-2015

TD 8

Exercice 1 – Généralités sur le clustering

Q 1.1 Supposons que l'on possède un corpus de N documents, combien de clustering différents peuvent être trouvés si on cherche K clusters ?

Q 1.2 Considérons que les N exemples appartiennent à K catégories différentes et que l'on connait la fonction d'appartenance (on sait, pour chaque exemple, à quelle catégorie il appartient, comme dans le cas de l'apprentissage supervisé). Proposer une mesure qui permette l'évaluation d'un système de clustering

Exercice 2 – Deux algorithmes de Quantification vectorielle

Algorithme des K-moyennes
La taille du dictionnaire K est fixée, c'est un paramètre de l'algorithme.

- Initialiser aléatoiremet les K prototypes.
- Répéter jusqu'à (critère d'arrêt) :
  - partitionner les exemples en les affectant aux prtotypes dont ils sont le plus proche ;
  - redéfinir les protoypes (i.e. centres de gravité des partitions).

Algorithme LBG (Linde-Buzo-Gray)

- y un vecteur initialisé aléatoirement de petite norme ;
- le dictionnaire est initialisé à un élément, le centre de gravité des exemples de l'ensemble d'apprentissage ;
- répéter jusqu'à (critère d'arrêt) :
  - dédoubler chaque prototype x du dictionnaire en $x + y$, $x - y$;
  - partitionner les exemples en les affectant aux prototypes dont ils sont le plus proche ;
  - redéfinir les prototypes (i.e. centres de gravité des partitions).

Q 2.1 Soit l'ensemble d'exemples en dimension 2 :
$D = \{(1, 3), (-2, 2), (-1, 2), (-3, 2), (1, 5), (1, 4), (2, 5), (2, 6)\}$
Faire tourner l'algorithme des K-moyennes en prenant pour dictionnaire de départ les deux centres de gravité des 4 premiers et dernièrs exemples.

Q 2.2 Faites tourner l'algorithme LBG sur les données D.

Q 2.3 Quels critères d'arrêt préconisez-vous pour les méthodes de QV ?

Exercice 3 – Carte auto-organisatrice (SOM)

Une carte auto-organisatrice est un réseau de neurones conçus pour l'apprentissage non-supervisé. Dans le cas d'une carte 2D, la carte est une grille de neurones $n_{i,j}$ tel que chaque neurone possède un vecteur de poids noté $w_{i,j}$. Nous considérerons une relation de voisinage entre les neurones $V : N \times N \to vrai, faux$ telle que $v(n_{i,j}, n_{k,l}) = vrai$ ssi $n_{i,j}$ et $n_{k,l}$ sont voisins. Typiquement, les voisins du neurone $n_{i,j}$ sont les neurones $n_{i-1,j}, n_{i+1,j}, n_{i,j-1}, n_{i,j+1}$. Chaque neurone représente une des classes de notre problème de clustering.

L'algorithme est le suivant :

- initialiser les poids des neurones aléatoirement.

- Répéter :
  — Pour chaque exemple x_k
    — choisir le neurone le plus proche : $argmin||x_k - w_{i,j}||$
    — pour tous les neurones du voisinage (y compris lui-même), actualiser les poids en utilisant la règle : $\Delta w_{i,j} = \epsilon(t) * (x_k - w_{i,j})$.

Faire tourner l'algorithme en utilisant une grille de en carré de 4 neuronnes de coordonnées $(-1, -1)$, $(1, -1)$, $(-1, 1)$, $(1, 1)$.

Exercice 4 – Clustering et mélange de lois

On souhaite estimer une densité de probabilité par un modèle de type mélange de gaussiennes. La probabilité d'une observation x est donnée par : $p(x) = \sum_{l=1}^{L} P_l.p(x|\lambda_l)$ où les P_l sont les probabilités a priori des lois et les $p(x|\lambda_l)$ sont des lois gaussiennes caractérisées par leur moyenne μ_l et leur variance σ_l , i.e. $\lambda_l = (\mu_l, \sigma_l)$.

Q 4.1 Dessiner la loi de probabilité pour $L = 2, P_1 = P_2 = 0.5$, et $\mu_1 = 1, \mu_2 = 3, \sigma_1 = 1, \sigma_2 = 10$.

Q 4.2 Quelles est la probabilité a posteriori qu'un exemple x aie été produit par la gaussienne l, $p(\lambda_l|x)$?

Q 4.3 Expliquer comment l'apprentissage d'un mélange de lois peut être utilisé pour faire du clustering.

Exercice 5 – Apprentisage d'un mélange de lois et maximum de vraissemblance

On souhaite apprendre le modèle de l'exerceice précedent avec un critère de maximum de vraisemblance (MV) sur une base d'apprentissage $E = \{x_i\}, i = 1..N$.

Q 5.1 Exprimer le logarithme de la vraisemblance des données par le modèle en supposant que les x_i sont indépendants.

Q 5.2 Montrer que maximiser ce logarithme ou la vraisemblance directement doit aboutir théoriquement à la même solution.

Q 5.3 On utilise un algorithme dit algorithme EM pour l'estimation de ce mélange de gaussiennes. Voici une des variantes de cet algorithme :

- initialiser les paramètres $(P_i, \mu_i, \sigma_i)_{i=1..L}$;
- Répéter :
  — déterminer pour chaque x_i la gaussienne qui l'a poduit avec la plus grande vraisseblance : pour $i = 1..N$, $I(x_i) = argmax_l\ p(\lambda_l|x_i)$;
  — ré-estimer les paramètres des lois à partir des exemples qui lui ont été affectés : pour $l = 1..L$, ré-estimer λ_l à partir des $\{x_i \in E|I(x_i) = l\}$

Q 5.4 Dans le cas où les matrices de covariance des lois sont fixées à l'identité, montrer que l'algorithme précédent est équivalent à un algorithme des K-Moyennes.