INF5081

Classification bayésienne naïve

bouguessa.mohamed@uqam.ca

Probabilité

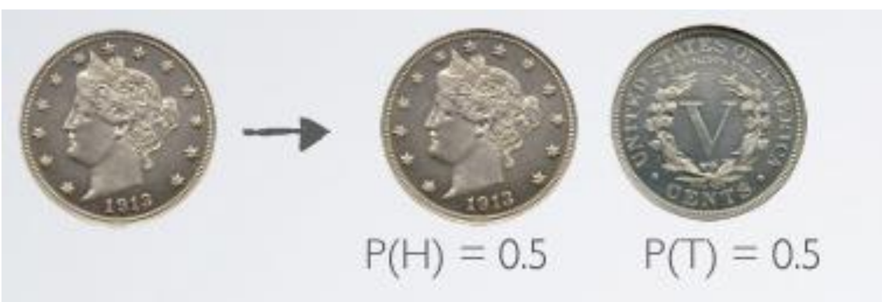
- Probabilités: une manière de quantifier l'incertitude associée aux évènements choisis au sein d'un univers d'évènements
- Par convention $P(E)$ signifie la probabilité de l'évènement E
- Les probabilités sont toujours positives
- La somme des probabilités de tous les évènements est 1

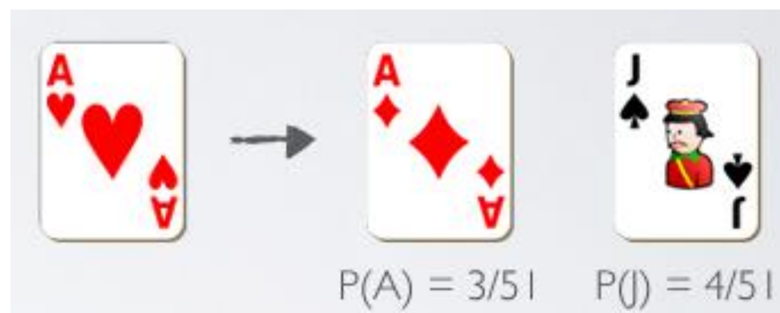
Indépendance

- Deux processus sont indépendants si la connaissance du résultat de l'un ne fournit aucune information sur le résultat de l'autre:

$$P(AB) = P(A)\,P(B)$$

- Les résultats de deux lancers de dés sont indépendants





Source: http://commons.wikimedia.org/wiki/File:1913_Eliasberg_Liberty_Head_Nickel.png

Source: http://openclipart.org/cgi-bin/navigate/recreation/games/cards/white

Probabilité conditionnelle

- Une **probabilité conditionnelle** est la probabilité d’un événement sachant qu'un autre événement a eu lieu
- Si une carte d'un jeu de cartes est tirée au hasard, on estime qu'il y a une chance sur quatre d'obtenir un cœur
- Si on aperçoit un reflet rouge sur la table, il y a maintenant une chance sur deux d'obtenir un cœur

Probabilité conditionnelle

		discipline		
		science informatique	mathématique	totale
sexe	femme	30	20	50
	homme	30	20	50
	totale	60	40	100

$$P(Info \mid F) = \frac{30}{50} = 0.6 \qquad P(Math \mid F) = \frac{20}{50} = 0.4$$

$$P(Info \mid M) = \frac{30}{50} = 0.6 \qquad P(Math \mid M) = \frac{20}{50} = 0.4$$

$$P(Info) = \frac{60}{100} = 0.6 \qquad P(Math) = \frac{40}{100} = 0.4$$

Probabilité conditionnelle

Si $P(A|B) = P(A)$ alors les événements A et B sont dits indépendants.

