APPRENTISSAGE STATISTIQUE

EXAMEN - CONTRÔLE DE CONNAISSANCES (DURÉE 1 HEURE 30)

Les notes de cours ne sont pas autorisées, l'usage d'ordinateurs ou tablettes est prohibé.

Notations. On se place dans le cadre du modèle de classification où X est un vecteur aléatoire sur $\mathbb{R}^d$, $d \geq 1$, de loi $\mu(dx)$ et Y est une variable aléatoire à valeurs dans $\{-1, +1\}$. On pose $\eta(X) = \mathbb{P}(Y = 1 \mid X)$, $p = \mathbb{P}\{Y = +1\} = \mathbb{E}[\eta(X)]$ et on suppose la v.a. $\eta(X)$ continue pour simplifier. Le risque d'un classifieur $g : \mathbb{R}^d \to \{-1, +1\}$ est défini par $L(g) = \mathbb{P}\{Y \neq g(X)\}$. On suppose que l'on dispose d'une collection d'exemples $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, copies indépendantes du couple générique (X, Y). On désigne par $\langle ., . \rangle$ et $||.||$ le produit scalaire et la norme euclidienne usuels sur $\mathbb{R}^d$. La fonction indicatrice d'un événement quelconque $\mathcal{E}$ est notée $\mathbb{I}\{\mathcal{E}\}$.

THÉORIE DE L'APPRENTISSAGE

1 Soit $\mathcal{A}$ une classe de sous-ensembles mesurables de $\mathbb{R}^d$. Définir son coefficient d'éclatement à l'ordre n, sa dimension de Vapnik-Chervonenkis.

2 Définir le risque empirique $\widehat{L}_n(g)$ d'un classifieur g calculé sur l'échantillon d'apprentissage $\mathcal{D}_n$.

3 Expliquer en quoi consiste le principe de minimisation du risque empirique appliqué à une classe $\mathcal{G}$ de classifieurs.

Pour chacune des affirmations ci-dessous, préciser si elle est vraie ou fausse (aucune justification n'est demandée).

4 Le problème de la minimisation du risque empirique $\widehat{L}_n(g)$ est NP-difficile et la plupart des algorithmes reposent en pratique sur un relaxation du problème original.

5 Pour mettre en oeuvre la sélection de modèle, on se fonde sur une estimation de l'erreur de généralisation calculée sur un échantillon de validation, par validation croisée ou par rééchantillonnage.

6 Le classifieur minimisant le risque à coût sensitif

$$L_w(g) = 2p(1-w)\mathbb{P}\{g(X) = -1 \mid Y = +1\} + 2(1-p)w\mathbb{P}\{g(X) = +1 \mid Y = -1\},$$

avec $w \in (0, 1)$, est donné par : $\forall x \in \mathbb{R}^d$,

$$g^*(x) = 2\mathbb{I}\{\eta(x) \geq w\} - 1.$$

ALGORITHMES "BASIQUES"

1. L'un des avantages de l'algorithme CART réside dans les propriétés d'interprétabilité des règles qu'il produit. Quelles sont-elles ?

2. Qu'entend-on par l'affirmation suivante : 'L'algorithme CART est peu stable' ? Quelle stratégie fondée sur le 're-échantillonnage' peut être utilisée pour corriger l'instabilité ? Cette stratégie peut-elle être appliquée à un autre algorithme ?
3. Comment peut on effectuer une régression logistique dans le cas où les variables sont catégorielles ?

Pour chaque affirmation ci-dessous, préciser si elle est vraie ou fausse (aucune justification n'est demandée).

3 Le modèle de la régression logistique linéaire pour la classification binaire requiert de stipuler que la loi de X sachant $Y = +1$ et la loi de X sachant $Y = -1$ sont des gaussiennes de même moyenne mais de matrices de covariance différentes.

4 La sortie de l'algorithme du Perceptron monocouche cesse d'évoluer au bout d'un nombre variable mais fini d'itérations lorsque les populations 'positives' et 'négatives' de l'échantillon d'apprentissage sont séparables par un hyperplan affine.

ALGORITHMES "AVANCÉS"

On se place toujours dans le cadre de la classification supervisée binaire déjà décrite plus haut.

