Modèles probabilistes pour l'ingénierie linguistique

Matthieu Constant
Université Paris-Est Marne-la-Vallée, LIGM

Plan

Plan

Probabilités

Probabilité

- Soit X un événement dans une expérience aléatoire
- P(X) est la probabilité que X se produise (valeur réelle entre 0 et 1)
- Comment estimer P(X) ?

Exemple : séquence de symboles

- Alphabet={a,b,c}
- Séquence d'apprentissage de taille L(=10) : ababcaabca
- P(a) est la probabilité que a apparaisse
- $P(a) = \frac{\#occ(a)}{L} = \frac{5}{10} = 0.5$

Probabilités (suite)

Probabilité de plusieurs événements

- Soit X et Y deux événements disjoints dans une expérience aléatoire
- P(X∪Y)= P(X,Y) est la probabilité que X et Y se produisent

Exemple : séquence de symboles

- Séquence d'apprentissage de taille L(=10) : ababcaabca
- P(a,b) est la probabilité que a apparaisse et que b apparaisse à la position suivante (sous-séquence ab)
- $P(a,b) = P(ab) = \frac{\#occ(ab)}{L-1} = \frac{3}{9} = 0.333$

Probabilités (suite)

Probabilité conditionnelle

P(X|Y) est la probabilité que X se produise étant donné que Y se produit

Exemple : séquence de symboles

- Séquence d'apprentissage de taille L(=10) : ababcaabca
- On note P(b|a) la probabilité que *b* apparaisse sachant que *a* le précède

$$P(b|a) = \frac{\#occ(ab)}{\#occ(a)} = \frac{3}{5} = 0.6$$

Probabilités (suite)

Indépendance entre deux événements

Si X et Y sont deux événements indépendants l'un de l'autre,

$$P(X \cup Y) = P(X).P(Y)$$

Généralisation

Si X_1, X_2, ..., X_n sont n événements indépendants les uns des autres,

$$P(X_1 \cup X_2 \cup ... \cup X_n) = P(X_1).P(X_2)...P(X_n)$$

Probabilités (suite)

Indépendance et probabilités conditionnelles

- Soient X, Y et Z trois événements
- Si X et Y sont deux événements indépendants l'un de l'autre,

$$P(X \cup Y|Z) = P(X|Z).P(Y|Z)$$

Généralisation

Si X_1, X_2, ..., X_n sont n événements mutuellement indépendants les uns des autres,

$$P(X_1 \cup X_2 \cup ... \cup X_n|Y) = P(X_1|Y).P(X_2|Y)...P(X_n|Y)$$

Plan

Formule de Bayes

Formule

$$P(Y|X) = \frac{P(Y).P(X|Y)}{P(X)}$$

Maximisation

$$argmax_Y P(Y|X) = argmax_Y P(Y).P(X|Y)$$

Classification supervisée naïve bayésienne

Motivation

On cherche à assigner la catégorie c la plus probable à un document d au moyen d'un modèle probabiliste

Formalisation du problème

- Soit C l'ensemble des catégories possibles
- P(c|d) est la probabilité d'avoir la catégorie c étant donné un document d
- Pour chaque nouveau document d, déterminer la catégorie ĉ définie par

$$argmax_{c \in C} P(c|d) = max_Y P(c).P(d|c)$$

Exemple

Collection d'apprentissage (APP)

Document	contenu	Catégorie
D1	aabd	oui
D2	abcd	non
D3	abbc	oui
D4	bc	oui

On considère que *a*, *b*, *c* et *d* sont des mots.

Classification d'un nouveau document

Trouver la meilleure catégorie (oui ou non) pour un nouveau document D=*abbd*

Estimation des probabilités

Notations

- APP(c) : ensemble des documents catégorisés c de APP
- APP : ensemble documents dans APP
- |E| : nombre d'éléments de l'ensemble E

Estimation de P(c)

- Formule

$$P(c) = \frac{|APP(c)|}{|APP|}$$

- Exemple : P(oui) = 3/4 = 0.75 ; P(non)=0.25

Estimation de P(d|c)

Comment faire ? ? ? ?

Calcul de P(X|Y)

Caractérisation de X

On considère que X est caractérisé par k traits X_1, X_2, ..., X_k.

Hypothèse naïve

On considère que les traits de X sont mutuellement indépendants les uns des autres.

Formule

$$P(X|Y) = P(X_1, X_2, ..., X_n|Y) = P(X_1|Y).P(X_2|Y)...P(X_n|Y)$$

Calcul de P(d|c)

Caractérisation d'un document d

Un document est caractérisé par ses mots.

Calcul des probabilités

$$P(D|oui) = P(a,b,b,d|oui) = P(a|oui).P(b|oui).P(b|oui).P(d|oui)$$

Apprentissage : estimation de $P(X_i|c)$

$P(X_i|c)$ est le nombre d'occurrences du mot X_i dans APP(c), divisé par le nombre total de mots dans APP(c)

Exercice

Collection d'apprentissage

Document	contenu	Catégorie
D1	aabd	oui
D2	abcd	non
D3	abbc	oui
D4	bc	oui

Questions

1. Estimer les probabilités du modèle (i.e. tous les $P(X_i|c)$)
2. Trouver la catégorie la plus probable pour le document D=*abbd*

Plan

Les n-grammes

Définition

Un n-gramme est une sous-séquence de n symboles
(n = 1 $\rightarrow$ unigramme ; 2 $\rightarrow$ bigramme ; 3 $\rightarrow$ trigramme)

Estimation des probabilités de n-grammes

- Utilisation d'un corpus d'apprentissage de taille L
- Formule :

$$P(m_1 m_2 \ldots m_n) = \frac{\#occ(m_1 m_2 \ldots m_n)}{L-n+1}$$

Modèle de n-grammes (Shannon)

Principe

La vraisemblance du prochain symbole dépend d'un historique de symboles de taille limitée à $n-1$ (et non pas de toute la sous-séquence des symboles précédents).

