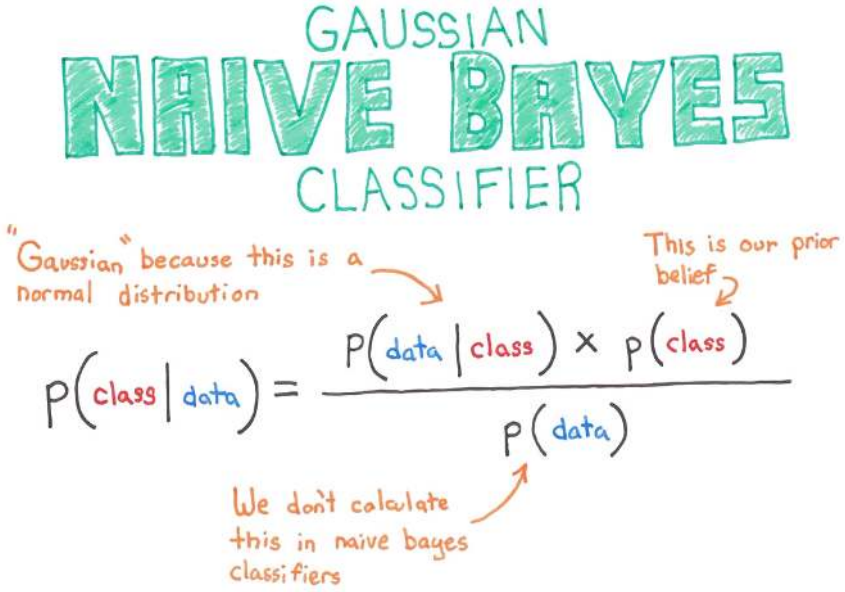
Les classifieurs bayésiens

AI for bio
J-P Comet
Introduction
Trees
Clustering
kNN
Bayes
intro
règle de Bayes
Estimations
Prédictions
Exemple
Valeur manquante
SVM
NN
Evaluation
Data Mining Project



Classifieur bayesien

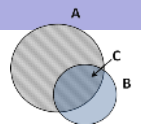
AI for bio
J-P Comet
Introduction
Trees
Clustering
kNN
Bayes
intro
règle de Bayes
Estimations
Prédictions
Exemple
Valeur manquante
SVM
NN
Evaluation
Data Mining Project

- méthode de classification reposant sur une approche probabiliste
- basée sur la règle de Bayes
- intérêt de cette approche : possibilité d'intégrer des connaissances *a priori*.
- Différence avec l'approche classique, dite fréquentiste :
  - classique : on estime $Pr[x]$,
  - bayésienne : on estime $Pr[x|w]$
    w joue le rôle d'une hypothèse
    w peut être un événement du genre *je possède telle connaissance*.

Dans le cas fréquentiste, on fait comme si l'occurence de l'événement x était un fait absolu et sa valeur absolue.
Dans le cas bayésien, on a une attitude pragmatique qui peut s'interpréter comme : *dans le monde où je vis, sachant que je suis dans telle situation notée w, j'estime la probabilite d'occurence de l'événement x*.

La règle de Bayes

AI for bio
J-P Comet
Introduction
Trees
Clustering
kNN
Bayes
intro
règle de Bayes
Estimations
Prédictions
Exemple
Valeur manquante
SVM
NN
Evaluation
Data Mining Project



- Probabilité conditionnelle :
  $Pr[A \cap B] = Pr[A|B].Pr[B] = Pr[B|A].P[A]$
- Théorème de Bayes : Soient A, B et C trois événements. On a :

$$Pr[A|B,C] = \frac{Pr[B|A,C]Pr[A|C]}{Pr[B|C]}$$

où :
- $Pr[B|A,C]$ est la vraisemblance de l'événement B si A et C sont vérifiés ;
- $Pr[A|C]$ est la probabilité *a priori* de l'événement A sachant C ;
- $Pr[B|C]$ est la probabilité marginale de l'événement B sachant C ;
- $Pr[A|B,C]$ est la probabilité *a posteriori* de A si B et C.

