Méthodes probabilistes pour le TAL
M1 Linguistique Informatique — Marie Candito
Université Paris 7 — 2009/2010

Correction Contrôle des connaissances – 10 mars 2010

1 Cours

1.1 Donnez le théorème de Bayes :**voir cours**

Définissez les concepts suivants :

1.2 Indépendance de deux évènements **: ne pas confondre avec l'exclusion mutuelle !!**

1.3 Variable aléatoire **:** fonction de $S \mapsto R$, i.e. qui a tout élément de S (tout résultat de l'expérience) associe un réel. **voir cours**

1.4 Variable de Bernouilli : variable aléatoire ne prenant que deux valeurs, qu'arbitrairement on note respectivement 1 et 0, et qu'on appelle succès et échec. Le paramètre d'une v.a. de Bernouilli est la probabilité de succès, c'est-à-dire P(X=1). On a une v.a. de Bernouilli dès qu'on utilise un v.a. qui représente une caractéristique binaire sur les résultats d'une expérience.

1.5 Variable binomiale et loi binomiale (i.e. donnez la loi de probabilité d'une variable binomiale). Bonus : redonnez le raisonnement amenant à la formulation mathématique de la loi binomiale)

Une variable aléatoire est binomiale de paramètres **n,p** ssi
- X est définie sur une expérience faite de **n** épreuves indépendantes
- on considère Y une variable de Bernouilli de paramètres **p**, définie sur chaque épreuve
- et la valeur de X pour une réalisation de ces n épreuves vaut le nombre de succès de Y

Donc on peut considérer une v.a. binomiale dès lors qu'il s'agit de compter le nombre de réalisations d'un événement E, lors de la répétition de n épreuves. La v.a. de Bernouilli étant alors définie comme Y=1 si l'événement E est réalisé, et Y=0 sinon.

La loi de probabilité d'une v.a. binomiale est de la forme : $p_X(k) = P(X = k) = C_n^k p^k (1-p)^{n-k}$
On obtient ce résultat ainsi :
- la probabilité d'obtenir un résultat contenant k succès vaut, du fait de l'indépendance des épreuves, $p^k(1-p)^{n-k}$
- mais P(X=k) correspond à la somme des probas de tous les résultats contenant k succès, peu importe comment sont distribués les succès et les échecs sur les n épreuves
- il y a C_n^k façons de distribuer les k succès sur les n positions possibles, d'où le résultat

1.6 Estimation par maximum de vraisemblance : on considère un corpus de phrases journalistiques anglaises contenant 25 phrases averbales sur un total de 10000 phrases. **Expliquez** le principe de l'estimation de paramètres par maximum de vraisemblance et **donnez** l'estimation par maximum de vraisemblance de la probabilité pour une phrase journalistique anglaise d'être averbale.

Correction : On considère un corpus, plus généralement des données, comme le résultat d'une expérience, elle-même constituée de la réalisation de n épreuves.
Par exemple : un corpus de n mots peut être vu comme le résultat de n tirages de mots, ou bien comme le résultat de n tirages de bigrammes de mots etc…
En effet, un modèle probabiliste décompose en général une probabilité complexe en probabilités plus élémentaires, en faisant notamment des hypothèses sur l'indépendance de certains évènements entre eux.

Par exemple, on peut décomposer l'événement « produire une phrase de x mots » en x tirages de mots, et le tirage d'un mot peut ne dépendre que du mot précédent (hypothèse de Markov d'ordre 1) ou des deux mots précédents (hyp. de Markov d'ordre 2) etc…
Ces probabilités plus élémentaires sont appelées les paramètres du modèle.

L'estimation par maximum de vraisemblance consiste à
- considérer la fonction *de vraisemblance* qui pour un corpus fixé (plus généralement des données fixées) prend comme argument des valeurs de paramètres, et leur associe la probabilité du corpus, calculée avec ces valeurs de paramètres
- les paramètres estimés par « maximum de vraisemblance » sont les valeurs qui rendent maximale la fonction de vraisemblance

Intuitivement, il s'agit de fixer les paramètres de telle sorte que ce que l'on connaît (le corpus) soit mathématiquement le plus probable (=ait la plus grande probabilité calculée avec ces paramètres).

