Cours de Data Science

Marc Tommasi

19 octobre 2022

Outline

Outline

Probabilités conditionnelles et Bayes

- Notations :
  - $P(A)$ la probabilité d'un événement A
  - $P(A \cap B)$ ou $P(A, B)$ la probabilité d'avoir à la fois les événements A et B.
  - $P(A \mid B)$ la probabilité d'avoir l'évenement A sachant B.

Definition des probabilités conditionnelles

$$P(A|B) = \frac{P(A,B)}{P(B)} \text{ ou encore par symétrie } P(B|A) = \frac{P(A,B)}{P(A)}$$

Donc

$$P(A,B) = P(A|B)P(B) = P(B|A)P(A)$$

Théorème de Bayes

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

Petit abus de notations

- Abus de notation : je note donc $P(X = \boldsymbol{x})$ simplement par

$P(\boldsymbol{x})$. (Soit : $P(\boldsymbol{x})$ pour dire qu'une variable aléatoire X prend comme valeur $\boldsymbol{x}$ donc la probabilité de l'événement $X = \boldsymbol{x}$)

- *Exemple* : Soit X soit la variable sur deux attributs x_1 et x_2 qui peuvent prendre uniquement des valeurs binaires. Alors

$P(X)$ désigne $P(X = (0, 0))$, $P(X = (0, 1))$, etc.. notés simplement $P(0, 0)$, $P(0, 1)$, etc. . .

- On va noter aussi $P(\boldsymbol{x})$ ou $P(x_1, x_2)$ pour désigner l'un de ces 4 cas.

Revenant à l'apprentissage

Rappels

- Les données sont générées par une probabilité jointe fixée mais inconnue d'avoir une description des données $\boldsymbol{x}$ et une classe y, écrite $P(\boldsymbol{x}, y)$.
- On veut chercher à résoudre le problème : trouver le meilleur y quand on observe un $\boldsymbol{x}$

La règle de Bayes

- C'est la meilleure règle qu'on puisse imaginer

$$\mathrm{argmax}_y P(y \mid \boldsymbol{x})$$

- Erreur de Bayes : Erreur de cette règle
- c'est la plus petite erreur qu'on puisse faire pour cet apprentissage si les exemples sont décrits par $\boldsymbol{x}$.

Difficile à calculer

- $P(y \mid \boldsymbol{x})$ ne peut être calculée car P est inconnue.
- Si on applique le principe ERM, par la règle de Bayes on a

$$P(y \mid \boldsymbol{x}) = \frac{P(y)P(\boldsymbol{x} \mid y)}{P(\boldsymbol{x})}$$

- On cherche la valeur de y qui maximise cette quantité. Mais $P(\boldsymbol{x})$ ne dépend pas de y. Il suffit de résoudre

$$\mathrm{argmax}_y P(y)P(\boldsymbol{x} \mid y)$$

- Problème : le calcul ne peut être fait efficacement $\boldsymbol{x}$ et des valeurs qu'il peut prendre. Exemple : Dans le cas binaire avec 5 attributs, on a 2^5 possibilités et donc 2 fois 2^5 quantités à estimer pour tous les cas de $P(\boldsymbol{x} \mid y)$.

La chain rule

$$
\begin{aligned}
P(x_1, x_2, \ldots x_n) &= P(x_1 \mid x_2, \ldots x_n) P(x_2, \ldots x_n) \\
&= P(x_1 \mid x_2, \ldots x_n) P(x_2, \mid x_3, \ldots x_n) P(x_3, \ldots x_n) \ldots
\end{aligned}
$$

- Obtenu en appliquant $P(A, B) = P(A|B)P(B)$ de façon répétée quand B est un événement qui peut être une conjonction d'événements
- Par récursion

$$
P(\boldsymbol{x} \mid y) = \prod_{k=1}^{n} P(x_k \mid x_{k-1}, \ldots, x_1, y)
$$

Indépendance Conditionnelle et Naive Bayes

- Si A et B sont indépendants alors $P(A \mid B) = P(A)$.
- Si on fait l'hypothèse que tous les attributs sont indépendants,

c'est-à-dire si x_i et x_j sont indépendants pour tout $i \neq j$, alors le produit s'écrit :

$$P(\boldsymbol{x} \mid y) = \prod_{k=1}^{n} P(x_k \mid y)$$

- Pour calculer chacun de ces $P(x_k \mid y)$ il suffit de compter !
- C'est une approximation forte !
- Mais le calcul est de faible complexité