- Intuitivement: savoir l'occurrence de B ne nous dit rien sur A
- Mathématiquement

$$P(A|B) = \frac{P(A\ et\ B)}{P(B)} = \frac{P(A)\ P(B)}{P(B)} = P(A)$$

Théorème de Bayes

- Permet d'inverser les probabilités conditionnelles
- Supposons que nous ayons besoin de la probabilité d'un certain évènement E conditionne par un autre évènement F, mais que nous n'ayons que des informations sur la probabilité de F conditionne par E

$$P(E|F) = \frac{P(E,F)}{P(F)} = \frac{p(F|E)\ P(E)}{P(F)}$$
$$= \frac{p(F|E)\ P(E)}{P(F|E)\,P(E) + P(F|\neg E)\,P(\neg E)}$$

Exemple 1

- Une maladie affecte 1 personne sur 10000 dans la population et il n'existe qu'un test de dépistage donnant le résultat correct (infecté si la personne est malade et non infecté si la personne est saine) 99% du temps. Le percentage des faux positives est limité à 1%.

Quel est la probabilité d'avoir la maladie (D) sachant que le test est positif (T+) ?

$$P(D|T+) = \frac{p(T+|D)\ P(D)}{P(T+|D)\,P(D) + P(T+|\neg D)\,P(\neg D)} = 0.98$$

Exemple 2

On considère un jeu de 52 cartes

- R : tirer une carte rouge (coeur ou carreau)
- P : tirer un pique
- T : tirer un trèfle
- D : tirer une carte appartenant d'un ensemble

	R	P	T	Effectif
D	4	1	0	5
D'	22	12	13	47
Effectif	26	13	13	52

Exemple 2

On considère un jeu de 52 cartes

- Probabilité de tirer une carte rouge (cœur ou carreau) ?
- Probabilité de tirer une carte appartenant à D sachant qu'elle est rouge $\mathrm{P}(D|R)$?
- Probabilité $\mathrm{P}(D|T)$?
- Probabilité $\mathrm{P}(R|D)$?

	R	P	T	Effectif
D	4	1	0	5
D'	22	12	13	47
Effectif	26	13	13	52

Exemple 3

Dans la salle des profs 60% sont des femmes ; une femme sur trois porte des lunettes et un homme sur deux porte des lunettes : quelle est la probabilité pour qu'un porteur de lunettes pris au hasard soit une femme ?

Théorème de Bayes

- Soit X et Y deux variables aléatoires.
- $P(X, Y)$: probabilité jointe.
- $P(X=x, Y=y)$: probabilité que la variable X prend la valeur x et la variable Y prend la valeur y.
- $P(Y=y|X=x)$: la probabilité que la variable $Y=y$ sachant que $X=x$. (probabilité conditionnelle)
- $P(X, Y) = P(Y|X) * P(X) = P(X|Y) * P(Y)$

Théorème de Bayes

Théorème de Bayes

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

$$P(X) = \sum_{i=1}^{k} P(X, Y_i) = \sum_{i=1}^{k} P(X|Y_i)P(Y_i)$$

$$X \in \{X_1, X_2, \ldots, X_k\}$$

$$Y \in \{Y_1, Y_2, \ldots, Y_k\}$$

Théorème de Bayes pour la classification

- Soit $X=\{X_1, X_2, \ldots, X_d\}$ un ensemble de *d* attributs.
- Soit Y l'attribut des classes.

→ Le but est de prédire la classe Y.

→ Spécifiquement, le but est d'estimer $P(Y|X)$.

cela revient à **estimer les probabilités a posteriori $P(Y|X)$** à partir des données d'apprentissage.

- En connaissant ces probabilités, un nouvel enregistrement X' peut être classé par l'identification de la classe Y' qui maximise la probabilité a posteriori $P(Y'|X')$.

Exemple

Soit l'ensemble d'apprentissage suivant :

classes

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- Prédire la classe de l'enregistrement suivant:

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
11	No	Married	120K	?

$\longrightarrow$

$$P(Y=Yes \mid X_{11}) = ?$$

$$P(Y=No \mid X_{11}) = ?$$

Exemple

- Pour prédire la classe de l'enregistrement précédent, on doit estimer les probabilités *a posteriori* P(**Yes** | X) et P(**No** | X) à partir des données d'apprentissage.

- Application du théorème de Bayes

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

- P(Y=Yes) est la probabilité *a priori* qui peut être estimée directement à partir des données d'apprentissage.

