

PJE – Analyse de comportement avec Twitter

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 classer 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 A_1 , A_2 et A_3 .

Dans un premier temps, nous produisons un classificateur en prenant $A_2 \cup A_3$ comme ensemble d’apprentissage. Et nous lui demandons de classer les instances de A_1 . Ensuite, nous mesurons le taux d’erreur, c’est-à-dire le nombre (ou la proportion) d’instances de A_1 dont la classe est mal prédite à l’aide de l’ensemble d’apprentissage $A_2 \cup A_3$. Notons ce taux d’erreur E_{A_1} .

Cette procédure est répétée pour A_2 en prenant $A_1 \cup A_3$ comme ensemble d’apprentissage, et pour A_3 en prenant $A_1 \cup A_2$ comme ensemble d’apprentissage. Ceci nous donne les taux d’erreur E_{A_2} et E_{A_3} , respectivement.

Le taux d’erreur de l’algorithme de classification est alors estimé par la moyenne des trois :

$$E = \frac{E_{A_1} + E_{A_2} + E_{A_3}}{3} \quad (1)$$

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 ?