Dans le cadre de la question posée,
- les données sont les 10000 phrases
- on les considère comme le résultat de 10000 tirages indépendants de phrases et réponses à la question : la phrase est-elle averbale ou pas. Admettons que l'on note A pour averbale, et V pour verbale, on considère donc une séquence de 10000 A ou V, contenant en tout 25 A
- la probabilité d'une telle séquence, si on note p la probabilité qu'une phrase soit averbale, est $p^{25}(1-p)^{10000-25}$
- ici le **paramètre** du modèle est la probabilité p qu'une phrase soit averbale
- le principe de l'estimation du maximum de vraisemblance consiste à choisir le p qui rende maximal la vraisemblance (notée L pour likelihood)
  - $L_{corpus}(p) = P_p(corpus) = p^{25}(1-p)^{10000-25}$
- Et on peut montrer (par exemple en cherchant le p qui rend nulle la dérivée du log de $L_{corpus}(p)$) que cet argmax est p = 25 / 10000, i.e. la **fréquence relative** de réalisation de l'événement dont p est la probabilité

2 Exercice

On considère un corpus taggé où l'on a mélangé deux corpus :
- un corpus de 10000 mots de phrases taggées par un humain,
- et un corpus de 90000 mots de phrases taggées par un tagger.

Le tagger a une précision (accuracy) de 97%. L'humain fait quant à lui 1% d'erreurs.

On choisit un mot au hasard dans le corpus. Quelle est la probabilité que le mot soit mal taggé ?
Correction : On note :
- H l'événement « le mot appartient à une phrase taggée par un humain »
- $\overline{H}$ et son complémentaire l'événement « le mot appartient à une phrase taggée par un tagger »
- E l'événement « le mot est mal taggé »

$$P(H) = 10000 / (10000 + 90000) = 0{,}1$$

L'énoncé nous donne : $P(\overline{H}) = 0{,}9$
$P(E \mid H) = 0{,}01$
$P(\overline{E} \mid \overline{H}) = 0{,}97$

Et on cherche **P(E)** => Par la formule des probabilités totales, on obtient :
(intuitivement : pour avoir une erreur on doit « passer » par un des deux évènements : le mot a été taggé par un humain ou bien il a été taggé par le tagger)
$P(E) = P(E \mid H)P(H) + P(E \mid \overline{H})P(\overline{H}) = 0{,}01 \times 0{,}1 + 0{,}03 \times 0{,}9 = 0{,}028$

On choisit un mot et il est mal taggé. Quelle est la probabilité qu'il provienne d'une phrase taggée par le tagger ?
Correction : On cherche $P(\overline{H} \mid E)$
que l'on obtient par Bayes : $P(\overline{H} \mid E) = \frac{P(E \mid \overline{H})P(\overline{H})}{P(E)} = \frac{0{,}03 \times 0{,}9}{0{,}028} \approx 0{,}9643$

3 Exercice

Un système de reconnaissance vocale reconnaît correctement un mot dans 60% des cas.
On teste la reconnaissance de messages constitués chacun de 10 mots : on fait lire à Gertrude des messages retranscrits par le reconnaisseur vocal.
On admet que la reconnaissance de chaque mot ne dépend que de ce mot. Et on admet que Gertrude comprendra un message s'il contient au plus deux erreurs.
Formulez mathématiquement la probabilité qu'un message soit compris par Gertrude.
Correction :
On considère l'expérience aléatoire E=« reconnaître 10 mots de suite », qui d'après l'énoncé est la réalisation de 10 épreuves indépendantes M=« reconnaître un mot ».
On considère
la variable de Bernouilli Y définit pour l'expérience M, qui vaut 1 si le mot est bien reconnu et 0 sinon. On a d'après l'énoncé p=P(Y=1) = 0,6
la variable X définit pour l'expérience E, qui compte le nombre de cas où Y vaut 1 : il s'agit d'une variable binomiale, de paramètre (10 ; 0,6)
On cherche la probabilité qu'un message soit compris, i.e.
$P(X > 7) = P(X = 8) + P(X = 9) + P(X = 10)$
sachant que pour une variable de paramètres (n ;p) on a $P(X = k) = C_n^k p^k (1-p)^{n-k}$
On obtient avec le logiciel R :

```
> k <- 8:10
> X <- dbinom(k,10,0.6)
> sum(X)
[1] 0.1672898
```

Soit environ 17% de chances pour que le message soit compris.

