Fouille de Données

Data Mining

Classification - Partie 3

Plan du cours

1. Classification Naive Bayésienne

Classification Bayésienne

- Basée sur des lois statistiques. Approche probabiliste.
- ➔ Probabilité conditionnelle & théorème de Bayes.
- Utilise la notion de « plus probable » sachant ….
- Connaissances *a priori* - Prévision du futur à partir du passé.
- Différente de l'approche basée sur les fréquences.
  - Fréquences : on estime la probabilité d'occurrence d'un événement.
  - Bayésienne : on estime la probabilité d'occurrence d'un événement **sachant** qu'une hypothèse préliminaire est vérifiée (connaissance).

Classification Bayésienne

Probabilité Conditionnelle

⇒ Quelle est la probabilité que quelque chose se produise, sachant que quelque chose d'autre s'est déjà passé.

On note : P(A|B) – Probabilité de A, sachant B.

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

Exemple :

Classification Bayésienne

Théorème de Bayes

$\Rightarrow$ à partir de la probabilité conditionnelle, on peut déduire que :

$$P(A \cap B) = P(A|B) * P(B) = P(B|A) * P(A)$$

Donc :

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

Classification Bayésienne

Théorème de Bayes

Cas Classification : Probabilité qu'un exemple X appartienne à une classe C.

fonction de vraisemblance de C;
calculée depuis le training set

Probabilité a priori;
précède toute info sur X

$$P(C|X) = \frac{P(X|C) * P(C)}{P(X)}$$

Probabilité Postérieure;
dépend directement de X

Classification Bayésienne

Théorème de Bayes

$$P(C|X) = \frac{P(X|C) * P(C)}{P(X)}$$

Cas Classification : Probabilité qu'un exemple X appartienne à une classe C.

Exemple :

X = (35ans, 40 000, **?**)

Age	Income	Buys Computer
….	…	…

P(**C**|**X**) - La probabilité que le client **X** achète (**C**) un ordinateur sachant que nous connaissons l'âge et le revenu du client.

P(**C**) - La probabilité qu'un client donné achète un ordinateur, quel que soit son âge, son revenu, ou toute autre information.

Classification Bayésienne

Théorème de Bayes

$$P(C|X) = \frac{P(X|C) * P(C)}{P(X)}$$

Cas Classification : Probabilité qu'un exemple X appartienne à une classe C.

Exemple :

X = (35ans, 40 000, **?**)

Age	Income	Buys Computer
….	…	…

P(**X**|**C**) - La probabilité qu'un client, **X**, ait 35 ans et gagne 40 000, sachant que nous savons que le client achètera un ordinateur.

P(**X**) - La probabilité qu'un client dans le Training Set ait 35 ans et gagne 40 000.

Classification Bayésienne

Classification Bayésienne Naïve

- *D* : base d'entrainement de *N* exemples avec leurs classes associées.
- Chaque exemple est décrit par *n* attributs : $A_1, A_2, A_3, \ldots, A_n$
- Chaque exemple X : $X = (x_1, x_2, x_3, \ldots, x_n)$
- *m* classes sont possibles : $C_1, C_2, C_3, \ldots, C_m$

$$P(C_i | x_1, .., x_n) = \frac{P(x_1, .., x_n | C_i) * P(C_i)}{P(x_1, .., x_n)}$$

Classification Bayésienne

Classification Bayésienne Naïve

$$P(C_i|X) = \frac{P(X|C_i) * P(C_i)}{P(X)}$$

- Approche naïve => **Indépendance des attributs.**

- D'où :

$$P(\boldsymbol{X}|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$

$$= P(x_1|C_i) \times P(x_2|C_i) \times \cdots \times P(x_n|C_i).$$

Classification Naïve Bayésienne

Exemple

- Training Set : 1000 exemples décrivant des fruits.
- Attributs :
  - Longueur (Long ou non),
  - Sucre (Sucré ou non),
  - Couleur (Jaune ou non).
- Classes possibles :
  - Banane,
  - Orange,
  - ou Autres.

Long	Sucré	Jaune	Classe
….	…	…	…

Classification Naïve Bayésienne

New Data X : Long, Sucré, Jaune, ?

Exemple

CPT : Conditional Probability Table.

