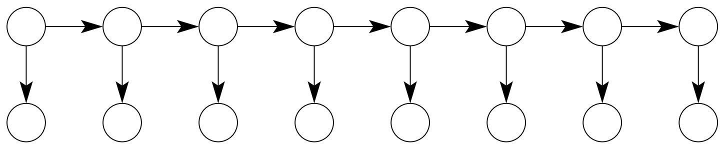
Mastère M2 MVA 2007/2008 - Modèles graphiques

Exercices à rendre pour le 7 Novembre 2007.

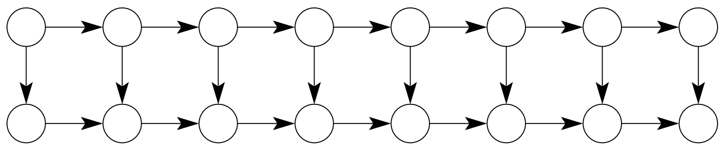
Ces exercices peuvent s'effectuer par groupe de deux élèves.

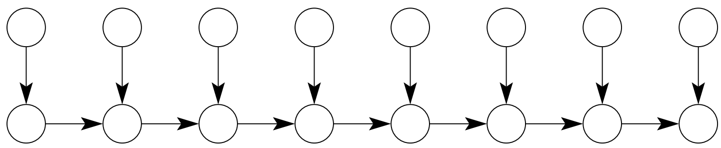
1 Elimination pour modèles dynamiques

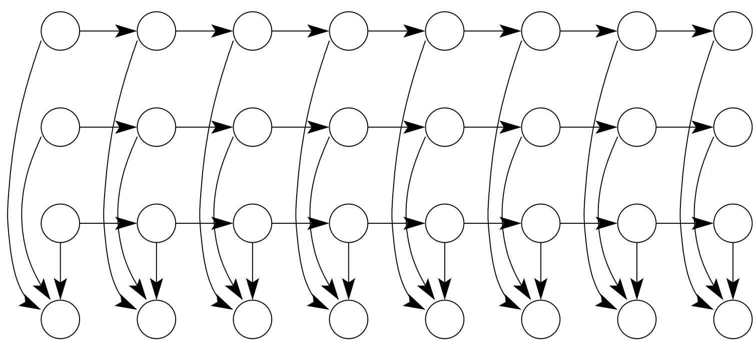
Pour chacun des 4 graphes suivants,
- Déterminer la largeur arborescente (treewidth) après moralisation. Preuve non exigée.
- Exhiber un ordre d'élimination et le graphe reconstitué correspondants



(exemple de modèle de Markov caché)







(exemple de modèle de Markov caché *factoriel*)

2 Apprentissage dans les modèles discrets

On considère le modèle suivant où z et x sont des variables discrètes pouvant prendre respectivement M et K valeurs : $p(z = m) = \pi_m$, $p(x = k|z = m) = \theta_{mk}$.

Calculer l'estimateur du maximum de vraisemblance de π et θ à partir de données i.i.d.

3 Classification linéaire

Le fichier "classification.dat" contient un ensemble de données (x_n, y_n) où $x_n \in \mathbb{R}^2$ et $y_n \in \{0, 1\}$. Le but de cet exercice est d'implémenter les méthodes de classification linéaire. Le langage de programmation est libre (MATLAB, Octave, Scilab ou R sont néanmoins recommandés). Le code source doit être remis avec les résultats.

1. Modèle génératif (LDA). Etant donnée la classe, les données sont normales avec des moyennes différentes et la même matrice de covariance :

   $$y \sim \text{Bernoulli}(\pi), \quad x|y = i \sim \text{Normale}(\mu_i, \Sigma).$$

   Calculer et implémenter le maximum de vraisemblance pour ce modèle et l'appliquer aux données. Représenter graphiquement les données ainsi que la droite définie par

   $$p(y = 1|x) = 0.5$$

   Indication : le modèle a été vu en cours mais pas le calcul du maximum de vraisemblance. On pourra s'inspirer de la Section 7.2 du polycopié (qui traite le cas où Σ est diagonale).

2. Régression logistique : implémenter la regression logistique en utilisant l'algorithme IRLS (Newton-Raphson) décrit en cours et dans le polycopié (ne pas oublier le terme constant). Représenter graphiquement les données ainsi que la droite définie par

   $$p(y = 1|x) = 0.5$$

3. Régression lineaire : en considérant la classe y comme variable réelle prenant les valeurs 0 et 1, implémenter la regression linéaire par résolution de l'équation normale. Représenter graphiquement les données ainsi que la droite pour laquelle la fonction de regression vaut 0.5.

4. Les données contenues dans le fichier "classification.test" ont été générées par la même distribution que les données dans "classification.dat". Tester les différents modèles sur ces données en calculant le taux d'erreur, et comparer les résultats.

5. BONUS : Modèle génératif (QDA). Etant donnée la classe, les données sont normales avec des moyennes et des matrices de covariance différentes :

$$y \sim \text{Bernoulli}(\pi), \quad x|y = i \sim \text{Normale}(\mu_i, \Sigma_i).$$

   Implémenter le maximum de vraisemblance pour ce modèle et l'appliquer aux données. Représenter graphiquement les données ainsi que la conique définie par

$$p(y = 1|x) = 0.5$$

4 BONUS : Equivalence factorisation/indépendances conditionnelles pour les modèles graphiques orientés

Soit $G = (V, E)$ un DAG et X un ensemble de variables aléatoires discrètes indexées par V et de loi jointe $p(x)$. Montrer par récurrence sur $\#(V)$ la proposition suivante :

Proposition 1 *Si pour tout $v \in V$, $X_v \perp X_{\text{non-descendants}(v)} | X_{\text{parents}(v)}$, alors $p(x)$ se factorise dans $L(G)$.*