4 Classificateur naïf bayésien :

On considère la tâche d'identification de la langue d'un texte. **Expliquez** le principe du classificateur naïf bayésien, dans le cadre de cette tâche (le modèle probabiliste et son estimation sur corpus).
Contraintes : l'unité considérée pour la représentation d'un document sera la lettre. C'est-à-dire que vous représenterez les documents comme des « bag of letters » : des ensembles de lettres avec leur nombre d'occurrences.
Correction :
Les différentes classes sont les langues pour lesquelles on dispose d'un corpus de documents avec langue identifiée. Le principe général consiste à retourner la classe :

$$\hat{c} = \arg\max_c P(c \mid D) = \arg\max_c \frac{P(D \mid c)P(D)}{P(D)} = \arg\max_c P(D \mid c)P(c)$$

Avec un modèle « bag of letters », on décompose la proba de l'événement « D » en la proba conjointe sur toutes les lettres apparaissant dans D, et on considère en outre que chaque lettre est indépendante des autres (! en fait plus précisément, on considère qu'il est inutile de modéliser les dépendances entre lettres pour capturer l'appartenance à une langue), on obtient alors :

$$\hat{c} = \arg\max_c P(c) \prod_{i=1}^{N} P(l_i \mid c)^{C_D(l_i)}$$

$$= \arg\max_c \log(P(c)) + \log(\prod_{i=1}^{N} P(l_i \mid c)^{C_D(l_i)})$$

$$= \arg\max_c \log(P(c)) + \sum_{i=1}^{N} C_D(m_i) \log(P(l_i \mid c))$$

On estime P(l | c) par fréquence relative : le nombre d'occurrences de la lettre l dans les documents de langue c, divisé par le nombre total d'occurrences de lettres dans les documents de langue c.
On estime P(l) par le nombre de documents de langue l divisé par le nombre total de documents.
D'où l'importance du corpus utilisé pour l'estimation : il doit être suffisamment grand pour que les estimations soit fiables, et la répartition des documents sur les différentes langues doit refléter la répartition attendue pour les documents dont la langue sera à identifier.

Pbs de lissage : voir cours

5 Chaîne de Markov

Exprimez la probabilité d'une phrase w1…wn dans le cadre d'une chaîne de Markov d'ordre 1 (=modèle bigramme), sans lissage : donnez d'abord P(w1…wn) en fonction de probabilités plus élémentaires (avec toutes les étapes). Et donnez ensuite l'estimation de ces probabilités élémentaires.

Soit le mini corpus (on ignore la ponctuation et la casse) :
Les lions adorent les antilopes.
Les antilopes adorent les herbes.
Les herbes chatouillent les lions.
Correction :
On ignore les différences de casse, et la ponctuation.
On pose w_0=début de phrase : un token fictif
et on ajoute un token f à chaque fin de phrase : on considère en fait non pas $P(w_1 \ldots w_n)$ mais

$$P(w_1^n f \mid w_0) = P(w_1 \mid w_0)P(w_2 \mid w_0 w_1)P(w_3 \mid w_0 w_1 w_2)...P(w_n \mid w_0^{n-1})P(f \mid w_0^n)$$
$$= (hypMarkov) = \left(\prod_{i=1}^{n} P(w_i \mid w_{i-1})\right) P(f \mid w_n)$$

Ensuite les $P(w_i|w_{i-1})$ sont estimables par fréquence relative, ce qui maximise la vraisemblance du corpus : la probabilité conjointe de toutes les phrases du corpus est maximale quand les $P(w_i|w_{i-1})$ sont ainsi estimés.

$$\forall w_i \in V \cup \{f\}, \forall w_{i-1} \in V \cup \{d\}$$
$$P_{MLE}(w_i \mid w_{i-1}) = \frac{C(w_{i-1}w_i)}{\sum_{w \in V \cup \{f\}} C(w_{i-1}w)} = \frac{C(w_{i-1}w_i)}{C(w_{i-1})}$$

Appliquez le modèle bigramme pour calculer la probabilité de *Les herbes chatouillent les antilopes.*

P(Les herbes chatouillent les antilopes f | d) =
P(les|d) P(herbes|les)P(chatouillent|herbes)P(les|chatouillent)P(antilopes|les)P(f|antilopes)
*=1 * 2/6 * 1/2 * 1 /1 *2/6 *1/2*
=1/36

Bonus : Proposez une méthode pour générer aléatoirement des phrases, étant données les probabilités estimées sur corpus, pour un modèle bigramme.
Correction : Méthode de Shannon pour unigrammes, bigrammes, … ngrammes :
Pour chaque mot x, on attribue à chaque mot y un intervalle d'entiers entre 1 et N, de telle sorte que la **taille de l'intervalle affecté à y soit proportionnelle à la probabilité P(y|x)**. Notons I la fonction qui associe un mot x et un entier n au mot y correspondant. I(x, n) = y
Le mot d est fixé au départ (pour commencer une phrase). Ensuite on tire au hasard un entier n_1 entre 1 et N, et on récupère le mot w_1=I(d,n_1). On réitère en tirant un entier n_2, et on récupère le mot w_2=I(w_1, n_2) etc…