Type	Long	Non Long	Sucré	Non sucré	Jaune	Non Jaune	Total
Banane	400	100	350	150	450	50	**500**
Orange	0	300	150	150	300	0	**300**
Autre	100	100	150	50	50	150	**200**
Total	**500**	**500**	**650**	**350**	**800**	**200**	**1000**

Classification Naïve Bayésienne

Exemple

New Data X : Long, Sucré, Jaune, **?**

Calculer :

- Probabilité postérieure - classe Banane : P(Banane|Long, Sucré, Jaune)
- Probabilité postérieure - classe Orange: P(Orange|Long, Sucré, Jaune)
- Probabilité postérieure - classe Autre: P(Autres|Long, Sucré, Jaune)

$$P(C_i|X) = \frac{P(X|C_i) * P(C_i)}{P(X)}$$

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$

Classification Naïve Bayésienne

New Data X : Long, Sucré, Jaune, ?

Exemple

$$P(C_i|X) = \frac{P(X|C_i) * P(C_i)}{P(X)}$$

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$

Probabilité postérieure - classe Banane :

$$P(Banane|Long, Sucré, Jaune) =$$

$$\frac{P(Long|Banane) * P(Sucré|Banane) * P(Jaune|Banane) * P(Banane)}{P(Long) * P(Sucré) * P(Jaune)}$$

Classification Naïve Bayésienne

New Data X : Long, Sucré, Jaune, ?

Exemple

$$P(C_i|X) = \frac{P(X|C_i) * P(C_i)}{P(X)}$$

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$

Probabilité postérieure - classe Orange:

$$P(Orange|Long, Sucré, Jaune) =$$

$$\frac{P(Long|Orange) * P(Sucré|Orange) * P(Jaune|Orange) * P(Orange)}{P(Long) * P(Sucré) * P(Jaune)}$$

Classification Naïve Bayésienne

New Data X : Long, Sucré, Jaune, ?

Exemple

$$P(C_i|X) = \frac{P(X|C_i) * P(C_i)}{P(X)}$$

$$P(\mathbf{X}|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$

Probabilité postérieure - classe Autre:

$$P(Autre|Long, Sucré, Jaune) =$$

$$\frac{P(Long|Autre) * P(Sucré|Autre) * P(Jaune|Autre) * P(Autre)}{P(Long) * P(Sucré) * P(Jaune)}$$

Classification Naïve Bayésienne

Exemple

$$P(C_i \mid X) = \frac{P(X \mid C_i) * P(C_i)}{P(X)}$$

Probabilité à priori : P(C)

- P(Banane) = 500/1000 = 0.5
- P(Orange) = 300/1000 = 0.3
- P(Autre)= 200/1000 = 0.2

Type	Long	Non Long	Sucré	Non sucré	Jaune	Non Jaune	Total
Banane	400	100	350	150	450	50	**500**
Orange	0	300	150	150	300	0	**300**
Autre	100	100	150	50	50	150	**200**
Total	**500**	**500**	**650**	**350**	**800**	**200**	**1000**

Classification Naïve Bayésienne

New Data X : Long, Sucré, Jaune, ?

Exemple

$$P(C_i \mid X) = \frac{P(X \mid C_i) * P(C_i)}{P(X)}$$

Probabilité : P(X)

- P(Long) = 500/1000 = 0.5
- P(Sucré) = 650/1000 = 0.65
- P(Jaune)= 800/1000 = 0.8

Type	Long	Non Long	Sucré	Non sucré	Jaune	Non Jaune	Total
Banane	400	100	350	150	450	50	**500**
Orange	0	300	150	150	300	0	**300**
Autre	100	100	150	50	50	150	**200**
Total	**500**	**500**	**650**	**350**	**800**	**200**	**1000**

Classification Naïve Bayésienne

New Data X : Long, Sucré, Jaune, ?

Exemple

Probabilité Vraisemblance: P(X|**Banane**)

$$P(C_i|X) = \frac{P(X|C_i) * P(C_i)}{P(X)}$$

- P(Long|**Banane**) = 400/500
- P(Sucré|**Banane**) = 350/500
- P(Jaune|**Banane**)= 450/500

Type	Long	Non Long	Sucré	Non sucré	Jaune	Non Jaune	Total
Banane	400	100	350	150	450	50	**500**
Orange	0	300	150	150	300	0	**300**
Autre	100	100	150	50	50	150	**200**
Total	**500**	**500**	**650**	**350**	**800**	**200**	**1000**

Classification Naïve Bayésienne

New Data X : Long, Sucré, Jaune, ?