- Application à la classification :
  $Pr[y|x,\mathcal{X}]$: probabilité d'observer la classe y si on observe la donnée x sachant que l'on dispose de l'ensemble d'exemples $\mathcal{X}$.
  C'est l'estimation de la probabilité que la donnée x soit de classe y étant donné que je dispose des exemples $\mathcal{X}$.
  En appliquant la règle de Bayes, on obtient :

$$Pr[y|x,\mathcal{X}] = \frac{Pr[x|y,\mathcal{X}].Pr[y|\mathcal{X}]}{Pr[x|\mathcal{X}]}$$

$Pr[y|\mathcal{X}]$ - estimation de la probabilité *a priori*

AI for bio
J-P Comet
Introduction
Trees
Clustering
kNN
Bayes
intro
règle de Bayes
Estimations
Prédictions
Exemple
Valeur manquante
SVM
NN
Evaluation
Data Mining Project

- $Pr[y|\mathcal{X}]$: probabilité *a priori*
  proba d'observer la classe y étant donné l'ensemble d'exemples $\mathcal{X}$.
- estimée par la proportion d'exemples de $\mathcal{X}$ qui sont de classe y.
- si pour une raison ou une autre, on dispose de cette information, on peut utiliser la proportion dans l'ensemble des données de la classe y

$Pr[x|y, \mathcal{X}]$ - estimation de la vraisemblance

- $Pr[x|y, \mathcal{X}]$: la vraisemblance de l'événement "*observer la donnée x*" si elle est de classe y, disposant des exemples $\mathcal{X}$.
  Terme plus difficile à estimer
  Hypothèse de Bayes naïve (HBN) :
  - la donnée x est une conjonction de valeur d'attributs
  - on suppose que les attributs sont des variables aléatoires indépendantes (pas corrélées).
  - Clairement, hypothèse jamais vérifiée ;
  - Cependant, elle permet de faire des calculs simplement et les résultats obtenus ne sont pas sans intérêt.
  - si l'on a des informations concernant ces corrélations entre valeurs d'attributs, on pourra les utiliser.
- Si on applique l'HBN, en supposant que la donnée x est décrite par P attributs notés a_j dont les valeurs sont notées v_j , on écrit :

$$\begin{aligned} Pr[x|y, \mathcal{X}] &\approx Pr[a_1 = v_1|y, \mathcal{X}] \times ... \times Pr[a_P = v_P|y, \mathcal{X}] \\ &= \prod_{i=1}^{i=P} Pr[a_i = v_i|y, \mathcal{X}] \end{aligned}$$

- Chaque terme $Pr[a_j = v_j|y, \mathcal{X}]$ est estimé à l'aide de l'ensemble d'exemples
- l'estimation de ce terme dépend de la nature, qualitative ou quantitative, de l'attribut. Les exemples donnés plus bas illustrent ce point ;

$Pr[x, \mathcal{X}]$ - estimation de la proba marginale

- $Pr[x \mid X]$ est la probabilité d'observer la donnée x, ayant l'ensemble d'exemples X : *a priori*, on ne voit pas comment calculer cette quantité. De fait il s'agit d'un facteur d'échelle...
- Astuce très simple : si la classe est binaire (on généralise sans difficulté aux autres cas), la somme de la probabilité d'observer une donnée x si elle est de la première classe et de la probabilité d'observer cette même donnée x si elle est de la seconde vaut 1. On peut donc écrire :

$$\sum_{y_i \in Y} Pr[y_i \mid x, \mathcal{X}] = 1$$

- Du coup :

$$\begin{aligned} Pr[y \mid x, \mathcal{X}] &= \frac{Pr[y \mid x, \mathcal{X}]}{1} = \frac{Pr[y \mid x, \mathcal{X}]}{\sum_{y_i \in Y} Pr[y_i \mid x, \mathcal{X}]} \\ &= \frac{\frac{Pr[x \mid y, \mathcal{X}] Pr[y \mid \mathcal{X}]}{Pr[x, \mathcal{X}]}}{\sum_{y_i \in Y} \frac{Pr[x \mid y_i, \mathcal{X}] Pr[y_i \mid \mathcal{X}]}{Pr[x, \mathcal{X}]}} \end{aligned}$$

soit finalement :

$$Pr[y \mid x, \mathcal{X}] = \frac{Pr[x \mid y, \mathcal{X}] Pr[y \mid \mathcal{X}]}{\sum_{y_i \in Y} Pr[x \mid y_i, \mathcal{X}] Pr[y_i \mid \mathcal{X}]} \quad (2)$$

Et maintenant, on peut tout calculer...

