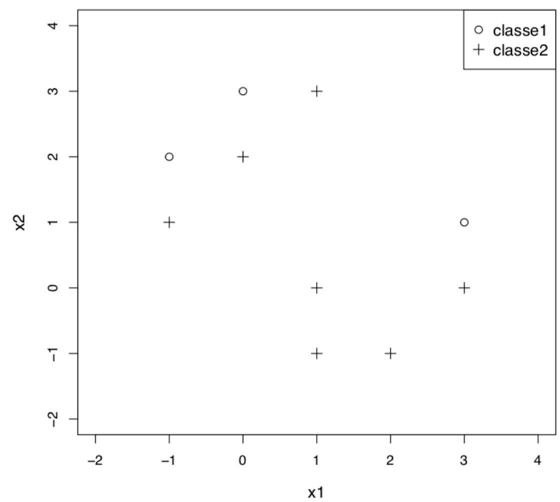
Examen "Grande Dimension et Apprentissage" - 11 Janvier 2019

Les documents ne sont pas autorisés. Le dernier exercice concerne la partie du cours assurée par M. Graczyk.

Exercice 1. On considère l'algorithme du 1-plus proche voisin sur le jeu de données ci-dessous. On suppose également qu'en cas de distances égales, le prédicteur choisit un des voisins avec probabilité 1/2.



1. Calculer l'erreur de validation croisée lorsque l'on choisit l'option "Leave-One-Out" (dont la définition se déduit facilement de la traduction). On détaillera le calcul.
2. Dans le cadre de la classification multi-classes, pouvez-vous en quelques lignes expliquer la propriété de consistance des k-plus proches voisins vue en cours ?

Exercice 2. On cherche à améliorer la qualité d'un estimateur par bagging.

1. En cours, on a vu l'application de cette méthode aux arbres de décision mais on sait qu'elle s'applique dans d'autres contextes. Pour un algorithme de prédiction général que nous noterons $\hat{f}_n$, rappelez le principe de sa mise en oeuvre dans le cadre de :
   (a) la classification binaire,
   (b) la régression (moindres carrés).
2. Dans un problème où l'objectif est la prédiction, le bagging est-il plutôt une méthode à préconiser ? Argumenter.
3. Même question lorsque l'objectif est la compréhension d'un problème via une approche "analyse de données".
4. Expliquez la nuance principale entre les forêts aléatoires et le bagging.

Exercice 3 (Classification binaire)**.** Soit un problème d'apprentissage où Y est à valeurs dans $\{-1,1\}$ et $Y|X=x$ suit une loi de Rademacher de paramètre $p(x)$ où $p(x)\in[0,1]$. Pour simplifier, on suppose que X est une variable discrète à valeurs dans un ensemble $\mathcal{X}$. On note $\mathcal{F}$ l'ensemble des fonctions f de $\mathcal{X}$ dans $\{-1,1\}$.

1. On note $\Phi_f(x)=\mathbb{P}(Y\neq f(X)|X=x)$.
   (a) Montrez que
$$\Phi_f(x)=p(x)\mathbf{1}_{\{f(x)=-1\}}+(1-p(x))\mathbf{1}_{\{f(x)=1\}}.$$

(b) En déduire que pour tout $x \in \mathcal{X}$

$$\min_{f \in \mathcal{F}} \Phi_f(x) = \min(p(x), 1 - p(x))$$

et déterminez $f^*(x) = \mathrm{Argmin}_{f \in \mathcal{F}} \Phi_f(x)$.

(c) On suppose dans cette question que X suit la loi uniforme sur $\{\frac{k}{10}, k \in \{1, \ldots, 10\}\}$ et que $p(x) = x$. Montrez que dans ce cas le risque minimal (risque de Bayes) est égal à $\frac{1}{4}$.

2. On considère une fonction $h : \{-1, 1\} \to \mathbb{R}$ telle que $\eta = h(-1) - h(1) > 0$. On pose

$$\Psi_f(x) = \mathbb{E}[h(Yf(X))|X = x].$$

(a) Montrez que pour tout $f \in \mathcal{F}$.

$$\Psi_f(x) = p(x)(h(f(x)) - h(-f(x))) + h(-f(x)).$$

(b) En déduire que

$$\Psi_f(x) \geq \begin{cases} \eta p(x) + h(1) & \text{si } p(x) \leq 1/2 \\ -\eta p(x) + h(-1) & \text{si } p(x) > 1/2. \end{cases}$$

(c) En déduire que

$$\Psi_{f^*}(x) = \mathrm{Argmin}_{f \in \mathcal{F}} \Psi_f(x).$$

Exercice 4 (Ridge/LASSO)**.** On note $\mathbf{y} = (y_1, \ldots, y_n)^T$ un vecteur de taille n et $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ une matrice $n \times p$ ($\mathbf{x}_j$ est un vecteur colonne de taille n). Pour un s fixé supérieur ou égal à 1, on considère la fonction L_s définie sur $\mathbb{R}^p$ par

$$L_s(\theta) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i \theta)^2 + \lambda \|\theta\|_s^s$$

où $\|\theta\|_s^s = \sum_{j=1}^{p} |\theta_j|^s$ et $\lambda > 0$.

