Apprentissage statistique

Examen - Contrôle de Connaissances (durée 3 heures)

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

Eléments de Théorie de l'Apprentissage Statistique

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 $\mathbb{I}\{\mathcal{E}\}$.

1. Donner l'expression du *classifieur de Bayes* (*i.e.* le minimiseur de $L(g)$ sur l'ensemble des classifieurs g) pour le problème de classification relatif au couple (X, Y). Exprimer son risque en fonction de η et de $\mu(dx)$.
2. Définir le risque empirique d'un classifieur calculé à partir de $\mathcal{D}_n$. Expliquer (de façon concise) le principe de la *Minimisation du Risque Empirique*.
3. 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.
4. Donner la dimension de Vapnik-Chervonenkis de la classe des séparateurs affines

$$\mathcal{A}_0 = \{\{x \in \mathbb{R}^d :\ \ \langle x, w\rangle + \theta > 0\},\ \ (w, \theta) \in \mathbb{R}^d \times \mathbb{R}\}.$$

5. Expliquez brièvement les garanties que donnent la théorie de Vapnik-Chervonenkis pour le principe de Minimisation du Risque Empirique.

Algorithmes - Sélection de Modèle

1. ("Plus proches voisins") On se place dans le cadre de la classification binaire décrit précédemment. On dispose de $M \geq 1$ métriques $d_1,\ \ldots,\ d_M$ sur $\mathbb{R}^d$. Proposer une méthode expérimentale pour choisir la métrique d_m et le nombre $k \in \{1,\ \ldots,\ n\}$ du classifieur des plus proches voisins dont l'erreur de généralisation est susceptible d'être la moindre avec une forte probabilité
   (a) dans le cas où la taille n de l'échantillon est très grande,
   (b) dans le cas où la taille n de l'échantillon est petite.

2. (Analyse Discriminante Linéaire) On se place dans le cadre de la classification binaire décrit précédemment. On pose

$$\begin{aligned}
\hat{p} &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{Y_i=+1\} = n_+/n = 1-n_-/n,\\
\hat{\mu}_+ &= \frac{1}{n_+}\sum_{i=1}^{n}X_i\mathbb{I}\{Y_i=+1\},\\
\hat{\mu}_- &= \frac{1}{n_-}\sum_{i=1}^{n}X_i\mathbb{I}\{Y_i=-1\},\\
\hat{\Gamma} &= \hat{p}\hat{\Gamma}_+ + (1-\hat{p})\hat{\Gamma}_-,
\end{aligned}$$

avec

$$\begin{aligned}
\hat{\Gamma}_+ &= \frac{1}{n_+}\sum_{i=1}^{n}{}^t(X_i-\hat{\mu}_+)(X_i-\hat{\mu}_+)\mathbb{I}\{Y_i=+1\},\\
\hat{\Gamma}_- &= \frac{1}{n_-}\sum_{i=1}^{n}{}^t(X_i-\hat{\mu}_-)(X_i-\hat{\mu}_-)\mathbb{I}\{Y_i=-1\}.
\end{aligned}$$

   (a) A partir des quantités ci-dessus, écrire la règle de classification $g_{LDA}(x)$ reposant sur le modèle de l'analyse disciminante linéaire et en supposant que la matrice $\hat{\Gamma}$ est inversible.

   (b) Sur quelles hypothèses repose le modèle de l'analyse discriminante linéaire ?

   (c) Peut-on qualifier le classifieur $g_{LDA}(x)$ de classifieur "plug-in" ?

3. (Perceptron monocouche) On se place dans le cadre de la classification binaire décrit précédemment.

   (a) Décrire l'algorithme du Perceptron monocouche de F. Rosenblatt.

   (b) Dans quelle situation cet algorithme converge (au sens où aucun critère d'arrêt n'a besoin d'être spécifié par l'utilisateur)

   (c) Cet algorithme est-il adapté au cas où les données sont observées de façon séquentielle ?

4. (Régression logistique linéaire) On se place dans le cadre de la classification binaire décrit précédemment.