1. Le problème d'optimisation résolu par l'algorithme SVM peut être formulé de façon à l'interpréter comme la minimisation d'un risque empirique pour la perte 'hinge' $(1+u)_+$ pénalisé.
2. L'"astuce du noyau" permet de déterminer une règle de décision affine dans l'espace de représentation (et non linéaire dans l'espace d'entrée original si le noyau n'est pas un produit scalaire dans l'espace d'entrée $\mathbb{R}^d$) sans avoir à spécifier la représentation afférente (*i.e.* "feature variables").
3. L'algorithme ADABOOST produit itérativement un classifieur $sgn(f(X))$ minimisant une version empirique de l'erreur exponentielle :

$$L_{\exp}(f) = \mathbb{E}[\exp(-Yf(X))].$$

RÉSEAUX DE NEURONES

Répondre aux questions suivantes de manière concise.

1. Quelle loss cherche-ton à minimiser dans un réseau de neurone de sortie binaire ? Donnez son nom, sa formule mathématique. On notera $\hat{y}^{(i)}$ la valeur de sortie pour l'exemple i et $y^{(i)}$ sa vérité terrain.
2. Justifier cette formule de la loss en vous appuyant sur la loi de probabilité de Bernouilli que suit y dont le cas d'un problème binaire
3. Nous considérons un réseau de neurone prenant en entrée un vecteur $\underline{x}^{(i)}$ de taille 50, ayant 4 couches cachées de 100 neurones chacune et dont la sortie est binaire. Quels sont les paramètres à apprendre de ce réseau ? Combien y en a t'il ?

4. Dans notre réseau à 4 couches cachées, nous avons choisi des non-linéarités de type sigmoide. Ce choix provoque un effet connu sous le nom de "vanishing gradient". A l'aide de la formule de la sigmoide, de la formule de sa dérivée et de l'algorithme de back-propagation expliquez cet effet. Nous noterons $\underline{z}_1$ la pré-activation de la couche 1, et $\underline{a}_1$ son activation, $\underline{z}_2$ la pré-activation de la couche 2, ... ainsi $\underline{z}_1 = \underline{\underline{W}}_1 \underline{x}$ ($\underline{x}$ est l'entrée du réseau), $\underline{z}_2 = \underline{\underline{W}}_2 \underline{a}_1$ et $\underline{a}_1 = \sigma(\underline{z}_1)$.

CLUSTERING

On considère le problème du clustering de n échantillons de $\mathbb{R}^d$, $x_1, \ldots, x_n$. On s'intéresse au score de concentration (CS) permettant de mesurer la qualité d'une partition $C_1, \ldots, C_K$ des échantillons :

$$\mathrm{CS} = 1 - \frac{\sum_{j=1}^{K} \sum_{i \in C_j} ||x_i - \mu_j||^2}{\sum_{i=1}^{n} ||x_i - \mu||^2},$$

avec

$$\mu_j = \frac{1}{|C_j|} \sum_{i \in C_j} x_i, \quad \mu = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

1. Montrer que $0 \leq \mathrm{CS} \leq 1$.
2. Dans quels cas a-t-on $\mathrm{CS} = 0$? $\mathrm{CS} = 1$?
3. Donner les avantages et inconvénients de cette métrique.

SÉRIES TEMPORELLES

1. On observe une quantité échantillonnée à une fréquence trimestrielle sous la forme d'une série temporelle $(X_t)_{t \in \mathbb{Z}}$. On souhaite éliminer une tendance périodique de période 1 an. Proposer un entier $k \geq 0$ et des coefficients $\psi_0, \ldots, \psi_k$ tels que $(Y_t)_{t \in \mathbb{Z}}$ défini par

$$Y_t = \sum_{j=0}^{k} \psi_j X_{t-j}$$

ne contienne plus cette tendance.

2. Quelle relation y-a-t-il entre le périodogramme

$$I_n(\lambda) = \frac{1}{2\pi n} \left| \sum_{k=1}^{n} (X_k - \hat{\mu}_n) \, \mathrm{e}^{-\mathrm{i}\lambda k} \right|^2$$

et la covariance empirique :

$$\hat{\gamma}_n(h) = \frac{1}{n} \sum_{k=1}^{n-|h|} (X_k - \hat{\mu}_n)(X_{k+|h|} - \hat{\mu}_n) ?$$

Classification basée sur la profondeur de données

Pour chaque affirmation ci-dessous, préciser si elle est vraie ou fausse (aucune justification n'est demandée).

