M1–Master d'informatique – 2015/2016

PJE – Analyse de comportement avec Twitter

octobre 2015

Classification bayésienne (3)

Le but de cette séance est d'analyser expérimentalement la performance des différentes variantes de l'algorithme de *classification bayésienne* permettant de classifier un *nouveau tweet* comme `positif`, `négatif` ou `neutre`.

Analyse expérimentale

Une méthode sophistiquée pour mesurer la qualité d'un classifieur est la "validation croisée". Considérons un ensemble d'apprentissage A. Il faut découper cet ensemble d'instances en $k > 3$ sous-ensembles mutuellement disjoints de même taille. Cette découpe peut se faire aléatoirement en prenant garde à ce que toutes les classes soient représentées avec la même fréquence dans chacun des sous-ensembles. Pour $k = 3$, nous obtenons donc 3 sous-ensembles $A1$, $A2$ et $A3$.

Dans un premier temps, nous produisons un classificateur en prenant $A2 \cup A3$ comme ensemble d'apprentissage. Et nous lui demandons de classifier les instances de $A1$. Ensuite, nous mesurons le taux d'erreur, c'est-à-dire le nombre (ou la proportion) d'instances de $A1$ dont la classe est mal prédite à l'aide de l'ensemble d'apprentissage $A2 \cup A3$. Notons ce taux d'erreur E_{A1}.
Cette procédure est répétée pour $A2$ en prenant $A1 \cup A3$ comme ensemble d'apprentissage, et pour $A3$ en prenant $A1 \cup A2$ comme ensemble d'apprentissage. Ceci nous donne les taux d'erreur E_{A2} et E_{A3}, respectivement.
Le taux d'erreur de l'algorithme de classification est alors estimé par la moyenne des trois :

$$E = \frac{E_{A1} + E_{A2} + E_{A3}}{3} \tag{1}$$

En général on prend $k = 10$. Cette technique est appelée validation croisée en k-plis (*k-fold cross-validation*).

Question 1 : Choisissez un de vos ensembles d'apprentissage et effectuez une validation croisée pour estimer le taux d'erreur des algorithmes de classification bayésienne suivants :

1. Présence, uni-gramme
2. Présence, bi-gramme
3. Présence, uni-gramme + bi-gramme
4. Fréquence, uni-gramme

5. Fréquence, bi-gramme
6. Fréquence, uni-gramme + bi-gramme

Lequel est le plus performant ?

Question 2 : En utilisant la même technique, comparez les performances du classifieur bayésien avec celles du classifieur par mot-clé et du classifieur k-NN.

Question 3 : Si vous avez terminé, vous pouvez considérer un autre ensemble d'apprentissage. Observez-vous des différences ?