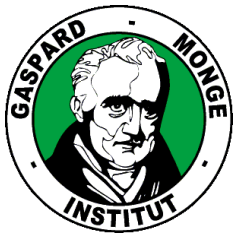
Master 1 Informatique – Université Marne-la-Vallée (IGM)
06/03/2014 – Cours 4
Ingéniérie Linguistique

Modèles probabilistes pour l'ingénierie linguistique



UNIVERSITÉ
PARIS-EST
MARNE-LA-VALLÉE

Philippe Gambette

Sources du cours

- Cours de Matthieu Constant, *Ingéniérie Informatique 1*

http://igm.univ-mlv.fr/ens/Master/M1/2010-2011/IngenierieLinguistique1/cours.php

- Cours de Guillaume Wisniewski, Université Paris-Sud

http://perso.limsi.fr/Individu/wisniews/enseignement/old/10-11/10-11_rdf_m1/

Plan

- Rappels sur les probabilités
- Modèles bayésiens
- Les *n*-grammes
- Le modèle du canal bruité
- Modèle de Markov caché

Plan

- Rappels sur les probabilités
- Modèles bayésiens
- Les *n*-grammes
- Le modèle du canal bruité
- Modèle de Markov caché

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

$$P(a) = \frac{\#occ(a)}{L} =$$

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

Probabilité de plusieurs événements

• Soient *X* et *Y* deux événements disjoints dans une expérience aléatoire

• $P(X \cap Y) = P(X,Y)$ 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)

$$P(a,b) = \frac{\#occ(ab)}{L\text{-}1} =$$

Probabilités

Probabilité de plusieurs événements

• Soient *X* et *Y* deux événements disjoints dans une expérience aléatoire

• $P(X \cap Y) = P(X,Y)$ 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)

$$P(a,b) = \frac{\#occ(ab)}{L\text{-}1} = \frac{3}{9} = 0.333$$

Probabilités

Probabilité conditionnelle

- P($X|Y$) est la probabilité que X se produise étant donné que Y se produit
- $P(X \cap Y) = P(Y).P(X|Y)$

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

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

dans les n-1 premières lettres

Probabilités

Probabilité conditionnelle

- $P(X|Y)$ est la probabilité que X se produise étant donné que Y se produit
- $P(X \cap Y) = P(Y).P(X|Y)$

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

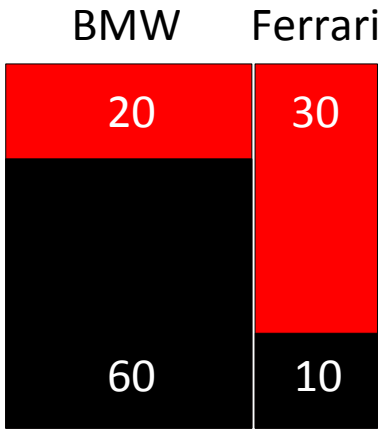
$$P(b|a) = \frac{\#occ(ab)}{\#occ(a)} = \frac{3}{4} = 0.75$$

Probabilités

Probabilité conditionnelle

- P($X|Y$) est la probabilité que X se produise étant donné que Y se produit
- $P(X \cap Y) = P(Y).P(X|Y)$

Exemple :



P(BMW)=

P(Ferrari)=

P(rouge)=

P(noir)=

P(rouge|BMW)=

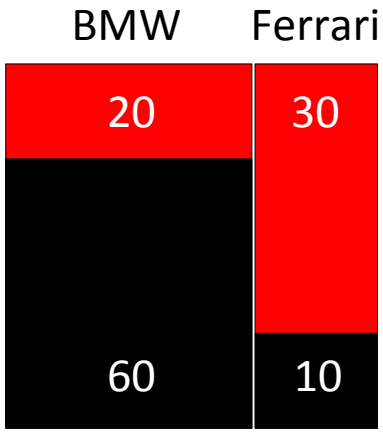
P(BMW rouge)=

Probabilités

Probabilité conditionnelle

- P($X|Y$) est la probabilité que X se produise étant donné que Y se produit
- P($X \cap Y$) = P(Y).P($X|Y$)

Exemple :



P(BMW)=80/120=2/3 P(Ferrari)=40/120=1/3

P(rouge)=50/120=5/12 P(noir)=70/120=7/12

P(rouge|BMW)=20/80=1/4

P(BMW rouge)=20/120=1/6=P(BMW).P(rouge|BMW)

Probabilités

Indépendance entre deux événements

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

$$P(X \cap Y) = P(X) \cdot P(Y)$$

Généralisation

- Si $X_1, X_2, \ldots, X_n$ sont n événements indépendants les uns des autres,

$$P(X_1 \cap X_2 \cap \ldots \cap X_n) = P(X_1) \cdot P(X_2) \ldots P(X_n)$$

Probabilités

Indépendance et probabilités conditionnelles

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

$$P(X \cap Y \mid Z) = P(X|Z) \,.\, P(Y|Z)$$

Généralisation

- Si $X_1, X_2, \ldots, X_n$ sont *n* événements indépendants les uns des autres,

$$P(X_1 \cap X_2 \cap \ldots \cap X_n \mid Y) = P(X_1 \mid Y) \,.\, P(X_2 \mid Y) \ldots P(X_n \mid Y)$$

Plan

- Rappels sur les probabilités
- Modèles bayésiens
- Les *n*-grammes
- Le modèle du canal bruité
- Modèle de Markov caché

Formule de Bayes

Formule

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

Démonstration :

Formule de Bayes

Formule

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

Démonstration :

$P(X \cap Y) = P(Y).P(X|Y)$ et $P(X \cap Y) = P(X).P(Y|X)$, donc $P(Y).P(X|Y)=P(X).P(Y|X)$.

