2024

TP02 - Rapport

62-62 DATA MINING

HÜNI VICTOR

Sommaire

Figures

Tableaux

Généralités sur le modèle Naïve Bayes

L'objectif du modèle de Naïve Bayes est bien de calculé P(C|A) avec C étant la variable cible et A tous les Attributs prédictifs. Le but étant de proposer une prédiction de C pour n'importe quelle combinaison de valeur dans A.

Pour cela, le modèle Naïves Bayes calcule au préalable

- La fréquence de chaque valeur de la variable cible avec lesquels on peut déterminer $P(C)$, la probabilité la variable cible
- Pour les attribut qualitatif, $P(A|C)$, la probabilité conditionnelle de A pour chaque valeur de C
- Pour les attributs quantitatif, μ_c et σ_C pour chaque valeur de C.

Ces probabilités et valeurs sont par la suite utilisées par le modèle Naïve Bayes pour calculer $P(C|A)$ en utilisant notamment

- La formule du théorème de Bayes $P(C|A) = \frac{P(A|C)\cdot P(C)}{P(A)}$
- L'hypothèse d'indépendance de tous les attributs prédictifs
- L'hypothèse de contribution de même importance au résultat (valeur de C)

Attributs Qualitatifs

Pour les attributs qualitatifs, dans le TP1, $P(C|A)$, la probabilité de la variable A sachant la valeur de la cible A, était calculé grâce à.

$$\mathrm{P}(C|A) = \frac{P(A\cap C)}{P(A)} \text{ ou } \mathrm{P}(A|C) = \frac{P(A\cap C)}{P(C)}$$

On sait donc par définition que

$$P(A \cap C) = P(C|A) \cdot P(A) \text{ ou } P(A \cap C) = P(A|C) \cdot P(C)$$

On a également appris que $P(A)$ est équivalent à la somme des probabilités jointes $P(A \cap C)$ pour toutes les valeurs de C. Ceci peut s'écrire

$$P(A) = \sum_C P(A \cap C)$$

Ainsi nous pouvons déduire que

$$P(A) = \sum_C P(A|C) \cdot P(C)$$

Et donc

$$\sum_C P(A \cap C) = \sum_C P(A|C) \cdot P(C)$$

Maintenant, dans la formule du théorème de Bayes, $P(C|A) = \frac{P(A|C)\cdot P(C)}{P(A)}$ Nous pouvons substituer $P(A)$ et cela nous donne

$$P(C|A) = \frac{P(A|C) \cdot P(C)}{\sum_C P(A|C) \cdot P(C)}$$

Ainsi la différence majeure entre le TP1 et le TP2 provient du fait que :

- Dans le TP1, $P(C|A)$ était directement calulé avec $\mathrm{P}(C|A) = \frac{P(A \cap C)}{P(A)}$ avec $P(A \cap C)$ découvert grâce au jeu de donné complet.
- Dans le TP2, $P(C|A)$ est cette fois ci indirectement déterminé grâce à la formule du théorème de Bayes et les données d'entrainement

Où

$$P(A|C) = \frac{Fréquence\ de\ A\ sachant\ C}{Fréquence\ total\ de\ C}$$

$$P(C) = \frac{Fréquence\ de\ C}{Nombre\ total\ d'instance}$$

$$P(A) = \sum_C P(A|C) \cdot P(C)$$

Exemple

Prenons un exemple, avec les chiffres résultant de l'entrainement du modèle. Utilisons la variable qualitative salary. Voici les probabilité $P(A|C)$ calculé par un modèle entrainé.