   (a) Décrire le modèle statistique de la régression logistique linéaire (Que modélise-t-on et comment ?).

   (b) Quel critère utilise-t-on pour ajuster ce modèle aux données d'apprentissage $\mathcal{D}_n = \{(X_1,\ Y_1),\ \ldots,\ (X_n, Y_n)\}$?

   (c) Proposer une méthode numérique pour optimiser le critère d'ajustement évoqué ci-dessus.

   (d) Comment exploiter le modèle statistique ajusté au données pour construire un classifieur ? Sur quel principe s'appuie cette approche ?

5. (Algorithme CART) On se place dans le cadre de la **régression** : Y est une v.a. réelle que l'on cherche à prédire par une fonction du vecteur aléatoire X, à valeurs dans $\mathbb{R}^d$. L'objectif est de construire une fonction (mesurable) $f : \mathbb{R}^d \to \mathbb{R}$ de risque quadratique $L_2(f) = \mathbb{E}[(f(X)-Y)^2]$ minimum à partir de copies indépendantes $(X_1,\ Y_1),\ \ldots,\ (X_n, Y_n)$ du couple (X, Y).

   (a) Quel type de fonction f produit l'algorithme CART ?

(b) Décrire la première phase de l'algorithme CART appliqué à la minimisation du risque quadratique.

(c) Expliquer comment quantifier l'impact de chacune des variables d'entrée (*i.e.* chaque coordonnée du vecteur aléatoire X) sur la règle produite.

Machine à Vecteurs Support

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

1. Quel problème d'optimisation faut-il résoudre dans l'espace primal pour trouver l'hyperplan de marge maximale (ou SVM linéaire) en faisant l'hypothèse de données bruitées ? Il ne vous est pas demandé de résoudre le problème dans l'espace dual mais d'expliquer la signification de chacun des termes décrivant le problème.
2. On fait maintenant l'hypothèse que les deux classes considérées sont fortement déséquilibrées. On considère alors qu'une erreur de prédiction sur la classe $+1$ n'est pas équivalente à une erreur de prédiction sur la classe -1. Proposer une modification simple du problème précédent pour en tenir compte.

Ensemble Learning

1. Expliquer le principe d'Adaboost, méthode d'ensemble dédiée à la classification supervisée binaire
2. Quelle est la fonction de coût sous-jacente ?

Clustering

Soit X un vecteur aléatoire sur $\mathbb{R}^d$ de loi $P(X)$ et $\mathcal{D}_n = \{X_1, X_2, \ldots, X_n\}$ un n-échantillon i.i.d tiré de cette loi. Soit K le nombre de clusters à déterminer. On cherche à estimer la loi P par un modèle de mélange de gaussiennes puis à en déduire une classification (non supervisée).

1. Définir l'expression de la densité de probabilité de ce modèle de mélange.
2. Exprimer la log-vraisemblance des données relativement au modèle.
3. Expliquer pourquoi il est difficile de minimiser cette fonction objectif.
4. Donner le principe de l'algorithme Expectation-Maximization et ses deux étapes
5. Comparer le type de clustering obtenu par l'estimation d'un modèle de mélange de gaussiennes et celui obtenu par par les k-moyennes

Analyse en Variables Latentes

1. Quel critère optimise t-on lors d'une analyse en composantes principales (ACP) ?
2. En quoi l'ACP est-elle pertinente pour réaliser une réduction de dimension ?
3. Comment pourrait-on quantifier la fraction de variance expliquée par les K premières composantes principales ?
4. Expliquer l'expression "non-Gaussian is independent" ?

5. Proposer un critère permettant de mesurer la "non-Gaussiennité" d'un échantillon de données ?
6. Quel est l'intérêt d'utiliser une fonction de coût autre que l'erreur quadratique dans le calcul d'une factorisation en matrices positives (NMF) ?
7. Dècrire une méthode heuristique pour la conception d'un algorithme de calcul de NMF utilisant des règles de mise à jour multiplicatives.
8. Quel est l'intérêt des méthodes de majoration-minimisation dans l'optimisation des paramètres d'un modèle NMF ?