Exemple

Probabilité Vraisemblance: P(X|**Orange**)

- P(Long|**Orange**) = 0/300
- P(Sucré|**Orange**) = 150/300
- P(Jaune|**Orange**)= 300/300

Type	Long	Non Long	Sucré	Non sucré	Jaune	Non Jaune	Total
Banane	400	100	350	150	450	50	**500**
Orange	0	300	150	150	300	0	**300**
Autre	100	100	150	50	50	150	**200**
Total	**500**	**500**	**650**	**350**	**800**	**200**	**1000**

Classification Naïve Bayésienne

New Data X : Long, Sucré, Jaune, ?

Exemple

Probabilité Vraisemblance: P(X|**Autre**)

- P(Long|**Autre**) = 100/200
- P(Sucré|**autre**) = 150/200
- P(Jaune|**Autre**)= 50/200

Type	Long	Non Long	Sucré	Non sucré	Jaune	Non Jaune	Total
Banane	400	100	350	150	450	50	**500**
Orange	0	300	150	150	300	0	**300**
Autre	100	100	150	50	50	150	**200**
Total	**500**	**500**	**650**	**350**	**800**	**200**	**1000**

Classification Naïve Bayésienne

Exemple

New Data X : Long, Sucré, Jaune, ?

- Probabilité postérieure - classe Banane : P(Banane|X) = 0.252 /P(X)
- Probabilité postérieure - classe Orange: P(Orange|X) = 0/P(X)
- Probabilité postérieure - classe Autre: P(Autres|X) = 0.01875 /P(X)

➔ **New Data X** : Long, Sucré, Jaune, **Banane**

Classification Naïve Bayésienne

Exemple 2:

New Data X : (a2, b1, c3, d1, **?**)

	A	B	C	D	Classe
E1	a1	b1	c1	d2	+
E2	a1	b2	c2	d2	+
E3	a1	b2	c3	d1	-
E4	a2	b1	c1	d1	-
E5	a2	b2	c1	d1	-
E6	a2	b2	c1	d2	+
E7	a1	b1	c1	d1	+
E8	a2	b1	c2	d2	-
E9	a3	b1	c3	d1	+
E10	a3	b2	c2	d2	+

- Probabilité postérieure - classe **+** : P(+|(a2, b1, c3, d1)) = ?
- Probabilité postérieure - classe **-** : P(-|(a2, b1, c3, d1)) = ?

Classification Naïve Bayésienne

Exemple 2:

New Data X : (a2, b1, c3, d1, **?**)

- Probabilité postérieure - classe **+** : P(+|X) =

(P(a2|+) * P(b1|+) * P(c3|+) * P(d1|+) * P(+)) / P(a2)*P(b1)*P(c3)*P(d1) =

(1/6 * 3/6 * 1/6 * 2/6 * 6/10) / (4/10 * 5/10 * 2/10 * 5/10) = 0,0027 / P(X)

- Probabilité postérieure - classe **-** : P(-|X) =

(P(a2|-) * P(b1|-) * P(c3|-) * P(d1|-) * P(-)) / P(a2)*P(b1)*P(c3)*P(d1) =

(3/4 * 2/4 * 1/4 * 3/4 * 4/10) / (4/10 * 5/10 * 2/10 * 5/10) = 0,02 / P(X)

➔ **New Data X** : (a2, b1, c3, d1, **-**)

Classification Naïve Bayésienne

Cas attributs numériques continus

- Discrétisation ou distribution des valeurs.
- **Distribution normale** des attributs : Calcul de la moyenne et de l'écart type.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Moyenne

$$\sigma = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2\right]^{0.5}$$

Écart type

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Distribution Normale

Classification Naïve Bayésienne

Cas attributs numériques continus

Temperature	Humidity	Class
Hot	86	Yes
Hot	96	Yes
Cool	80	Yes
Cool	65	Yes
Hot	70	Yes
Cool	80	Yes
Hot	70	Yes
Hot	90	Yes
Cool	75	Yes
Cool	85	No
Hot	90	No
Cool	70	No
Hot	95	No
Cool	91	No

P(Yes | Hot, 74) = [(Hot | Yes) * **P(74 | Yes)** * P(Yes)] / P(X)

Classification Naïve Bayésienne

Cas attributs numériques continus

- Distribution normale des attributs : Calcul de la moyenne et de l'écart type.