- P(X | Y=Yes) peut-être estimé en utilisant : un classifieur bayésien naïve.

Classification bayésienne naïve

- Soit X = $\{X_1, X_2, \ldots, X_d\}$ un ensemble de *d* attributs.
- Soit y l'étiquette d'une classe.
- La classification bayésienne naïve suppose que les attributs sont indépendants. Formellement :

$$P(X|Y=y)=\prod_{i=1}^{d}P(X_i|Y=y)$$

- Pour classifier un objet, le classifieur estime la probabilité *a posteriori* pour chaque classe Y comme :

$$P(Y|X)=\frac{P(Y)\prod_{i=1}^{d}P(X_i|Y)}{P(X)}$$

Classification bayésienne naïve

- Pour classer un objet, il suffit de choisir la classe qui maximise le terme

$$P(Y)\prod_{i=1}^{d} P(X_i|Y=y)$$

→ Comment estimer les probabilités P(Y) et $P(X_i|Y=y)$ à partir des données d'apprentissage ?

Estimation des probabilités

- **Estimation des probabilités *a priori***

  $P(Y=y) = n_y / n$

  $P(Y=Yes) = 3 / 10$

  $P(Y=No) = 7 / 10$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Estimation des probabilités

❑ Estimation des probabilités conditionnelles

Cas des attributs catégorique

Pour un attribut catégorique Xi, la probabilité conditionnelles P(Xi=xi|Y=y) = le nombre des objets de la classe *y* tel que leurs valeurs dans l'attribut Xi = xi divisé par le nombre total des objets de la classe y

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

P(Status=Married|No) = 4/7

P(Refund=Yes|Yes) = 0

Estimation des probabilités

❑ Estimation des probabilités conditionnelles

Cas des attributs numérique

Deux façons possibles:

1. Discrétisation des attributs numériques continus pour les transformer en attributs catégoriques (décomposition en plusieurs intervalles). Deux situations se présentent:

   1.1. Si le nombre d'intervalles est large, il y'aura peu de données d'apprentissage dans chaque intervalle pour fournir une estimation fiable de $P(Xi|Y)$.

   1.2. Si le nombre d'intervalles est petit, certains intervalles vont contenir des objets de classes différentes ce qui affecte la précision de la classification.

Estimation des probabilités

❑ Cas des attributs numérique

2. On suppose que les valeurs d'un attribut suivent une certaine distribution statistique.

• En général, la distribution **Gaussienne** est souvent utilisée.

• Les paramètres de cette distribution sont : la moyenne μ et la variance σ.

Estimation des probabilités

Cas des attributs numérique

- Pour chaque classe y_j, la probabilité conditionnelle pour l'attribut Xi est définie par

$$P(X_i = x_{ij} \mid Y = y_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{(x_{ij} - \mu_j)^2}{2\sigma_j^2}\right)$$

$$\mu_j = \frac{\sum_{i=1}^{n} x_{ij}}{n}$$

$$\sigma_j^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(x_{ij} - \mu_j\right)^2$$

Estimation des probabilités

❑ Cas des attributs numérique

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

$$P(Income = 120 \mid No) = \frac{1}{\sqrt{2\pi \times 2975}} \exp\left(-\frac{(120-110)^2}{2(2975)} \right) = 0.0072$$

Un exemple de la classification bayésienne naïve

Ensemble d'apprentissage

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Un objet X à classer

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
11	No	Married	120K	?

- On doit estimer P(Yes | X) et P(No | X)

$$P(Yes|X) = \frac{P(Yes)\prod_{i=1}^{d} P(X_i|Yes)}{P(X)}$$

- Estimation des probabilités *a priori*

P(Yes) = 0.3 et P(No) = 0.7

Un exemple de la classification bayésienne naïve

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
11	No	Married	120K	?