Tableau 1 : P(C)

	0	**1**
C (left)	76.00%	24.00%

Tableau 2 : P(A|C)

A (salary) / **C (left)**	**High**	**Low**	**Medium**
0	9.89%	44.45%	45.65%
1	1.94%	6.17%	36.38%

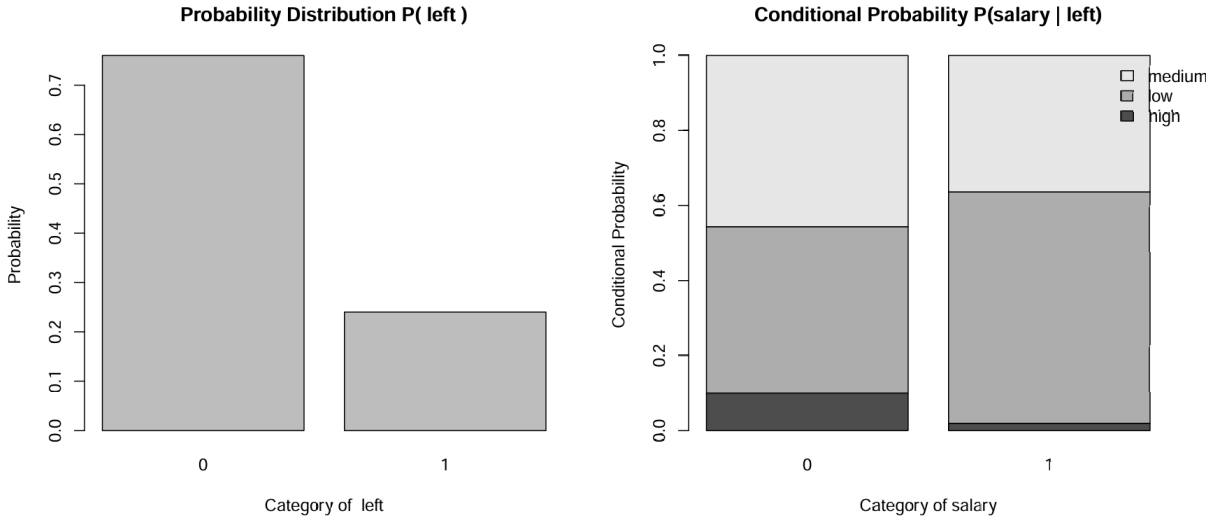

Figure 1 : P(left) vs P(salary left)

$$P(left|salary) = \{P(left = 1|salary), P(left = 0|salary)\}$$

$$P(left = 0|salary) = \frac{P(salary|left = 0) \cdot P(left = 0)}{P(salary)}$$

$$P(left = 1|salary) = \frac{P(salary|left = 1) \cdot P(left = 1)}{P(salary)}$$

Grâce à l'hypothèse d'indépendance, on peut exprimer les termes $P(salary|left = 1)$ et $P(salary|left = 0)$ en faisant le produit les probabilités conditionnelles de chaque classe de la variable salary. Par exemple

$$P(salary|left = 1) = \prod_{salary} P(salary|left = 1)$$

Ce qui grâce au résultat du modèle revient à

$$P(salary|left = 1) = P(high|left = 1) \cdot P(medium|left = 1) \cdot P(low|left = 1)$$

Comme découvert dans notre démonstration précédente, la $P(salary)$ peut être exprimé comme la somme du produit des probabilité conditionnelle et la probabilité de la variable cible :

$$P(salary) = \sum_{left} P(salary|left) \cdot P(left)$$

$$\begin{aligned} P(salary) = {} & P(high|left = 0) \cdot P(left = 0) + P(medium|left = 0) \cdot P(left = 0) \\ & + P(low|left = 0) \cdot P(left = 0) + P(high|left = 1) \cdot P(left = 1) \\ & + P(medium|left = 1) \cdot P(left = 1) + P(low|left = 1) \cdot P(left = 1) \end{aligned}$$

Attributs Quantitatifs

Pour les attributs quantitatifs, le modèle Naïves Bayes calcule et montre la moyenne et l'écart type de l'attribut pour chacune des valeurs de la variable cible. En effet pour calculer $P(C|A)$ pour des variables continues, nous allons injecter ces valeurs dans une distribution gaussienne afin d'estimer la probabilité.