Classifieur bayésien : prédiction

Classe MAP. Une fois que avoir calculé $Pr[y \mid x, \mathcal{X}], \forall y \in Y$, on peut prédire sa classe comme étant celle qui **maximise** la probabilité *a posteriori* : c'est la classe **MAP (Maximal A Posteriori)**, soit :

$$y_{MAP} = \underset{y \in Y}{argmax}\ Pr[y|x, \mathcal{X}]$$

qui peut s'écrire aussi en appliquant l'éq. 2 :

$$y_{MAP} = \underset{y \in Y}{argmax}\ Pr[x \mid y, \mathcal{X}] Pr[y \mid \mathcal{X}] \quad (3)$$

Classe ML.. Si on ne tient pas compte de $Pr[y|\mathcal{X}]$ et qu'on ne tient compte que de la vraisemblance $Pr[x|y]$, on obtient la classe **ML (Maximal Likelihood)**, soit :

$$y_{ML} = \underset{y \in Y}{argmax}\ Pr[x \mid y, \mathcal{X}]$$

Clairement, si les exemples sont uniformément répartis entre toutes les classes, soit $Pr[y|\mathcal{X}] = \frac{1}{|Y|}$, la classe ML et la classe MAP sont les mêmes.

Exercice 1 : TD sur les classifieurs bayesiens "jouer au tennis".

Classification homme/femme : apprentissage

- On cherche à classifier chaque personne en tant qu'homme ou femme, selon les caractéristiques mesurées. Les caractéristiques comprennent la taille, le poids, et la pointure.
- Entraînement

Sexe	Taille (cm)	Poids (kg)	Pointure (cm)
masculin	182	81.6	30
masculin	180	86.2	28
masculin	170	77.1	30
masculin	180	74.8	25
féminin	152	45.4	15
féminin	168	68.0	20
féminin	165	59.0	18
féminin	175	68.0	23

- hypothèse de distribution Gaussienne pour les lois de probabilités des caractéristiques :

Sexe	Espérance (taille)	Variance (taille)	Espérance (poids)	Variance (poids)	Espérance (pointure)	Variance (pointure)
masculin	178	2.9333×10^1	79.92	2.5476×10^1	28.25	5.5833×10^0
féminin	165	9.2666×10^1	60.1	1.1404×10^2	19.00	1.1333×10^1

- On suppose pour des raisons pratiques que les classes sont équiprobables, à savoir $P(masculin) = P(\text{féminin}) = 0,5$ (selon le contexte, cette hypothèse peut être inappropriée). Si l'on détermine $P(C)$ d'après la fréquence des échantillons par classe dans l'ensemble de données d'entraînement, on aboutit au même résultat.

Classification homme/femme : prédiction (1)

AI for bio
J-P Comet
Introduction
Trees
Clustering
kNN
Bayes
intro
règle de Bayes
Estimations
Prédictions
Exemple
Valeur manquante
SVM
NN
Evaluation
Data Mining Project

Sexe	Taille (cm)	Poids (kg)	Pointure (cm)
inconnu	183	59	20

- Quelle probabilité *a posteriori* est la plus grande ?
  $Pr[(183, 59, 20)|feminin]$ ou $Pr[(183, 59, 20)|masculin]$?

$$\begin{aligned} P_p(M) &= P(M)P(\text{taille}|M)P(\text{poids}|M)P(\text{pointure}|M)/\text{évidence} \\ P_p(F) &= P(F)P(\text{taille}|F)P(\text{poids}|F)P(\text{pointure}|F)/\text{évidence} \end{aligned}$$

- Le terme évidence (constante de normalisation) peut être calculé car la somme des probas *a posteriori* vaut 1.

$$\begin{aligned} \text{évidence} &= P(M)P(\text{taille}|M)P(\text{poids}|M)P(\text{pointure}|M) \\ &+ P(F)P(\text{taille}|F)P(\text{poids}|F)P(\text{pointure}|F) \end{aligned}$$

  Toutefois, on peut ignorer ce terme puisqu'il s'agit d'une constante positive (les lois normales sont toujours positives).
- On peut à présent déterminer le sexe de l'échantillon avec :

$$f_{j,k}(x) = \frac{1}{\sqrt{2\pi\sigma_{k,j}^2}} \exp\left(\frac{-1}{2\sigma_{k,j}^2}(x-\mu_{k,j})^2\right)$$

pour une variable j dans la classe k.

Classification homme/femme : prédiction (2)

AI for bio
J-P Comet
Introduction
Trees
Clustering
kNN
Bayes
intro
règle de Bayes
Estimations
Prédictions
Exemple
Valeur manquante
SVM
NN
Evaluation
Data Mining Project

- Pour la variable taille (t) dans le groupe masculin (m) on a donc :

$$\begin{aligned} P(\text{taille}|M) &= f_{t,m}(x) \\ &= \frac{1}{\sqrt{2\pi \times 2,9333 \times 10^1}} \exp\left(\frac{-1}{2 \times 2.9333 \times 10^1}(183-178)^2\right) \end{aligned}$$

- On réalise ce calcul pour chacune des variables et des groupes :

$P(M)$	=	0.5	$P(F)$	=	0.5
$P(taille\|M)$	=	4.8102×10^{-2}	$P(taille\|F)$	=	7.2146×10^{-3}
$P(poids \mid M)$	=	1.4646×10^{-5}	$P(poids\|F)$	=	3.7160×10^{-2}
$P(pointure\|M)$	=	3.8052×10^{-4}	$P(pointure\|F)$	=	1.1338×10^{-1}
$P_p(M)$	=	1.3404×10^{-10}	$P_p(F)$	=	1.5200×10^{-5}

Comme la proba *a posteriori* pour la classe "féminin" est supérieure à la proba *a posteriori* pour la classe "masculin", l'échantillon est plus probablement de sexe féminin.

