Ph. Leray
A. Rakotomamonjy

Data-Mining
Examen 2003/2004

Architecture des Systèmes d'Information
4^{eme} année

- Durée : 3h
- Calculatrice autorisée
- Documents autorisés :
  - supports de cours et notes personnelles
  - votre voisin n'est pas un document
- Barême indicatif sur 20 points
- **Les parties (1-2) et (3-4) sont à rendre sur des feuilles séparées**

1 SVM et Lagrangien (7 points)

On cherche à résoudre un problème de maximisation de marge avec des données non- séparable, en introduisant une variable de relachement qui est pénalisée quadratiquement.

$$\min_{w,b,\xi_i} \tfrac{1}{2}\|w\|^2 + \tfrac{C}{2}\textstyle\sum_i \xi_i^2$$
sous contraintes
$$y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i$$

1. Expliquer quel est le role de C
2. Justifier pourquoi les contraintes de type $\xi_i \geq 0$ n'ont plus lieu d'etre
3. Montrer que le Lagrangien associé à ce problème s'écrit :

$$\mathcal{L}(w, b, \xi_i) = \frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_i \xi_i^2 - \sum_{i=1}^{n} \alpha_i(y_i(\langle w, x_i\rangle + b) - 1 + \xi_i)$$

4. En déduire les conditions d'optimalités associés à w, b et ξ_k
5. En déduire l'expression du problème dual.
6. Démontrer que la marge géométrique γ s'écrit :

$$\gamma = \left(\sum_{i\in sv} \alpha_i^* - \frac{1}{C}\sum_{i\in sv}(\alpha_i^*)^2\right)^{-1/2}$$

où α_i^* sont les solutions du problème dual.

2 Classification (5 points)

On cherche à établir une règle de discrimination à partir des données de $\mathbb{R}^2$ deux classes suivant :

$$x^+ = \begin{pmatrix} 1 & 1 \\ 1.5 & 0.5 \\ 0.8 & 1.2 \\ 1.5 & -0.2 \\ 0.5 & -0.3 \end{pmatrix} \quad x^- = \begin{pmatrix} -1 & -0.8 \\ -0.2 & 0.5 \\ 0.2 & -1 \\ -0.5 & -0.2 \\ 0.3 & -0.5 \end{pmatrix} \quad x_T = \begin{pmatrix} 0.5 & 0.5 \\ 0.2 & -0.3 \\ -0.2 & 0.3 \\ -0.8 & -0.2 \\ 0.3 & 1 \end{pmatrix}$$

où $x+$ représentent les 5 exemples de la classe positive ($y = +1$) et x^- les 5 exemples de la classe négative ($y = -1$) et x_T les points à tester

1. Classer les données x_T suivant la règle des 1 et 3 plus-proches voisins en utilisant la distance issue de la norme infinie.
2. Si on considère que les deux classes proviennent de deux lois normales de meme variance mais de moyennes différentes, attribuer un label à chaque point x_T de façon à minimiser l'erreur de classification, en supposant que la matrice de covariance des lois est la matrice identité I et que les deux classes sont équiprobables.

3 Classification hiérarchique ascendante (4 points)

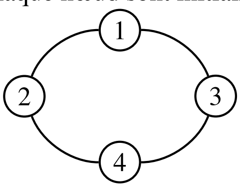
On considère une partie des données de l'exercice précédent, en oubliant la variable classe y :

$$\mathcal{D} = \begin{pmatrix} 1 & 1 \\ 1.5 & 0.5 \\ 0.8 & 1.2 \\ -1 & -0.8 \\ -0.2 & 0.5 \\ 0.2 & -1 \end{pmatrix}$$

1. Soit la mesure suivante entre deux points a et b de $\mathcal{D}$: $\mathcal{M}(a,b) = \max_i |a_i - b_i|$. Est-ce une mesure de similarité ou de dissimilarité ?
2. On utilise le diamètre maximum comme mesure de distance entre 2 clusters. Comment peut-on calculer facilement la nouvelle matrice de mesures M après fusion de 2 clusters ?
3. Effectuer une classification hiérarchique ascendante sur $\mathcal{D}$ et tracer le dendogramme correspondant

4 Cartes auto-organisatrices (4 points)

Soit la carte de Kohonen de la figure suivante, avec 4 nœuds et un voisinage unidimensionnel en anneau. Les poids de chaque nœud sont initialisés aux valeurs suivantes :



$$W_1 = \begin{pmatrix} -1 \\ -1.5 \\ 0.5 \end{pmatrix} \quad W_2 = \begin{pmatrix} 2 \\ -2 \\ 5.2 \end{pmatrix} \quad W_3 = \begin{pmatrix} 1.5 \\ 6 \\ 4.3 \end{pmatrix} \quad W_4 = \begin{pmatrix} -4 \\ 7 \\ 0.6 \end{pmatrix}$$

Les données à notre disposition sont $a = \begin{pmatrix} -1.4 \\ 2.3 \\ 0.2 \end{pmatrix}$ $b = \begin{pmatrix} -4 \\ 2.5 \\ 4 \end{pmatrix}$

1. Appliquer les 3 premières itérations de l'algorithme d'apprentissage compétitif des SOM avec les caractéristiques suivantes :
   - itération 1 : présentation de a puis b avec un pas d'apprentissage de 0.5 et un voisinage incluant tous les nœuds,
   - itération 2 : présentation de b puis a avec un pas d'apprentissage de 0.25 et un voisinage incluant les voisins immédiats de chaque nœud vainqueur,
   - itération 3 : présentation de a puis b avec un pas d'apprentissage de 0.125 et un voisinage restreint au neurone vainqueur.
2. Où sont projetées les données de test suivantes : $c = \begin{pmatrix} -1.6 \\ 3 \\ 0.5 \end{pmatrix}$ $d = \begin{pmatrix} -3 \\ 3 \\ 5 \end{pmatrix}$ $e = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$?