Contrôle de Connaissance
Master Recherche Informatique, parcours AIC - Université Paris-Saclay
TC1 : Apprentissage

16 novembre 2016
Durée : 3h00
Documents autorisés : supports et notes de cours
Les parties I et II sont indépendantes. Merci de les rédiger sur des copies distinctes.

Partie I : Validation, Arbres de décision, Réduction de problèmes

1 Validation (3 points)

Soit $\mathcal{E} = \{(x_i, y_i), x_i \in \mathcal{X}, y_i \in \{1, -1\}$ une base d'apprentissage supervisée à deux classes ($+1$ et -1). On suppose que l'espace des instances $\mathcal{X}$ est l'espace des vecteurs réels de dimension d ($\mathcal{X} = \mathbb{R}^d$).

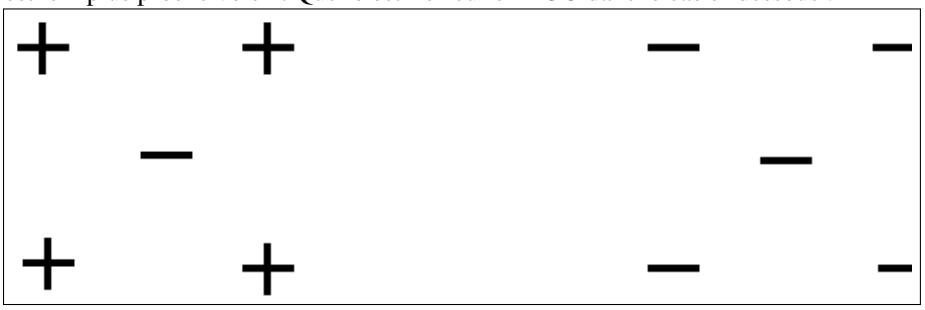
1. Qu'est-ce que le sur-apprentissage (overfitting) ? Comment le détecte-t-on ?
2. Vous apprenez un polynome $h(x)$ de degré p pour séparer les deux classes.

   $$y = signe(h(x))$$

   Doit-on augmenter ou diminuer le degré p du polynome en cas d'overfitting ? Pourquoi ?
3. La validation croisée partitionne les données $\mathcal{E}$ en k sous-ensembles $\mathcal{E}_1, \ldots \mathcal{E}_k$ (k est souvent égal à 10). Une hypothèse h_i est apprise à partir de

   $$\mathcal{E}_{-i} = \mathcal{E}_1 \bigcup \ldots \bigcup \mathcal{E}_{i-1} \bigcup \mathcal{E}_{i+1} \bigcup \ldots \bigcup \mathcal{E}_k$$

   et on calcule son erreur sur $\mathcal{E}_i$. L'erreur de l'apprentissage est estimée par la moyenne des erreurs des h_i. Quelles précautions faut-il prendre lorsqu'on partitionne $\mathcal{E}$?
4. Le leave-one-out (LOO) généralise la validation croisée : si $\mathcal{E}$ contient n exemples, on définit n sous-ensembles. L'estimation de l'erreur par LOO par rapport à la validation croisée est-elle
   — plus optimiste,
   — plus pessimiste
   — ni l'un ni l'autre ?
   Pourquoi ?
5. Votre algorithme est le 1 plus proche voisin. Quelle est l'erreur en LOO dans le cas ci-dessous ?

   Est-ce mieux de prendre un 3-plus proche voisin ?

2 Arbres de décision (5 points)

Vous avez observé des objets de deux classes ($+$ et $-$).

forme	taille	couleur	classe
rond	petit	blanc	$+$
carré	petit	rouge	$-$
rond	petit	vert	$+$
carré	moyen	blanc	$+$
rond	petit	rouge	$-$
rond	petit	jaune	$+$
rond	moyen	jaune	$-$
rond	moyen	blanc	$+$
carré	grand	blanc	$+$
carré	grand	rouge	$-$
carré	grand	vert	$-$

— Quelle quantité d'information est donnée par *couleur* = *blanc* ?
— Quel est le gain d'information de chaque attribut ?
— Déroulez l'algorithme d'arbre de décision et montrez l'arbre obtenu.
— Transformer l'arbre obtenu en une base de règles.
Inversement, peut-on transformer toute base de règles en un arbre de décision ?

3 Réduction de problèmes d'apprentissage (3 points)

3.1 Classification binaire $\rightarrow$ classification multi-classe

Vous disposez d'un algorithme $\mathcal{A}$ d'apprentissage binaire (applicable quand les labels sont $\{+1, -1\}$). Votre problème d'apprentissage comprend 3 classes. Construisez un algorithme $\mathcal{B}$ pour apprendre une hypothèse à 3 classes en utilisant $\mathcal{A}$ (plusieurs réponses possibles).
Supposons que l'erreur moyenne commise par $\mathcal{A}$ soit de 10%. Peut-on borner l'erreur de $\mathcal{B}$? Comment l'erreur évolue (grandir ou diminuer) si le nombre de classes augmente ?