- Estimation des probabilités cond. : P(Xi | Yes) et P(Xi | No)

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- **pour les attributs catégoriques**

P(Owner=Yes|No) = 3/7
P(Owner=No|No) = 4/7
P(Owner=Yes|Yes) = 0
P(Owner=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No)=1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/3
P(Marital Status=Divorced|Yes)=1/3
P(Marital Status=Married|Yes) = 0

- **pour l'attribut numérique « Annual Income »**

$$P(Income = 120K \mid No) = \frac{1}{\sqrt{2\pi \times 2975}} \exp\left(-\frac{(120-110)^2}{2(2975)} \right) = 0.0072$$

$$P(Income = 120K \mid Yes) = \frac{1}{\sqrt{2\pi \times 25}} \exp\left(-\frac{(120-90)^2}{2(25)} \right) = 1.2 \times 10^{-9}$$

Un exemple de la classification bayésienne naïve

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
11	No	Married	120K	

- $P(X|No) = P(Owner{=}No|No)$
  $\times P(\text{Marital Status} = \text{Married}|No)$
  $\times P(\text{Income}{=}120K|No)$
  $= 4/7 \times 4/7 \times 0.0072 = 0.0024$
- $P(X|Yes) = P(Owner{=}No|Yes)$
  $\times P(\text{Married}|\ Yes)$
  $\times P(\text{Income}{=}120K|Yes)$
  $= 1 \times 0 \times 1.2 \times 10^{-9} = 0$

- Puisque $P(X|No)P(No) > P(X|Yes)P(Yes)$

doc $P(No|X) > P(Yes|X)$

→ Conclusion : l'enregistrement sera placé dans la classe No

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
11	No	Married	120K	**No**

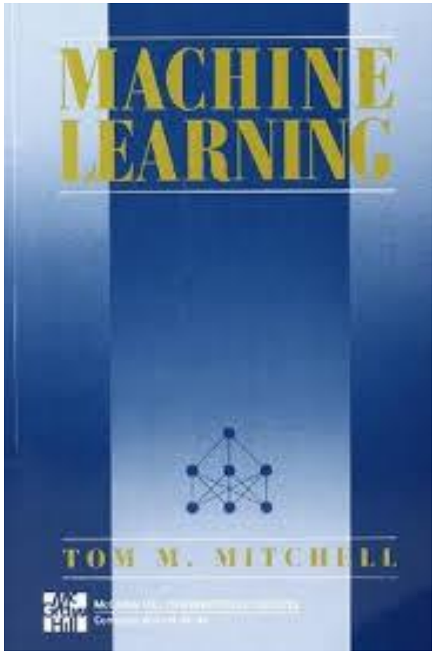
Exemple : Classification du texte

• Classer les documents comme en deux classes :
Document → {+, -}

• Phase d'apprentissage : estimation de

- P(+)
- P(-)
- P(Doc|+)
- P(Doc|-)

Algorithme tiré du livre de Tom Mitchell



Learn_Naive_Bayes_Text(TrainData, Y)

//Collecter tous les mots dans TrainData

- *vocabulaire* ← l'ensemble de mots TrainData //ne pas considérer les mots vides. Considérer qu'une seule occurrence par mot.

//Calcule de la probabilité à priori de chaque classe $P(Y=y_j)$ ainsi que les probabilités conditionnels associés au mots $P(w_i|yj)$

- **pour** chaque classe y_j dans Y **faire**
  - *$docs_j$* ← l'ensemble de mots appartenant à la classe y_j;
  - Calcule de la probabilité à priori de la classe y_j $P(y_j) \leftarrow \frac{|docs_j|}{|TrainData|}$
  - *W* ← l'ensemble de mots qui sont à la fois dans *$docs_j$* ET *vocabulaire*
  - **pour** chaque mot w_i dans *W* **faire**
    - Calcule des probabilités $P(wi|y_j) \leftarrow \frac{n_i + 1}{n + |vocabulaire|}$
      - n_i ← la fréquence du mot w_i dans $docs_j$
      - n ← le nombre de mots distinct dans $docs_j$

- Return *Décision* tel que $Décision = \max_{y_j \in Y}\left[P(y_j) \prod_{w_i} P(w_i|y_j)\right]$