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.

Le coefficient d'éclatement à l'ordre n de la classe $\mathcal{A}$ est donné par :

$$S_{\mathcal{A}}(n) = \max_{\{x_1,\ \ldots,\ x_n\} \subset \mathbb{R}^d} card\,\{A \cap \{x_1,\ \ldots,\ x_n\} :\ \ A \in \mathcal{A}\}\,.$$

La dimension de Vapnik-Chervonenkis de la classe $\mathcal{A}$ est donnée par :

$$dim_{VC}\mathcal{A} = \sup\{k :\ \ S_{\mathcal{A}}(k) = 2^k\} \in \mathbb{N}^* \cup \{+\infty\}.$$

2 Définir le risque empirique d'un classifieur g calculé sur l'échantillon d'apprentissage $\mathcal{D}_n$.
Le risque empirique d'un classifieur g est donné par

$$\widehat{L}_n(g) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{Y_i \neq g(X_i)\}.$$

(La statistique $\widehat{L}_n(g)$ est un estimateur naturel du risque $L(g)$. Le principe de minimisation du risque empirique consiste à remplacer le problème de minimisation $\min L(g)$ par le problème $\min_{g\in\mathcal{G}} \widehat{L}_n(g)$ sur une classe $\mathcal{G}$ de complexité contrôlée. L'excès de risque du minimiseur du risque empirique $\widehat{g}_n = \arg\min_{g\in\mathcal{G}} \widehat{L}_n(g)$ satisfaisant

$$L(\widehat{g}_n) - L(g^*) \leq \sup_{g\in\mathcal{G}} |\widehat{L}_n(g) - L(g)| + \inf_{g\in\mathcal{G}} L(g) - L(g^*), \tag{1}$$

un bon choix pour la classe $\mathcal{G}$ conduirait à un equilibre entre les termes stochastique et de biais.)

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

3 Pour mettre en oeuvre la sélection de modèle, on se fonde toujours seulement sur l'erreur d'apprentissage.
FAUX

4 Le classifieur optimal (*i.e.* de risque minimum) est donné par : $\forall x \in \mathbb{R}^d$,

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

VRAI

ALGORITHMES "BASIQUES"

1. On dispose d'une métrique $D(.,.)$ sur l'espace d'entrée $\mathbb{R}^d$. Soit $k \in \{1, \ldots, n\}$. Pour le problème de la classification binaire, explicitez la règle des k-plus proches voisins fondée sur la métrique D et l'échantillon $\mathcal{D}_n$.
Pour tout $x \in \mathbb{R}^d$, on désigne par σ_x une permutation de $\{1, \ldots, n\}$ telle que :

$$D(x, X_{\sigma_x(1)}) \geq \ldots \geq D(x, X_{\sigma_x(n)}).$$

La règle est donnée par :

$$g_k(x) = 2\mathbb{I}\{\sum_{i=1}^{k} Y_{\sigma_x(i)} \geq 0\} - 1.$$

2. On désigne par $n_+ = \sum_{k=1}^{n} \mathbb{I}\{Y_i = +1\} = n - n_-$ le nombre de données de l'échantillon d'apprentissage avec un label positif. Préciser si l'assertion suivante est vraie ou fausse (aucune justification n'est demandée) : "Pour $k = n$, l'erreur d'apprentissage de la règle des plus proches voisins est donnée par $\min\{n_+/n,\ n_-/n\}$".

VRAI

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 des formes paramétriques pour la loi de X sachant $Y = +1$ et la loi de X sachant $Y = -1$.

FAUX (seule la loi conditionnelle de Y sachant X est modélisée).

4 Le modèle de l'analyse discriminante 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 gaussiennes de même matrice de covariance.

VRAI

5 La sortie de l'algorithme du Perceptron monocouche cesse d'évoluer au bout d'un nombre fini d'itérations lorsqu'il existe un hyperplan affine séparant les données d'apprentissage avec un label positif des données d'apprentissage avec un label négatif.

VRAI

6 La notion d'*importance relative* permet de quantifier l'impact d'une variable explicative sur une règle prédictive produite par l'algorithme CART.

VRAI

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 comme un problème d'optimisation quadratique sous contraintes linéaires.

   VRAI

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").

   VRAI

3. L'algorithme AdaBoost produit itérativement des classifieurs à partir d'échantillons pondérés qui sont combinés *in fine* via un simple vote à la majorité (*i.e.* une prédiction est positive si la majorité des classifieurs prédisent un label positif).

   FAUX (le vote est pondéré, le poids d'un 'classifieur faible' est égal à $\log((1-err)/err)$, err désignant son erreur pondérée).

Clustering

Soit X un vecteur aléatoire sur $\mathbb{R}^d$ de loi μ et $\mathcal{D}_n = \{X_1, X_2, \ldots, X_n\}$ un n-échantillon i.i.d tiré de cette loi. Soit $K \in \{1, \ldots, n\}$ le nombre de clusters désirés.

1. L'algorithme K-means vise à déterminer une partition $C_1, \ldots, C_K$ du nuage de points minimisant
$$\sum_{k=1}^{K} \sum_{(X_i,X_j)\in C_k^2} ||X_i - X_j||^2.$$

   VRAI

2. L'algorithme K-means vise à déterminer une partition $C_1, \ldots, C_K$ du nuage de points maximisant
$$\sum_{1\leq k\neq l\leq K} \sum_{(X_i,X_j)\in C_k\times C_l} ||X_i - X_j||^2.$$

   VRAI (la somme de ces deux critères est indépendante de la partition, égale à la dispersion totale du nuage de points.)

3. Les sorties de l'algorithme K-means cessent d'évoluer au bout d'un nombre fini d'itérations.
   VRAI (le critère évolue de façon monotone et ne prend qu'un nombre fini de valeurs).

Analyse en Variables Latentes

1. Le cadre de validité de l'analyse en composantes indépendantes stipule que le vecteur observé soit gaussien.

   FAUX (seule une composante au maximum peut être gaussienne).

2. Les composantes obtenues par l'Analyse en Composantes Principales sont des combinaisons linéaires des composantes originales du vecteur analysé.

   VRAI