1. Montrez que dans le cas $s = 2$, L_s admet un unique minimum que l'on déterminera.
2. Considérons le cas $s \in]1, 2[$.
   (a) Montrez que la fonction admet au moins un minimum.
   (b) Par un argument de convexité, montrez que ce minimum est unique. On le notera $\hat{\theta}$.
   (c) Ecrire l'équation satisfaite par $\hat{\theta}$.
   (d) Dans le cas $s = 3/2$ et en dimension 1, déterminez ce minimum.
3. $s = 1$, $p = 1$. Déterminez l'unique point critique de L_1. Pourquoi est-ce un minimum ? Sous quelles conditions a-t-on un seuillage de l'estimateur non pénalisé ?

Exercice 5 (SVM)**.** Avec les notations du cours sur les SVM dans le cadre de la classification binaire (à valeurs dans $\{-1, 1\}$), on considère le problème d'optimisation suivant :

$$\begin{cases} \min_{\beta, \beta_0} \frac{1}{2} \|\beta\|^2 \text{ sous la contrainte} \\ \forall i \in \{1, \ldots, N\}, \quad y_i(\langle \beta, x_i \rangle + \beta_0) \geq 1. \end{cases} \tag{1}$$

1. Illustrer par un dessin les différentes variables mises en jeu dans ce problème.
2. Ecrire la formulation Lagrangienne du problème.
3. Quelles formes de généralisation peut-on envisager ? (Ecrire quelques lignes sur le sujet)

Exercice 6 (P. Graczyk)**.** Soit X un caractère statistique Gaussien de dimension 4 centré

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} \sim N(0, \Sigma)$$

avec les matrices de covariance $\Sigma = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 \\ 1 & 2 & 3 & 4 \end{pmatrix}$ et de précision $K = \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{pmatrix}$.

1. Quelle est la relation entre les matrices Σ et K ?

2. Y a-t-il des composantes X_i indépendantes entre elles ? Si oui, lesquelles ?

3. Y a-t-il des composantes X_i conditionellement indépendantes sachant les autres ? Si oui, lesquelles ?

4. Dessiner le graph de dépendence $\mathcal{G}$ de X.

5. Déterminer la loi marginale de $(X_1, X_2)^T$.

6. Déterminer la loi conditionnelle $(X_1, X_2)^T | (X_3 = u, X_4 = v)$.

7. Déterminer la corrélation conditionnelle $\rho_{X_1, X_2 | (X_3 = u, X_4 = v)}$.

8. Quels ensembles de vertices de $\mathcal{G}$ sont séparés par $S = \{2\}$? Qu'en déduit-on sur la prédiction de X_1 ?

9. Le graphe $\mathcal{G}$, est-il triangulé(décomposable) ?

10. (a) Pourquoi les graphes décomposables sont-ils importants en Modèles Graphiques statistiques ?

    (b) Donner un exemple d'un graphe non-décomposable.

Rappels de cours. *Soit X un vecteur gaussien $N(\xi, \Sigma)$ dans $\mathbf{R}^d$ avec Σ inversible.*
On partitionne $X = \begin{pmatrix} X_A \\ X_B \end{pmatrix}$ en sous-vecteurs $X_A \in \mathbf{R}^r$ et $X_B \in \mathbf{R}^s$, avec $r + s = d$. On partitionne $\xi = \begin{pmatrix} \xi_A \\ \xi_B \end{pmatrix}$,
$\Sigma = \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{pmatrix}$, $K = \begin{pmatrix} K_{AA} & K_{AB} \\ K_{BA} & K_{BB} \end{pmatrix}$ *en blocs* $\begin{pmatrix} r \times r & r \times s \\ s \times r & s \times s \end{pmatrix}$.

La **loi conditionnelle** $X_A | \ (X_B = x_B) \ \sim N(\xi_{A|B}, \Sigma_{A|B})$ *où* $\xi_{A|B} = \xi_A + \Sigma_{AB}\Sigma_{BB}^{-1}(x_B - \xi_B)$ et $\Sigma_{A|B} = K_{AA}^{-1}$.

La **corrélation conditionnelle** $\rho_{lm|V \setminus \{l,m\}} = -\tilde{\kappa}_{lm} = -\frac{\kappa_{lm}}{\sqrt{\kappa_{ll}}\sqrt{\kappa_{mm}}}$,