Mastere M2 MVA 2010/2011 - Modèles graphiques

Exercices à rendre pour le 17 novembre 2010.

Ces exercices peuvent s'effectuer par groupe de deux élèves.

1 Distributions factorisables dans un graphe

(a) Prouver les propositions suivantes énoncées en cours :
- Soit $G = (V, E)$ un DAG et soit $i \to j$ une *arête couverte*, i.e. telle que $\pi_j = \pi_i \cup \{i\}$, soit $G' = (V, E)$, avec $E' = (E \backslash \{i \to j\}) \cup \{j \to i\}$, alors on a $\mathcal{L}(G) = \mathcal{L}(G')$.
- Soit G un arbre orienté sans structure en v et G' l'arbre non-orienté de même ensemble d'arêtes (son symmétrisé), alors $\mathcal{L}(G) = \mathcal{L}(G')$.

(b) Donnez le plus petit graphe non-orienté G tel qu'il n'existe pas de graphe orienté G' tel que $\mathcal{L}(G) = \mathcal{L}(G')$.

2 d-Séparation

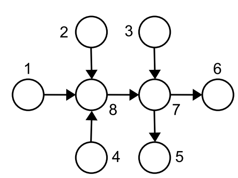
(a) Pour quels graphes orientés acycliques la d-séparation se réduit-elle à la séparation usuelle dans le graphe symmétrisé ?

(b) Soit G un DAG et G_M le graphe moralisé. Soit A, B, S trois ensembles de noeuds disjoints montrer que si A et B sont séparés par S dans G_M, ils sont d-séparés par S dans G.

(c) Soit A, B, S et T des sous-ensembles de noeuds non-vides d'un DAG G. Si A et B sont d-séparés par S et A et S sont d-séparés par T, est-ce que A et B sont d-séparés par T ?

(d) Soit le graphe G ci-dessous. Quelles assertions sont vraies pour toute distribution $p \in \mathcal{L}(G)$ parmi les suivantes :

- $X_{\{1,2\}} \perp\!\!\!\perp X_4 \mid X_3$
- $X_{\{1,2\}} \perp\!\!\!\perp X_4 \mid X_5$
- $X_1 \perp\!\!\!\perp X_6 \mid X_{\{2,4,7\}}$



3 Implémentation - mélange de Gaussiennes

Le fichier "EMGaussienne.dat" contient un ensemble de données (x_n, y_n) où $(x_n, y_n) \in \mathbb{R}^2$. Le but de cet exercice est d'implémenter l'algorithme EM pour certains mélanges de K Gaussiennes dans $\mathbb{R}^d$ (dans cet exercice, $d = 2$ et $K = 4$), avec des données IID. (NB : dans cet exercice, il n'est pas nécessaire de démontrer les formules utilisées).

Le langage de programmation est libre (MATLAB, Octave, Scilab ou R sont néanmoins recommandés). Le code source doit être remis avec les résultats. Le corrigé sera codé en MATLAB.

(a) Implémenter l'algorithme K-means (chapitre 10 du polycopié). Représenter graphiquement les données d'apprentissage, les centres obtenus et les différents groupes ("clusters"). Essayer plusieurs initialisations et comparer les résultats (centres et mesures de distortions).

(b) Implémenter l'algorithme EM pour un mélange de Gaussiennes avec des matrices de covariance proportionnelles à l'identité (initialiser l'algorithme EM avec les moyennes trouvées par K-means).

Représenter graphiquement les données d'apprentissage, les centres et les covariances obtenus (une manière élégante de représenter graphiquement est de représenter l'ellipse contenant un certain pourcentage (e.g., 90%) de la masse de la Gausienne). Estimer et représenter la variable latente pour chaque point (pour le jeu de paramètres appris par EM).

(c) Implémenter l'algorithme EM pour un mélange de Gaussiennes avec des matrices de covariance générales. Représenter graphiquement les données d'apprentissage, les centres et les covariances obtenus. Estimer et représenter la variable latente pour chaque point (pour le jeu de paramètres appris par EM).

(d) Commenter les différents résultats obtenus. En particulier, comparer les log-vraisemblances des deux modèles de mélanges, sur les données d'apprentissage, ainsi que sur les données de test (dans "EMGaussienne.test").