Méthodes Empiriques et Langages de Script
TP 7

G. A. Musillo
musillo4@etu.unige.ch

December 19, 2005

1 Exercice 1: un classifieur Naive Bayes pour la classification de documents textuels

On vous demande de programmer un classifieur bayésien naïf capable d'apprendre à classer des documents textuels. Le pseudo-code que vous devez implémenter en Perl vous a récemment été présenté par Paola. Vous pouvez consulter le document (`http://www.latl.unige.ch/mels/cours7-bayes.pdf`) qui décrit ce pseudo-code.
Le pseudo-code de Paola spécifie deux procédures: la procédure d'apprentissage et la procédure de classification. Bien que le pseudo-code de ces deux procédures soit correcte, son implémentation exige quelques modifications. En effet, l'algorithme bayésien naïf demande de calculer un produit dont les facteurs sont des probabilités comprises entre 0 et 1. Or, le produit de nombres compris entre 0 et 1 pourrait bien être plus petit que le plus petit nombre qu'une machine puisse représenter. Pour résoudre ce problème d'*underflow*, il suffit de substituer à une probabilité son logarithme (la fonction logarithme étant croissante, maximiser un produit de probabilités équivaut à maximiser son logarithme). Rappelez-vous que la classe retournée par le Naive Bayes est la classe c qui maximise le produit

$$P(c) \times \prod_i P(w_i|c)$$

Or le logarithme d'un produit de facteurs est la somme des logarithmes de ces facteurs. Par conséquent, votre programme devra retounrner la classe c qui maximise la somme

$$\log P(c) + \sum_i \log P(w_i|c)$$

Les données que vous utiliserez sont stockées dans l'archive

`http://www.latl.unige.ch/mels/tp7_newgroups.tar.gz`.

Cette archive contient des messages de newsgroups. Sauvegardez cette archive et désarchivez-la au moyen de la commande `tar -xzvf tp7_newgroups.tar.gz`. Il en résultera 7 répertoires: `comp.sys.mac.hardware`, `comp.windows.x`, etc. Chacun de ces répertoires correspond à une classe de newsgroups qui contient des messages. Partionnez ces messages en messages d'entraînement (99 % des messages de chaque répertoire) et messages de test (1 % des messages de chaque répertoire). Les messages n'ont pas été prétraités: ils contiennent donc des mots (ou d'autres signes) qui ne sont pas utiles à la classification. Tâchez d'éliminer ces mots ou signes inutiles.
Rapportez également toutes les mesures de performance qui vous paraissent les plus pertinentes pour évaluer votre classifieur.
Vous devez me faire parvenir votre code et vos résultats au plus tard le 18 janvier 2005 à midi.