Accès à l'information - Préparation du cours 2

François Yvon

0.1 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 Bernoulli paramétré par α).

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 $vtheta$ (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 θ_w.

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 :
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`