Cette distribution se définit par

$$N(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Exemple

Prenons comme exemple le niveau de satisfaction (satisfaction_level). Les valeurs calculées par le modèle Naïve Bayes sont les suivantes

Tableau 3 : Moyenne et écart type conditionnelle de satisfaction level

	$\boldsymbol{\mu}$	$\boldsymbol{\sigma}$
Left = 0	0.668	0.2156
Left = 1	0.44	0.2644

Nous pouvons injecter ces données dans la formule de la loi normale pour connaitre la densité gaussienne associé. Puis nous déduirons par la suite les probabilité conditionnelles $\mathrm{P(C|A)}$

$$N\left(\text{satisfaction\_level};\mu_{\text{satisfaction\_level}|\text{left}=1},\sigma^2_{\text{satisfaction\_level}|\text{left}=1}\right) = \frac{1}{\sqrt{2\pi\cdot 0.2644^2}}\exp\left(-\frac{(\text{satisfaction\_level}-0.44)^2}{2\cdot 0.2644^2}\right)$$

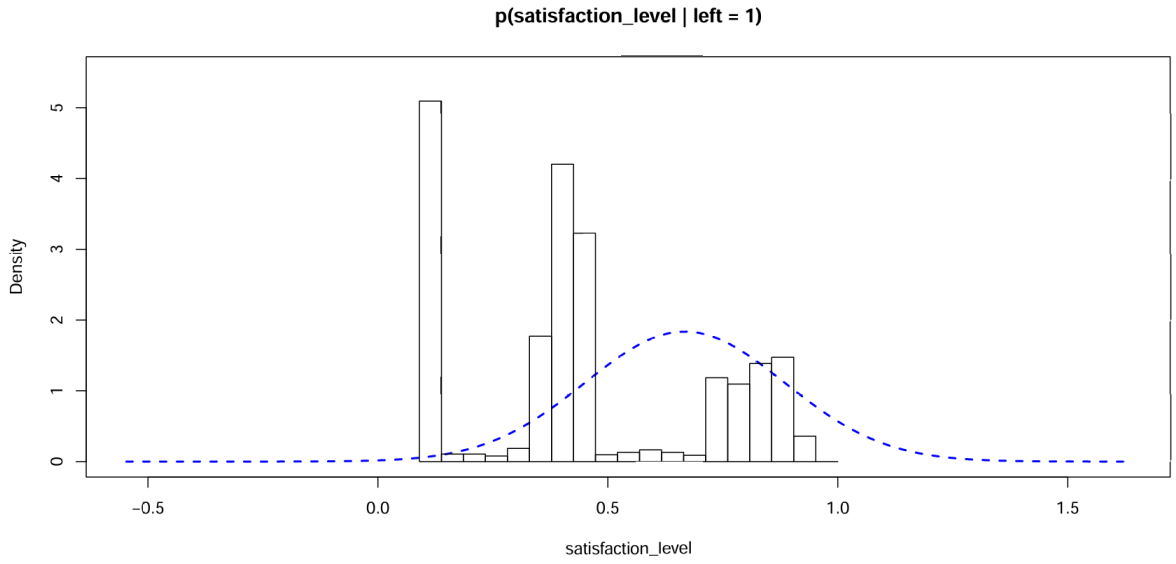

Figure 2 : P (satisfaction_level|left = 1)

$$P(C|A) = \frac{P(A|C) \cdot P(C)}{P(A)}$$

$$\mathrm{P}(\text{left } = 1|\text{satisfaction\_level}) = \frac{P(\text{satisfaction\_level}|\text{left} = 1) \cdot P(\text{left} = 1)}{P(\text{satisafction\_level})}$$

La logique pour déterminer chacun des membres de cette formule est exactement la même que dans le cas des attributs qualitatif. L'unique différence réside dans la façon dans laquelle la Probabilité conditionnelle est calculé. En effet, en lieu et place des probabilités conditionnelles par classes calculé par le modèle, nous devons ici utiliser la moyenne et l'écart types conditionnels par classe pour les déterminer.

