq 0 \end{array}$$

où ν est un paramètre réel appartenant à $[0, 1]$ (contrairement au paramètre $c \in \mathbb{R}$ de la formulation vue en cours).

Q1- Déterminez le Lagrangien de p où vous dénoterez les multiplicateurs de Lagrange de la façon suivante : le vecteur α est associé aux contraintes (i), le vecteur μ est associé aux contraintes (ii) et le réel δ est associé à la contrainte (iii).

Q2- Déterminez les relations issues des équations normales qui sont telles que les dérivées partielles du Lagrangien du primal par rapport aux variables du primal soient nulles.

Q3- Sachant que les multiplicateurs de Lagrange sont positifs, déduisez de la question précédente une inégalité simple entre $\sum_i \alpha_i$ d'une part et le paramètre ν d'autre part. Expliquez alors pourquoi cette méthode qui a pour paramètre $\nu \in [0, 1]$, permet de mieux contrôler le pourcentage de vecteurs supports obtenus par l'optimisation.

Q4- En utilisant les relations obtenus à la question Q2, déterminez le problème dual dont la fonction objectif ne dépend que du vecteur α et dont les contraintes sont par conséquents les relations impliquant cette variable.

Exercice 6

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'apprentissage d'un arbre décisionnel à ce problème de catégorisation binaire. Vous donnerez à chaque itération la mesure de l'entropie justifiant le choix de la variable discriminante. Vous représenterez le résultat obtenu par un arbre.

Q2- 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 $P(X^j = m_k|C_2)$.

Q3- 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).

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

Q4- Déterminez la prédiction obtenue par chacune des méthodes étudiées en Q1 et Q2 pour les 3 nouvelles observations de Q3 (explicitez vos calculs). Commentez les résultats obtenus.

Exercice 9

Soit un problème de régression multiple où Y est la variable à expliquer et X^1 et X^2 sont les variables explicatives. Toutes ces variables sont continues. Nous avons 9 individus et les mesures observées sont données ci-dessous :

$$\mathbf{y} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 4 \\ 4 \\ 4 \\ 2 \\ 2 \\ 2 \end{pmatrix} \quad ; \quad \mathbf{x}^1 = \begin{pmatrix} 1 \\ 2 \\ 1 \\ 2 \\ 1 \\ 2 \\ 6 \\ 6 \\ 6 \end{pmatrix} \quad ; \quad \mathbf{x}^2 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 4 \\ 5 \\ 4 \\ 5 \\ 4 \\ 5 \end{pmatrix}$$

Q1- On se propose de modéliser Y en fonction de X^1 et X^2 par un arbre binaire de régression. Appliquez l'algorithme de contruction d'un arbre de régression avec comme fonction objectif à minimiser, la variance de chaque noeud m (donnant deux branches m^-, m^+) que nous rappelons ci-dessous :

$$err(m) = \frac{|m^-|}{|m|} \sum_{\mathbf{x}_i \in m^-} (y_i - g^{m^-})^2 + \frac{|m^+|}{|m|} \sum_{\mathbf{x}_i \in m^+} (y_i - g^{m^+})^2$$

où $g^{m^-} = \frac{1}{|m^-|} \sum_{\mathbf{x}_i \in m^-} y_i$ est la valeur moyenne de la variable Y sur les individus $\mathbf{x}_i$ appartenant à la branche m^- de cardinal $|m^-|$ issue du noeud m. Vous prendrez $\theta = 0$ comme valeur du seuil pour décider de l'arrêt du développement d'une branche de l'arbre. Il vous est demandé d'indiquer à chaque étape de l'algorithme les éléments définissant une nouvelle branche et un nouveau noeud.

Indication : L'arbre devant être binaire, chaque noeud correspond à un découpage en deux sous-ensembles qui sont de part et d'autre d'une valeur de la variable choisie pour la séparation. Par exemple, dans le cas de la variable X^1, il existe trois possibilités de découpage qui sont les sous ensembles $\{\mathbf{x}_i \in m : \mathbf{x}_i \leq \delta\}$ et $\{\mathbf{x}_i \in m : \mathbf{x}_i > \delta\}$ avec $\delta = 1$ ou 2 ou 6.

Q2- Représentez par un schéma l'arbre obtenu en indiquant les conditions associées à chaque branche ainsi que les étiquettes associées à chaque feuille.