3.2 Classification une-classe

On dispose d'un ensemble d'instances non supervisé $\mathcal{G} = \{x_1, \ldots, x_n, x_i \in \mathbb{R}^d\}$ tirés selon une distribution P, et on suppose qu'on a un algorithme capable à partir de $\mathcal{E}$ d'apprendre une distribution de probabilités Q sur $\mathbb{R}^d$:

$$Q : \mathbb{R}^d \mapsto [0, 1] \; t.q. \int Q(x)dx = 1$$

Comment évaluer la qualité de Q ?

On dispose maintenant d'un ensemble supervisé, $\mathcal{E} = \{(x_i, y_i), x_i \in \mathbb{R}^d, y_i \in \{-1, 1\}\}$. Comment se servir de l'algorithme non supervisé ci-dessus pour séparer les classes $+1$ et -1 ?
Indication : on pourra commencer par supposer que les deux classes sont de même taille. On rappellera ensuite le théorème de Bayes, et on traitera le cas général en supposant qu'on connait la probabilité *a priori* des deux classes.

Partie II : Apprentissage probabiliste

4 Bayésien naïf (5 points)

Nous souhaitons mettre en place un système qui à partir d'un mail, détermine son degré d'importance. Nous supposons qu'il y a 3 degrés d'importance (0 : à lire un jour ; 1 : à lire aujourd'hui ; 2 à lire maintenant). Pour cela, on considère un vocabulaire $\mathcal{V}$ contenant V mots. Un mail est alors représenté par un vecteur de V caractéristiques binaires :

$$\mathbf{x} = (x_1, ..., x_i, ..., x_V), \text{ avec } x_i = 1 \text{ ou } 0$$

La composante x_i indique si le $i$ème mot du vocabulaire est présent (1) ou absent (0). La classe associée à un mail est notée y et peut prendre 3 valeurs. Nous supposons que **x** et y sont respectivement la réalisation du vecteur aléatoire $\mathbf{X}$ et de la variable aléatoire Y. Nous souhaitons réaliser ce système grâce à un classifieur Bayésien.

1. Ecrire la règle de décision dite du "maximum a posteriori" dans le cadre général et nommer chacun des termes utilisés.
2. Chaque composante de **x** étant indépendante, nous proposons d'écrire :

$$P(X = \mathbf{x}|Y = y) = \prod_{i=1}^{i=V} \theta_{i,y}^{x_i}(1 - \theta_{i,y})^{(1-x_i)} \tag{1}$$

   — Expliquer brièvement cette formule et à quoi correspondent les $\theta_{i,y}$ (vous pouvez prendre par exemple un exemple avec un vocabulaire fictif de 3 mots et $\mathbf{x} = (1, 0, 1)$)
   — Quel nom porte ce type de distribution ?
3. Combien de paramètres sont nécessaires pour ce modèle ? Exprimer ce nombre en fonction de V.
4. Nous disposons d'un ensemble d'exemples d'apprentissage. Pour cela nous considérons connus : tous les compteurs de la forme $c(i, y)$ qui indique dans combien de mails de la classe y le mot i est présent ; le nombre total d'exemples N et le nombre d'exemples pour chacune des classes N_k. Comment calculer les paramètres de la question précédente ? Donner la formule d'estimation selon le maximum de vraisemblance.
5. Est-il nécessaire de calculer le terme $P(\mathbf{X} = \mathbf{x})$ pour décider ? Pourquoi ?
6. Si un mot n'apparait pas dans les textes d'une classe, quelle est la vraisemblance pour cette classe ? Est-ce acceptable ? Et le cas échéant quel remède proposez-vous ?

5 Algorithme E.M (3 points)

Soit un ensemble de mails décrits comme dans l'exercice précédent, à la différence que désormais nous ne disposons pas de la classe des mails. Nous souhaitons regrouper ces mails en $k = 3$ classes de manière non-supervisée. Nous allons pour cela utiliser l'algorithme E.M vu en cours, en considérant qu'un cluster n'est pas modélisé par une loi gaussienne, mais selon l'équation 1.

Expliquer comment mettre en œuvre l'algorithme de clustering E.M (plusieurs possibilités). Suggestion :

— On pourra supposer que la distribution des données est un mélange de distributions (voir Eq. 1). Expliciter la probabilité d'un exemple selon ce modèle de mélange.
— En déduire l'ensemble des paramètres à apprendre.
— Poser le calcul permettant de calculer la probabilité d'affectation d'un exemple.
— Détailler rapidement les 2 étapes de l'algorithme E.M.

6 Les K-moyennes (3 points)

Considérons l'espace vectoriel $\mathbb{R}^2$ muni du produit scalaire canonique et donc de la distance Euclidienne. On dispose des points d'apprentissage suivants :

$$\left\{\begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 3 \\ 2 \end{pmatrix}, \begin{pmatrix} 4 \\ 4 \end{pmatrix} \begin{pmatrix} 5 \\ 5 \end{pmatrix}, \begin{pmatrix} 6 \\ 9 \end{pmatrix}, \begin{pmatrix} 7 \\ 8 \end{pmatrix}\right\}$$

Appliquer l'algorithme des k-moyenne avec $k = 3$ sur ce jeu de données (suggestion : faire un dessin). Détailler la mise en place de l'algorithme ainsi que l'intialisation et la première itération de l'algorithme.