# Accès à l'information - Homeworks

François Yvon

## Représentations : Variantes du bayésien naïf

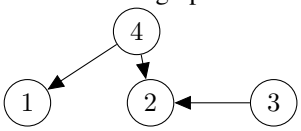
1. Refaire tous les calculs du cours pour le cas où les distributions *a priori* ne sont pas uniformes (on supposera deux classes, $\mathrm{P}(y)$ suit une loi de Bernoulli paramétrée par $\alpha$).

Dériver les estimateurs pour le Bayésien naïf dans le cas où :

1. chaque document $\boldsymbol{x}$ est représenté par le vecteur de compte : $x_w$ correspond au nombre d'occurrences de $w$ dans $\boldsymbol{x}$. En notant $l_{\boldsymbol{x}}$ le nombre total d'occurrences, on peut modéliser $\boldsymbol{x}$ comme le résultat de $l_{\boldsymbol{x}}$ tirages d'une loi multinomiale paramétrisée par $\boldsymbol{\theta}$ (de dimension $n_w$).
   (a) quel est l'estimateur ML pour $\boldsymbol{\theta}$ ?
   (b) quel est l'estimateur MAP pour $\boldsymbol{\theta}$ (on choisira la loi Dirichlet, qui est la loi conjuguée de la loi multinomiale, comme loi *a priori*) ?
   (c) quelle est la loi prédictive ?
2. idem lorsque l'on considère que le vecteur de comptes $\boldsymbol{x}$ résulte de $n_w$ tirages, chacun dans une loi de Poisson paramétrée par $\theta_w$ (optionnel)

## Manipuler des MG

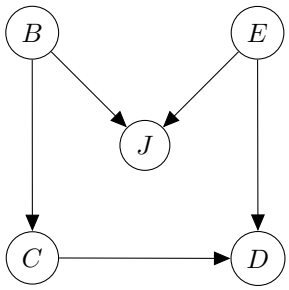
1. Considérez le graphe suivant :




   Écrire la factorisation de la loi jointe induite par ce graphe. Montrez par le calcul que $X_1 \perp\!\!\!\perp X_2 \,|\, X_4$

## Analyse d'un modèle graphique

On considère le graphe suivant :



avec la sémantique suivante (dans le domaine des voitures) : $B$ à la charge de la batterie (**p**lein ou **v**Ide), $J$ correspond au niveau de la jauge d'essence (**p**lein ou **v**Ide), $E$ au contenu du réservoir (**p**lein ou **v**Ide), $C$ indique si le contact se fait (**o/n**), et $D$ si la voiture démarre (**o/n**).

1. Écrire la factorisation de la loi jointe.
2. En utilisant les valeurs numériques ci-dessous, calculez la probabilité que le réservoir soit vide ($E = v$) quand la voiture ne démarre pas.

Valeurs numériques :

| | |
|---|---|
| $\mathrm{P}(B = v) = 0.02$ | $\mathrm{P}(E = v) = 0.05$ |
| $\mathrm{P}(G = v \mid B = p, E = p) = 0.04$ | $\mathrm{P}(G = v \mid B = p, E = v) = 0.97$ |
| $\mathrm{P}(G = v \mid B = v, E = p) = 0.1$ | $\mathrm{P}(G = v \mid B = v, R = v) = 0.99$ |
| $\mathrm{P}(C = n \mid B = v)$=0.98 | $\mathrm{P}(C = n \mid b = \mathit{good}) = 0.03$ |
| $\mathrm{P}(D = n \mid C = y, E = p) = 0.01$ | $\mathrm{P}(D = n \mid C = n, E = p) = 1.0$ |
| $\mathrm{P}(D = n \mid C = y, E = v) = 0.92$ | $\mathrm{P}(D = n \mid C = n, E = v) = 0.99$ |

## Regarder des films

Par exemple Daphne Koller sur Coursera : `[https://fr.coursera.org/course/pgm/lecture]` (les fondamentaux des modèles graphiques, au moins les 5 premiers extraits).

## Lire un article

Pour la prochaine séance, (essayer de) lire jusqu'à la section 4 :
Thomas Hofmann. *U*nsupervised Learning by Probabilistic Latent Semantic Analysis. Machine Learning 42(1/2) : 177-196 (2001)

Téléchargeable ici : `http://www.cs.helsinki.fi/u/vmakinen/stringology-k04/hofmann-unsupervised_learning_by_probabilistic_latent_semantic_analysis.pdf`