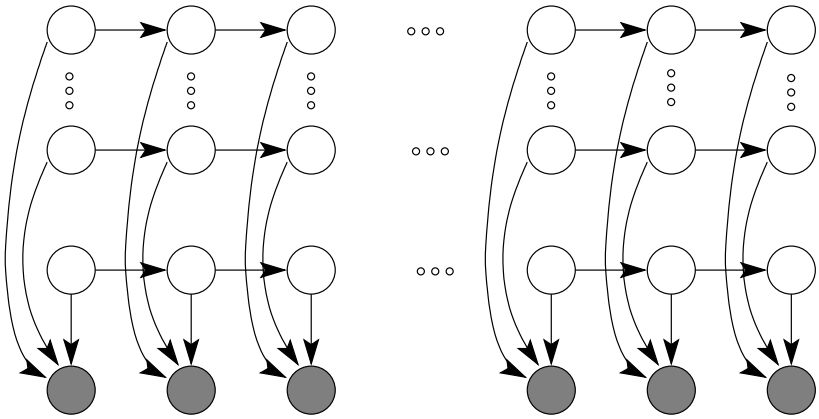
Cours MVA 2005 - Modèles graphiques

Exercices à rendre pour le 13 Décembre 2005.

1 Modèles de Markov cachés factoriels

On considère le modèle de Markov caché factoriel suivant, avec M chaînes de Markov cachées de longueur T.



1. Quelle est la complexité de l'algorithme de l'arbre de jonction pour l'inférence? (on supposera que la i-ième chaîne a m_i états).
2. Quelle serait la complexité en utilisant (naivement) une seule chaîne de Markov avec $\prod_i m_i$ états? (si $x_1^t, \ldots, x_M^t$ sont les états des M chaînes à l'instant t, l'état de cette chaîne est $x^t = (x_1^t, \ldots, x_M^t)$).

2 Mélange d'analyses factorielles

Afin de modéliser des données vectorielles $x \in \mathbb{R}^d$, on considère un modèle de mélange de M modèles d'analyses factorielles (avec chacun le même nombre de facteurs K).

1. Quel est le modèle graphique associé à ce modèle?
2. A quoi peut servir ce modèle?
3. Quels en sont les paramètres?
4. Calculer les équations d'estimation (E-step et M-step) par l'algorithme EM.

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 et R sont néanmoins recommandés, R peut être téléchargé gratuitement à partir de `http://www.r-project.org/`). Le code source doit être remis avec les résultats.

1. 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).
2. 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'apprentisage, 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).
3. Implémenter l'algorithme EM pour un mélange de Gaussiennes avec des matrices de covariance générale Représenter graphiquement les données d'apprentisage, les centres et les covariances obtenus. Estimer et représenter la variable latente pour chaque point (pour le jeu de paramètres appris par EM).
4. 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").

4 Implémentation - HMM

On considère les mêmes données d'apprentissage dans le fichier "EMGaussienne.dat", mais cette fois-ci en considérant la structure temporelle, i.e., les données sont de la forme $u_t = (x_t,y_t)$ où $u_t = (x_t,y_t) \in \mathbb{R}^2$, pour $t = 1,\ldots,T$. Le but de cet exercice est d'implémenter l'inférence dans les HMM ainsi que l'algorithme EM pour l'apprentissage des paramètres.

On considère le modèle HMM suivant avec une chaîne (q_t) à K=4 états et matrice de transition $a \in \mathbb{R}^{4\times4}$, et des "probabilités d'émission" Gaussiennes: $u_t|q_t = i \sim \mathcal{N}(\mu_i,\Sigma_i)$.

1. Implémenter les récursions α et β vues en cours et dans le polycopié pour estimer $p(q_t|u_1,\ldots,u_T)$ et $p(q_t,q_{t+1}|u_1,\ldots,u_T)$.
2. Calculer les équations d'estimation de EM.
3. Implémenter l'algorithme EM pour l'apprentissage (on pourra initialiser les moyennes et les covariances avec celles trouvées en 3.3).
4. Implémenter l'inférence pour estimer la séquence d'états la plus probables, i.e. $\arg\max_q p(q_1,\ldots,q_T|y_1,\ldots,y_T)$, et représenter le résultat obtenu avec les données (pour le jeu de paramètres appris par EM).
5. Commenter les différents résultats obtenus (exercices 3 et 4). En particulier, comparer les log-vraisemblances, sur les données d'apprentissage, ainsi que sur les données de test (dans "EMGaussienne.test").

5 Projet de fin de cours

Décrire le projet en quelques lignes (thème, articles de référence, implémentations prévues, type de données, etc...).