Example:

		Humidity	*Mean*	*StDev*
Play Golf	yes	86 96 80 65 70 80 70 90 75	79.1	10.2
	no	85 90 70 95 91	86.2	9.7

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\sigma = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2\right]^{0.5}$$

Classification Naïve Bayésienne

Cas attributs numériques continus

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

P(74 | Yes)

		Humidity	Mean	StDev
Play Golf	yes	86 96 80 65 70 80 70 90 75	79.1	10.2
	no	85 90 70 95 91	86.2	9.7

$$P(\text{humidity} = 74 \mid \text{play} = \text{yes}) = \frac{1}{\sqrt{2\pi}(10.2)} e^{-\frac{(74-79.1)^2}{2(10.2)^2}} = 0.0344$$

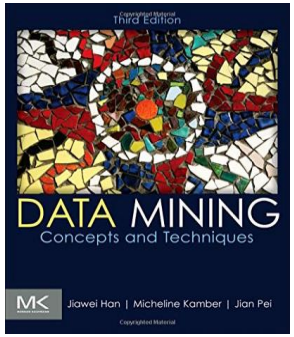
$$P(\text{humidity} = 74 \mid \text{play} = \text{no}) = \frac{1}{\sqrt{2\pi}(9.7)} e^{-\frac{(74-86.2)^2}{2(9.7)^2}} = 0.0187$$

Classification Naïve Bayésienne

Quelques domaines d'application

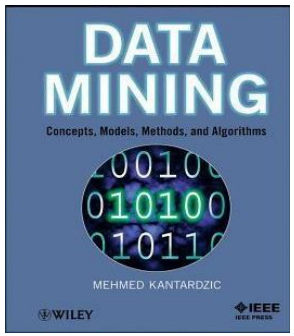
- Prédiction temps réel. – rapidité.
- **Classification textuelle**.
- Sentiment Analysis - Opinion Mining.
- Filtrage bayésien du spam et des courriers indésirables.
- Systèmes de recommandation.

Ressources



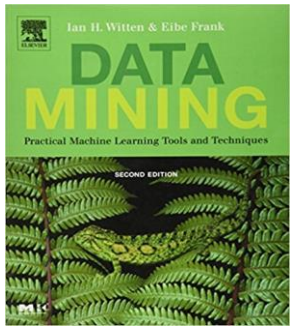
Data Mining : concepts and techniques, 3rd Edition

- ✓ Auteur : Jiawei Han, Micheline Kamber, Jian Pei
- ✓ Éditeur : Morgan Kaufmann Publishers
- ✓ Edition : Juin 2011 - 744 pages - ISBN 9780123814807



Data Mining : concepts, models, methods, and algorithms

- ✓ Auteur : Mehmed Kantardzi
- ✓ Éditeur : John Wiley & Sons
- ✓ Edition : Aout 2011 – 552 pages - ISBN : 9781118029121



Data Mining: Practical Machine Learning Tools and Techniques

- ✓ Auteur : Ian H. Witten & Eibe Frank
- ✓ Éditeur : Morgan Kaufmann Publishers
- ✓ Edition : Juin 2005 - 664 pages - ISBN : 0-12-088407-0

Ressources

Cours – Abdelhamid DJEFFAL – Fouille de données avancée

- ✓ www.abdelhamid-djeffal.net

WekaMOOC – Ian Witten – Data Mining with Weka

- ✓ https://www.youtube.com/user/WekaMOOC/featured

Cours - PJE : Analyse de comportements avec Twitter Classification supervisée - Arnaud Liefooghe

- ✓ http://www.fil.univ-lille1.fr/~liefooghe/PJE/bayes-cours.pdf

✓ http://stackoverflow.com/questions/10059594/a-simple-explanation-of-naive-bayes-classification