INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES ROUEN

Data-Mining
2018/2019

DÉPARTEMENT ARCHITECTURE DES SYSTÈMES D'INFORMATION

— Durée : 2h
— *Documents autorisés : cours, codes des TD machine*
— **A rendre : pour chaque exercice, une archive contenant le script Matlab ou Python avec toutes les fonctions nécessaires pour l'exécuter**

1 Regroupement de pays (7 points)

Le fichier `datacountries.mat` contient des données d'un sondage de l'ONU auprès de 194 pays. Le sondage portait sur les sujets politiques prioritaires comme le changement climatique, la santé, l'emploi, l'égalité... La matrice $X \in \mathbb{R}^{N\times 16}$ comprend les données et la variable `pays` le nom des pays.

1. Écrire un programme qui réalise, à partir de X, le regroupement des pays en $K = 5$ clusters par la méthode des K-Means. Les centres initiaux seront les K premières lignes de X.
2. On veut visualiser les clusters via l'ACP. Compléter votre programme pour :
   — calculer la matrice de projection P de l'ACP,
   — projeter les données en 2D en utilisant P,
   — visualiser la projection en 2D des points des clusters avec une couleur/un symbole différent pour chaque cluster.
   — annoter chaque point avec le nom du pays correspondant (on utilisera la fonction `text(posx, posy, texte)` de matlab ou matplotlib.pyplot)

2 Vache espagnole (13 points)

Pour évaluer leur niveau en anglais, des étudiants ont écrit un texte. De chaque texte on extrait des indicateurs (orthographique, lexique, grammaire, vocabulaire...). On a ces données pour les étudiants de niveau B2 (label $y = 1$) et ceux de niveau A2 (label $y = -1$). Les fichiers `englevelapp.mat, englevelval.mat, engleveltest.mat` contiennent les données et labels et serviront respectivement pour l'apprentissage, la validation et le test. On cherche un SVM linéaire $f(x) = \mathbf{w}^\top \mathbf{x} + b$. .

1. Donner le nombre de points par classe. Que constatez-vous ?
2. Soit les coûts de mauvaise classification $L_{FP} = 2$ et $L_{FN} = 7$. **Le critère à utiliser pour sélectionner le modèle** sera l'erreur de classification pondérée

$$E = \frac{L_{FP} \times FP + L_{FN} \times FN}{N}$$

   $FP =$ nombre de faux positifs, $FN =$ nombre de faux négatifs, $N =$ nombre de points.
   Ecrire une fonction $E =$`moncout(`$Y_{\texttt{vrai}}, Y_{\texttt{predit}}$`,` L_{FP}`,` L_{FN}`)` qui calcule la matrice de confusion et en déduit E.
3. Ecrire un programme qui réalise les opérations suivantes :
   — apprendre un modèle SVM linéaire. On mettra en évidence la sélection du "meilleur modèle" (en utilisant la fonction précédente),

— évaluer ce SVM sur les données d'apprentissage et de test en affichant les matrices de confusion et les erreurs de classification pondérées E.

4. Supposons qu'on souhaite faire un SVM asymétrique

$$\begin{array}{lll} \min_{w\in\mathbb{R}^d, b\in\mathbb{R}, \xi_i} & \frac{1}{2}\|w\|^2 + \sum_{i=1}^N C_i\xi_i & \\ \text{s.c.} & y_i\left(w^\top x_i + b\right) \geq 1 - \xi_i & \forall \quad i = 1, \ldots, N \\ & \xi_i \geq 0 & \forall \quad i = 1, \cdots, N \end{array}$$

où

$$C_i = \left\{ \begin{array}{ll} C \times r & \text{si} \quad y_i = -1 \\ C \times (1-r) & \text{si} \quad y_i = 1 \end{array} \right.$$

et $r = \frac{N_+}{N}$ avec $N_+ =$ nombre de points dont le label est $y = 1$.
Proposer une méthodologie pour sélectionner le meilleur modèle correspondant à ce SVM (*Nota : aucun code n'est demandé*).