Formule de Bayes

Formule

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

Maximisation

$$\text{argmax}_Y \, P(Y|X) = \text{argmax}_Y \, P(Y) \,.\, P(X|Y)$$

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 $\hat{c}$ définie par

$$\hat{c} = \text{argmax}_{c \in C} P(c|d) = \max_{c \in C} P(c) \,.\, P(d|c)$$

Classification supervisée naïve bayésienne

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) = ; P(non)=

Estimation de P($d|c$)

Comment faire ?

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

$$P(X|Y)=P(X_1, X_2,..., X_k|Y) =P(X_1|Y) \cdot P(X_2|Y) \dots P(X_k|Y)$$

Calcul de P($d|c$)

Caractérisation d'un document *d*

Un document est caractérisé par ses mots.

Calcul des probabilités

$P(D|\text{oui}) = P(a, b, b, d|\text{oui}) = P(a|\text{oui}).P(b|\text{oui}).P(b|\text{oui}).P(d|\text{oui})$

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 $APP(c)$

Rappel de la formule de Bayes : $P(c|d) = \frac{P(c) \,.\, P(d|c)}{P(d)}$

Classification supervisée bayésienne

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. On prend en compte les fréquences de ces mots à l'intérieur des documents pour le calcul des probabilités.

Questions

1. Estimer les probabilités du modèle, c'est-à-dire tous les $P(X_i|c)$

2. Trouver la catégorie la plus probable pour le document D=abbd

Classification supervisée bayésienne

Collection d'apprentissage (APP)

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

Questions

1. Estimer les probabilités du modèle, c'est-à-dire tous les $P(X_i|c)$

   P(a|oui) = 3/10 ; P(b|oui) = 4/10 ; P(c|oui) = 2/10 ; P(d|oui) = 1/10 ;

   P(a|non) = 1/4 ; P(b|non) = 1/4 ; P(c|non) = 1/4 ; P(d|non) = 1/4.

2. Trouver la catégorie la plus probable pour le document D=abbd

   P(oui).P(abbd|oui) = 3/4 * 3/10 * 4/10 * 4/10 * 1/10 = 9/2500

   P(non).P(abbd|non) = 1/4 * 1/4 * 1/4 * 1/4 * 1/4 = 1/1024

   → la catégorie la plus probable est oui.

Plan

- Rappels sur les probabilités
- Modèles bayésiens
- Les *n*-grammes
- Le modèle du canal bruité
- Modèle de Markov caché

Les *n*-grammes

Définition

Un *n*-gramme est une sous-séquence de *n* symboles

(*n* = 1 → unigramme ; 2 → bigramme ; 3 → trigramme)

Estimation des probabilités de *n*-grammes

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

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

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

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

Exemple

Corpus d'apprentissage

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

Dénombrement

- 1-grammes :
- 2-grammes :
- 3-grammes :

Exemples de probabilités

- P(a) = ; P(ab) = ; P(aab) =
- P(a|a) = ; P(b|aa) =

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 (2), 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) = ; P(ab) = ; P(aab) =
- P(a|a) = ; P(b|aa) =

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 (2), ba (1), ac (1), ca (1)
- 3-grammes : aab (2), aba (1), baa (1), aac (1), aca (1), caa (1)

Exemples de probabilités

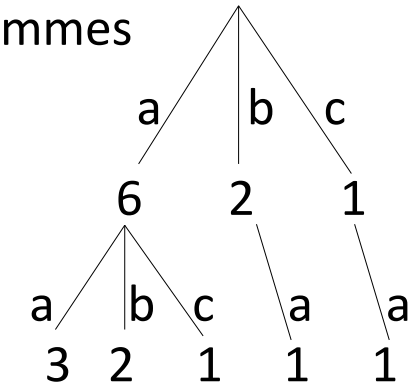
- P(a) = 6/9 = 2/3 ; P(ab) = 2/8 = 1/4 ; P(aab) = 2/7
- P(a|a) = 3/6 = 1/2 ; P(b|aa) = 2/3

Exemple

Corpus d'apprentissage

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

stockage des n-grammes et de leur nombre d'occurrences dans une structure d'arbre :



Dénombrement

- 1-grammes : a (6 occ.), b (2), c(1)
- 2-grammes : aa (3), ab (2), 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) = 6/9 = 2/3 ; P(ab) = 2/8 = 1/4 ; 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 : $P(m) = P(m_1) \cdot P(m_2|m_1) \ldots P(m_k|m_k\text{-}1)$

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

Deviner un symbole illisible

Corpus d'apprentissage

Alphabet de 3 lettres {a,b,c}

Texte = aabaacaab (L=9)

Question

Soit le message m(*)=a*ab avec * symbolisant une lettre invisible.

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

Deviner un symbole illisible

Corpus d'apprentissage

Alphabet de 3 lettres {a,b,c}

Texte = aabaacaab (L=9)

Question

Soit le message m(*)=a*ab avec * symbolisant une lettre invisible.

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

P(m(*)) = P(a) . P(*|a) . P(a|*) . P(b|a)

P(m(a)) =
P(m(b)) =
P(m(c)) =