⇒ Finir l'exercice 1 du TD sur les classifieurs bayesiens "jouer au tennis" .

Valeur d'attribut manquante

AI for bio
J-P Comet
Introduction
Trees
Clustering
kNN
Bayes
intro
règle de Bayes
Estimations
Prédictions
Exemple
Valeur manquante
SVM
NN
Evaluation
Data Mining Project

- Absence de la valeur d'un attribut dans une donnée dont on veut prédire la classe.
  ⇒ on ne prend pas cet attribut en compte dans l'estimation de la probabilité.
  Par exemple, si on veut prédire la classe de la donnée $x = (Taille = 178, poids = 86)$, on écrira

$$Pr[x|y, \mathcal{X}] \approx Pr[Taille = 178|y, \mathcal{X}] \times Pr[poids = 86|y, \mathcal{X}].$$

- Absence de la valeur d'un attribut dans le jeu d'apprentissage
  - Si, pour une certaine classe, une certaine caractéristique ne prend jamais une valeur donnée dans l'ensemble de données d'entraînement, alors l'estimation de probabilité basée sur la fréquence aura pour valeur zéro.
  - PB puisqu'on aboutit à l'apparition d'un facteur nul lorsque les probabilités sont multipliées.

  ⇒ corriger les estimations de probabilités avec des probabilités fixées à l'avance.

  - D'un point de vue conceptuel, cela n'a pas de sens d'estimer que cette probabilité soit nulle.
  - D'un point de vue pratique, ce 0 pose problème. Il est dû au manque d'exemples correspondants.

Estimateur de Laplace

AI for bio
J-P Comet
Introduction
Trees
Clustering
kNN
Bayes
intro
règle de Bayes
Estimations
Prédictions
Exemple
Valeur manquante
SVM
NN
Evaluation
Data Mining Project

Lorsqu'on a un effectif égal à 0 (pour une classe donnée, et pour un attribut a donné) :

- on ajoute une valeur (par exemple 1) à chaque décompte de la table des effectifs (pour la classe considérée). Il faudra ensuite considérer qu'il y a k exemples de plus (k : nb de valeurs possibles a)
- L'idée générale est :
  - d'ajouter une valeur μ à chaque dénominateur pour l'attribut considéré a et la classe considérée
  - d'ajouter $\frac{\mu}{k}$ à l'effectif associé à chaque valeur de l'attribut considéré et classe considérée
  - Cette quantité, $\frac{\mu}{k}$ de l'attribut considéré, peut être vue comme une probabilité *a priori* de l'observation de chacune des valeurs de l'attribut.
- On n'est donc pas obligé d'avoir une même probabilité *a priori* pour chacune des valeurs de l'attribut, mais des valeurs $p_1, p_2, ...p_n$ pour les n valeurs possibles de l'attribut considéré, du moment que les $p_i \in [1, n]$ sont positifs et que leur somme vaut 1.
- On intègre une connaissance *a priori* dans la méthode de classification.

⇒ Faire l'exercice 2 du TD sur les classifieurs bayesiens "jouer au tennis".

Matrice de confusion

AI for bio
J-P Comet
Introduction
Trees
Clustering
kNN
Bayes
intro
règle de Bayes
Estimations
Prédictions
Exemple
Valeur manquante
SVM
NN
Evaluation
Data Mining
Project

classification

référence	Classe 1	Classe i	Classe n	total lignes
Classe 1	$\mathbf{x}_{11}$	x_{1i}	x_{1n}	N_1
Classe i	x_{i1}	$\mathbf{x}_{ii}$	x_{in}	N_i
Classe n	x_{n1}	x_{ni}	$\mathbf{x}_{nn}$	N_n
total colonnes	M_1	M_i	M_n	N

Généralement, deux classes :

Classe estimée par le classificateur

classe réelle	courrier	pourriel	total lignes
courriel	95 (vrais positifs)	5 (faux négatifs)	100
pourriel	3 (faux positifs)	97 (vrais négatifs)	100
	98	102	200