1. La profondeur de données de Tukey (Tukey depth) d'un point $\mathsf{x} \in \mathbb{R}^d$ par rapport à un ensemble de données $\mathsf{X} = \{\mathsf{x}_1, \mathsf{x}_2, ..., \mathsf{x}_n\} \subset \mathbb{R}^d$ peut être formulée de la manière suivante (où $\sharp$ désigne le cardinal d'un ensemble, S^{d-1} la sphere unité et $^\top$ l'opérateur de transposition) :
$$D^{Tuk(n)}(\mathsf{x}|\mathsf{X}) = \frac{1}{n} \min_{\mathsf{u} \in S^{d-1}} \sharp\{i : \mathsf{u}^\top \mathsf{x}_i \geq \mathsf{u}^\top \mathsf{x}\}.$$

2. La profondeur de données de Mahalanobis (Mahalanobis depth) d'un point $\mathsf{x} \in \mathbb{R}^d$ par rapport à un ensemble de données $\mathsf{X} = \{\mathsf{x}_1, \mathsf{x}_2, ..., \mathsf{x}_n\} \subset \mathbb{R}^d$ peut être formulée de la manière suivante (où $^\top$ désigne l'opérateur de transposition) :
$$D^{Mah(n)}(\mathsf{x}|\mathsf{X}) = \frac{1}{1 + (\mathsf{x} - \mu_\mathsf{X})^\top \Sigma_\mathsf{X}^{-1} (\mathsf{x} - \mu_\mathsf{X})}$$
avec $\mu_\mathsf{X} = \frac{1}{n} \sum_{i=1}^n \mathsf{x}_i$ et $\Sigma_\mathsf{X} = \frac{1}{n-1} \sum_{i=1}^n (\mathsf{x}_i - \mu_\mathsf{X})(\mathsf{x}_i - \mu_\mathsf{X})^\top$.

3. Soient $\mathsf{X} = \mathsf{X}_0 \cup \mathsf{X}_1$ l'ensemble d'apprentissage constitué de deux sous-ensembles : $\mathsf{X}_0 = \{\mathsf{x}_1, \mathsf{x}_2, ..., \mathsf{x}_m\}$ et $\mathsf{X}_1 = \{\mathsf{x}_{m+1}, \mathsf{x}_{m+2}, ..., \mathsf{x}_{m+n}\}$ ayant les labels 0 et 1 respectivement. Le classifieur de profondeur maximale (maximum depth classifier) est défini de la manière suivante :
$$g(\mathsf{x}) = \begin{cases} 1 & \text{si } D(\mathsf{x}|\mathsf{X}_0) \leq D(\mathsf{x}|\mathsf{X}_1), \\ 0 & \text{sinon}. \end{cases}$$
où D est une profondeur de données et le traitement des ex-æquo est prévu.

ACP/AFCM

1. Lorsque l'on fait l'analyse en composantes principales d'une matrice $X \in \mathbb{R}^{n \times p}$, avec des colonnes préalablement centrées et standardisées, comment sont définis les axes principaux ?
2. Dans le cadre de la question précédente, que représente la carte des individus obtenue avec les packages standard (ex. FactoMineR)
3. Pour l'analyse des correspondances de deux variables discrètes, lorsque l'on analyse la première variable, sur quelle matrice effectue-t-on une ACP ? quelle est la métrique utilisée ? quel poids attribue -t-on à chaque catégorie ? où $\tilde{d}_j$ est l'effectif de la catégorie j, $\tilde{d}_j = \sum_{i=1}^{m_1} N_{i,j}$.

Text-Mining

On se place dans le cadre de l'utilisation du bayésien naïf pour la classification de documents.

1. Quelle est l'hypothèse (naïve) que l'on fait sur les mots d'un document ?
2. Expliquez comment, à partir de comptes $T(w, c)$ pour tous les mots w d'un vocabulaire $\mathcal{V}$ et pour chaque classe $c \in C$, et des probabilités à priori $P(c)$, on peut prédire à laquelle de ces classes un document inconnu appartient ?