Modèles Probabilistes pour l'Informatique: Classification de documents

Georges Schwing, Maxime Raynal

2 mars 2017

0 Objectifs

L'objectif du projet est de développer un modèle probabiliste pour la classification de documents, dans le cas oú on dispose d'une base d'apprentissage étiquetée contenant un certain nombre de documents par classe que l'on veut identifier. Cela veut dire : base d'apprentissage pour trois classes :

— Classe 1 : sport
— Classe 2 : médecine
— Classe 3 : fait divers

1 Prérequis a l'exécution de notre programme

Dans le but de compiler notre programme vous devez disposer des programmes suivant :

— flex
— flex-devel
— g++
— gcc

Si vous disposez de tout les prérequis vous pouvez compiler l'exécutable a l'aide du Makefile fournis dans le dossier src de l'archive.

make install permet également de télécharger et de décompresser l'archive des statistiques tirées du site d'information Reuter.

2 Questions demandées

Hypothèse : on peut représenter un documents (c'est a dire le caractériser) sous forme de vecteurs de variables aléatoires discrète de Bernoulli.

1. Quel modèle probabiliste peut-on utiliser pour la caractérisation des documents. Les documents sont représentés par des vecteurs qui caractérisent le nombre de termes d'un vocabulaire dans ces documents :
   Exemple :
   $d = vecteur\ de\ mots\ dans\ le\ vocabulaire = (tf_{1,d}, tf_{2,d}, ..., tf_{v,d})$
   $vocabulaire = \{des\ mots\ qui\ apparaissent\ dans\ le\ document\} = \{t_1, t_2, ..., t_v\}$ il s'agit d'un ensemble de termes.
   $classes = \{1, 2, ..., K\} = \mathcal{Y}$ il s'agit d'un ensemble d'étiquettes.
   $tf_{1,d} = \{mots\ d\ indice\ 1\ du\ vocabulaire\ dans\ d\}$

2. On s'intéresse plus à la présence ou non des mots qu'à leur fréquence d'apparition. Ainsi, on transforme notre vecteur en vecteur de booléens $v = (present_{1,d}, absent_{2,d}, ..., present_{v,d})$
3. Modèle utilise : pour chaque classe on la représente par une étiquette $k \in \{1, ..., K\}$. On cherche la probabilité d'un document, représenté par un vecteur de booléens, sachant une classe (très proche du modèle unigramme du projet précédent).
   Hypothèse : les termes du vocabulaire présents dans un document sont interdépendants.
   Ainsi, $P(d = (present_{1,d}, ..., present_{v,d})|jd) = \prod_{j=1}^{v}(P(present_{j,d} = \prod_{j=1}^{v} \theta_{j|k}(1 - \theta_{j|k}).$
   On utilisera un modèle binomial $\rightarrow$ produit de lois de Bernoulli.
4. A qui correspond le $\theta_{j|k}$?, comment l'estimer ?
   On va utiliser la base d'apprentissage pour déterminer les $\theta_{j|k}$.
   $\theta_{j|k} \Rightarrow$ probabilité de présence du terme d'indice j du vocabulaire dans la classe k.
   $\theta_{j|k} = \frac{card(document\ de\ la\ base\ de\ classe\ k\ avec\ le\ mot\ j)+1}{card(document\ de\ la\ classe\ k)+2}$
   A partir de la base on créer notre notre modèle de travail.
5. Au final on utilise la règle de Bayes pour prendre une décision $P(y = k|d) \approx p(\lambda|y = k)(p(y = k)$

3 Exploitation des sources statistiques

Dans le but d'exploiter le gros fichier contenant l'ensemble des statistiques issu des articles de *Reuter* nous avons décidé en d'utiliser *flex* en lieu et place d'une bonne vielle expression régulière avec *grep* ou de *sed*.

L'avantage d'utiliser *flex* est que outre le fait de mettre en application pratique les cours d'analyse syntaxique du semestre 5 est la rapidité d'exécution de *flex* pour parser le fichier.

4 blabla du prof

4.1 blabla1

$S = \{(d_i, y_i), i \in [i, n]\}$

Objectif : Déterminer les paramètres de Bernoulli en maximisant la vraisemblance complète des observations de S.

Donc, pour tout documents d, on associe un vecteur indicateur de classe : $c_i = (0, ..., 0, 1, 0, ..., 0)$

Dans ce cas, la vraisemblance complète des données de l'apprentissage est : $\mathcal{L}_c((q_1, ..., q_k) = p((d_1, y_1), ..., (d_m, y_m))$=, $\mathcal{L}_c$ signifiant *"LikelyHood"* complète.

Trouver les Q qui maximisent cette vraisemblance :

— on les notes $Q^* = \{q_1^*, ..., q_k^*\}$ qui maximisent $\mathcal{L}_c(S, Q)$

— incalculable sans l'hypothèse de l'indépendance (identiquement indépendamment distribue).

Ainsi,

$$\mathcal{L}_c(S, Q) = \prod_{i=1}^{m} p((d_i, y_i)|Q)$$

$$\mathcal{L}_c(S, Q) = \prod_{i=1}^{m}(\prod_{k=1}^{K} p((d_i, y_i = k|Q)^{c_{k,i}})$$

On peut chercher Q^* qui maximise $ln(\mathcal{L}_c(S, Q))$. On a :

$$ln(\mathcal{L}_c(S,Q)) = \sum_{i=1}^{m}(\sum_{k=1}^{k} ln(p((d_i, y_i = k)|Q)))$$
$$ln(\mathcal{L}_c(S,Q)) = \sum_{i=1}^{m}(\sum_{k=1}^{k} c_{k,i} ln(p|d_i, y_i = k))$$

k classe de d $\Leftrightarrow ln(p(k|d)) = \max_{l\in\{1,...,k\}}(\sum_{j=i}^{V} X_{j,d} ln(q_{j,k}) + (1 - X_{j,d})ln(1-q) + ln(p(l)))$

$$\max_{l\in\{1,...,k\}} ln(p(l|d)) = \max_{l\in\{1,...,k\}} ln(p(d|l)p(l)$$
$$\Rightarrow \max_{l\in\{1,...,k\}} ln(p(d|l) + ln(p(l))$$

4.2 blabla2

$$(t|y = k) = \frac{1+nb\ de\ documents\ contenant\ de\ classe\ k}{1+nb\ de\ documents\ de\ classe\ k} = \Theta_{t|k}$$

$$p(y = k) = \frac{nb\ de\ documents\ de\ classe\ k}{nb\ de\ documents\ totaux}$$

$$p(y = l|d) = (p(d|y = l)p(y = l))/p(d)$$

$$p(d|y = l) = \sum_{t\in\mathcal{V}} \Theta_{t|p}^{presence\ t\in d}(1 - \Theta_{t|e})^{1-presence\ t\in d}$$