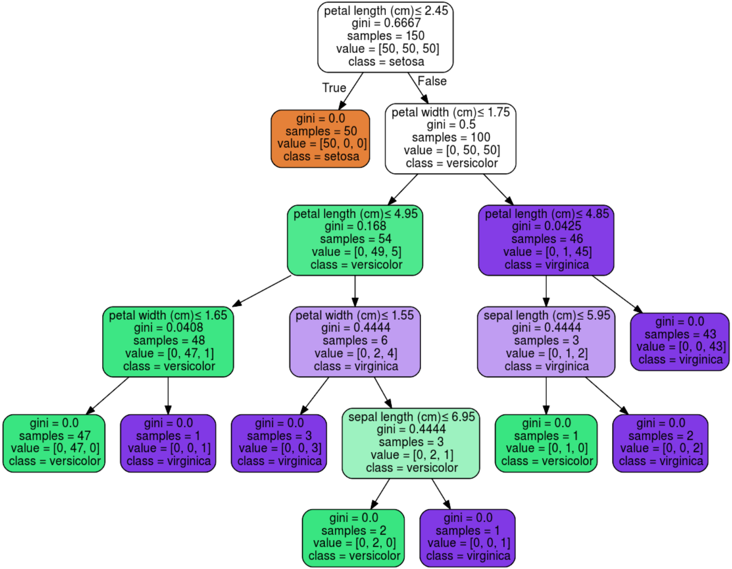
$$\mathrm{argmax}_y P(y) \prod_{k=1}^{n} P(x_k \mid y)$$

Outline

Principes

- Classifieur de $\mathcal{X}$ dans $\mathcal{Y}$
- Modèle sous forme d'un ensemble de règles de décision successives, représentées dans un arbre
- Modèle simple à interpréter quand l'arbre est petit
- Exemple sur le jeu de données iris, 3 classes : setosa, virginica, versicolor ; 150 exemples ; attributs sepal length, sepal width, petal length, petal width.

Exemple



petal length (cm)≤ 2.45
gini = 0.6667
samples = 150
value = [50, 50, 50]
class = setosa
True
False
gini = 0.0
samples = 50
value = [50, 0, 0]
class = setosa
petal width (cm)≤ 1.75
gini = 0.5
samples = 100
value = [0, 50, 50]
class = versicolor
petal length (cm)≤ 4.95
gini = 0.168
samples = 54
value = [0, 49, 5]
class = versicolor
petal length (cm)≤ 4.85
gini = 0.0425
samples = 46
value = [0, 1, 45]
class = virginica
petal width (cm)≤ 1.65
gini = 0.0408
samples = 48
value = [0, 47, 1]
class = versicolor
petal width (cm)≤ 1.55
gini = 0.4444
samples = 6
value = [0, 2, 4]
class = virginica
sepal length (cm)≤ 5.95
gini = 0.4444
samples = 3
value = [0, 1, 2]
class = virginica
gini = 0.0
samples = 43
value = [0, 0, 43]
class = virginica
gini = 0.0
samples = 47
value = [0, 47, 0]
class = versicolor
gini = 0.0
samples = 1
value = [0, 0, 1]
class = virginica
gini = 0.0
samples = 3
value = [0, 0, 3]
class = virginica
sepal length (cm)≤ 6.95
gini = 0.4444
samples = 3
value = [0, 2, 1]
class = versicolor
gini = 0.0
samples = 1
value = [0, 1, 0]
class = versicolor
gini = 0.0
samples = 2
value = [0, 0, 2]
class = virginica
gini = 0.0
samples = 2
value = [0, 2, 0]
class = versicolor
gini = 0.0
samples = 1
value = [0, 0, 1]
class = virginica

Fonctionnement

- On part de la racine, avec un exemple (e.g. 3, 2, 3, 2)
- On passe à travers les tests, chaque test coupe l'espace en 2 parties.
- On désigne donc un partitionnement récursif de l'espace de description
- Chaque feuille donne une étiquette à une partie.
- Bonne visualisation dans le livre de Jake VandenPlas.