M1–Master d'informatique – 2015/2016

PJE – Analyse de comportement avec Twitter

octobre 2015

Classification baysienne

L'objectif du TP est de développer un algorithme de *classification bayésienne* permettant de classifier un *nouveau tweet* comme `positif`, `négatif` ou `neutre`.
Une base d'apprentissage contenant des tweets déjà *étiquetés* et *nettoyés* est considérée comme existante.

Notons que la problématique ressemble beaucoup à la *classification de textes*. Disposant d'un ensemble de textes (tweets) et d'un ensemble de classes (`positif`, `négatif` ou `neutre`), le but est de déterminer la classe la plus probable pour un nouveau tweet.

1 Représentation d'un tweet

Un tweet est une suite de caractères alphanumériques. Il doit être mis sous une certaine forme pour pouvoir être traité par un classificateur bayésien. Pour cela, chaque donnée (un tweet) doit être décrite par un certain nombre d'attributs. Plusieurs représentations sont possibles :

- Un attribut est une position dans le texte : mot 1, mot 2, ... Donc, un attribut prend sa valeur parmi un ensemble de mots possibles. De toute évidence, les tweets ne seront pas strictement de même longueur, le nombre d'attributs va donc être différent pour chaque tweet.
- Un attribut est un mot du "vocabulaire". On lui affecte une valeur booléenne qui indique la présence d'un mot particulier dans le tweet. Ainsi, on ne se préoccupe ni de l'ordre des mots, ni de leur organisation dans le texte, ni de leur nombre d'occurrences.

C'est cette deuxième alternative que nous allons considérer. Notons que celle-ci fait abstraction de l'ordre des mots, de la syntaxe et de la sémantique des textes.

2 Règle de Bayes

L'objectif est de déterminer la classe $c \in C$ la plus probable pour le tweet considéré t.

$$c(t) = \arg\max_{c \in C} P(c|t) \tag{1}$$

L'utilisation de la règle de Bayes nous amène donc à calculer la probabilité d'observer une certaine classe c pour un tweet t.

$$P(c|t) = \frac{P(t|c) \cdot P(c)}{P(t)} \tag{2}$$

Quelque soit la classe considérée, $P(t)$ reste constant et peut donc être "oublié" de la formule. Il reste donc à estimer $P(t|c)$ et $P(c)$.
$P(c)$ peut être estimé par la proportion de tweets de la classe c dans l'ensemble d'apprentissage. Pour estimer $P(t|c)$, nous allons utiliser l'hypothèse naïve de Bayes (indépendance entre les attributs).

$$P(t|c) = \prod_{m \in t} P(m|c) \tag{3}$$

où $P(m|c)$ est la probabilité d'occurrence du mot m dans un texte de la classe c.

On pourrait alors estimer la probabilité $P(m|c)$ à l'aide de l'ensemble d'apprentissage de la façon suivante :

$$P(m|c) = \frac{n(m,c)}{n(c)} \tag{4}$$

où $n(c)$ est le nombre total de mots des tweets de la classe c, et où $n(m,c)$ est la probabilité d'occurrence du mot m dans un texte de la classe c.
Cependant, $n(m,c)$ peut être nul. On utilisera donc un estimateur de Laplace.

$$P(m|c) = \frac{n(m,c)+1}{n(c)+N} \tag{5}$$

où N est le nombre total de mots des tweets de l'ensemble d'apprentissage.

Récapitulons avec un exemple. La probabilité pour un tweet t d'être neutre est estimée comme suit.

$$P(\texttt{neutral}|t) = \prod_{m \in t} P(m|\texttt{neutral}) \cdot P(\texttt{neutral}) \tag{6}$$

Il en va de même pour les classes `positif` et `négatif`.

Question 2.1 : Réalisez une application qui permet, à partir de votre base d'apprentissage, de déterminer la classe la plus probable d'un nouveau tweet.