Estimation des probabilités conditionnelles

$$P(m_n|m_1 m_2 \ldots m_{n-1}) = \frac{\#occ(m_1 m_2 \ldots m_n)}{\#occ(m_1 \ldots m_{n-1})}$$

Exemple

Corpus d'apprentissage

- Alphabet de 3 lettres {*a*,*b*,*c*}
- Texte = *aabaacaab* (*L*=9)

Dénombrement

- 1-grammes : *a* (6 occ.),*b* (2),*c*(1)
- 2-grammes : *aa* (3),*ab* (1),*ba* (1),*ac* (1),*ca* (1)
- 3-grammes : *aab* (2),*aba* (1),*baa* (1),*aac* (1),*aca* (1),*caa* (1)

Exemples de probabilités

- P(a) = 2/3 ; P(ab)=1/8 ; P(aab)=2/7
- P(a|a)=3/6=1/2 ; P(b|aa)=2/3

Calcul de la probabilité d'une séquence

Principe

- Soit une séquence $m = m_1 m_2 \ldots m_k$
- Plus k est grand, moins le calcul "classique" de la probabilité de m est fiable (ou possible)
- Solution : principe du modèle des n-grammes

Formule

- n=2 : $P(m) = P(m_1).P(m_2|m_1)....P(m_k|m_{k-1})$
- n=3 :
  $P(m) = P(m_1).P(m_2|m_1).P(m_3|m_1 m_2)...P(m_k|m_{k-2} m_{k-1})$

Exercice : deviner un symbole illisible dans un message

Corpus d'apprentissage

- Alphabet de 3 lettres $\{a,b,c\}$
- Texte = *aabaacaab* (L=9)

Questions

Soit le message $a*ab$ avec * symbolysant une lettre invisible.

1. Deviner la lettre la plus probable pour * avec le modèle bigramme.

Plan

Canal bruité

Principe

Une séquence source s inconnue, est émise et transmise à travers un "canal bruité". En sortie, la séquence observée o est altérée. L'objectif est de décoder la séquence observée, i.e. retrouver la séquence source.

Formalisation

- Trouver la séquence ŝ qui maximise la probabilité P(s|o) parmi toutes les séquences s possibles
- Par la formule de Bayes :

$$\hat{s} = argmax_s P(s|o) = argmax_s P(o|s).P(s)$$

Verrous

Estimation des probabilités

- ex. Modèle de Markov caché

Maximisation efficace de P(s|o)

- Le nombre de séquences s candidates croît exponentiellement en fonction de la longueur
- Utilisation de la programmation dynamique (cf. prochain cours)

Plan

Etiquetage grammatical

But

Associer à une séquence $w = w_1 ... w_k$ de mots, une séquence $t = t_1 ... t_k$ d'étiquettes appartenant à un jeu d'étiquettes J.

Point de vue probabiliste

Trouver la séquence la séquence d'étiquettes $\hat{t}$ qui maximise $P(t|w)$ parmi l'ensemble des séquences d'étiquettes possibles.

$$\hat{t} = argmax_t P(t|w) = argmax_t P(w|t).P(t)$$

Modèle de Markov caché d'ordre N

Principe

- Symbole visible : mot
- Symbole caché à découvrir : étiquette
- Hypothèses d'indépendance

Hypothèses de Markov pour le calcul de $P(w|t).P(t)$

- P(w|t) ? un symbole observé (mot) ne dépend que du symbole caché associé (étiquette)
- P(t) ? un symbole caché (étiquette) ne dépend que de ses N précédents

Modèle de Markov caché d'ordre N

Calcul de $P(w|t)$

$$P(w|t) = P(w_1|t_1).P(w_2|t_2).....P(w_n|t_n)$$

Calcul de $P(t)$ pour N=1

$$P(t) = P(t_1).P(t_2|t_1).P(t_3|t_2).....P(t_n|t_{n-1})$$

Calcul de $P(t)$ pour N=2

$$P(t) = P(t_1).P(t_2|t_1).P(t_3|t_1t_2).....P(t_n|t_{n-2}t_{n-1})$$

Estimation des probabilités

Corpus d'apprentissage

Le corpus utilisé pour l'apprentissage des probabilités de base est un corpus annoté : chaque token est associé à une catégorie grammaticale.

Calcul des probabilités d'émission $P(w_i|t_i)$

$$P(w_i|t_i) = \frac{\#occ(w_i,t_i)}{\#occ(t_i)}$$

Calcul des probabilités de transitions $P(t_i|t_{i-1})$

$$P(t_i|t_{i-1}) = \frac{\#occ(t_{i-1}t_i)}{\#occ(t_{i-1})}$$