Ceci revient à écrire $P(\text{left } = 1|satisfaction\_level)$ sous la forme suivante :

$$\frac{N\left(\text{satisfaction\_level}; \mu_{\text{satisfaction\_level}|\text{left}=1}, \sigma^2_{\text{satisfaction\_level}|\text{left}=1}\right) \cdot P(\text{left} = 1)}{P(\text{satisafction\_level})}$$

Classification des instances

De manière générale, nous avons vu que le modèle de Naïve Bayes calcule les probabilités a posteriori pour chaque classe étant donnée les valeurs d'une instance $\mathrm{x} = \{\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3, \ldots, \mathrm{x}_\mathrm{n}\}$. Pour ce faire Naive bayes utilise l'hypothèse d'indépendance et le théorème de Bayes :

$$P(\, C_k \mid \mathbf{x}) = \frac{P(C_k) \cdot \prod_{i=1}^{n} P(\, x_i \mid C_k\,)}{P(\mathbf{x})}$$

$P(\mathrm{C_k})$ est la probabilité de la variable cible pour une classe donnée

$P(\, x_i \mid C_k\,)$ est la probabilité conditionnelle de l'attribut $\mathrm{x_i}$ sachant la classe de la variable cible. :

- Pour **les attributs qualitatifs**, cette probabilité est directement calculée comme la fréquence de $x_\mathrm{i}$four la classe C_k dans le jeu de donnée d'entrainement
- Pour **les attributs quantitatifs**, cette probabilité est calculée en utilisant la loi normale et sa distribution

$$P(\, x_1 \mid C_k\,) = \frac{1}{\sqrt{2\pi\sigma^2_{C_k}}} \exp\left(-\frac{\left(x_1 - \mu_{C_k}\right)^2}{2\sigma^2_{C_k}}\right)$$

À la suite de ces calculs, la probabilité à posteriori la plus grande permet d'assigné/prédire la classe d'une instance.

Exemple

Reprenons nos deux variables : satisfaction_level et salary

Le modèle va donc calculer $P(\text{left} = 1 \mid \text{satisfaction\_level}, \text{salary})$ en combinant le tout :

$$\frac{P(\text{left} = 1) \cdot N\left(\text{satisfaction\_level}; \mu_{\text{satisfaction\_level}|\text{left}=1}, \sigma^2_{\text{satisfaction\_level}|\text{left}=1}\right) \cdot P(\text{salary}|left = 1)}{P(\text{satisfaction\_level}, \text{salary})}$$

Puis $P(\text{left} = 0 \mid \text{satisfaction\_level}, \text{salary})$ sera également calculé et la plus grande des probabilités permettra au modèle de prédire la classe de l'instance.

Prenons une instance au hasard du jeux de test avec

satisfaction_level = 0.83

salary = medium

Probabilité à priori

Le modèle va tout d'abord calculer

$\text{P}(\text{left} = 0) = 76\%$

$\text{P}(\text{left} = 1) = 24\%$

Probabilités conditionnelles (Quantitatifs)

Puis le modèle va déterminer la probabilité des variables quantitative en utilisant la loi normale en utilisant les moyenne et écart type conditionnelle.

$$P(\text{satisfaction\_level} = 0.83 \mid \text{left} = 0) = \frac{1}{\sqrt{2\pi \cdot 0.2156^2}} \exp\left(-\frac{(0.83 - 0.668)^2}{2 \cdot 0.2156^2}\right)$$

$$P(\text{satisfaction\_level} = 0.83 \mid \text{left} = 1) = \frac{1}{\sqrt{2\pi \cdot 0.2644^2}} \exp\left(-\frac{(0.83 - 0.44)^2}{2 \cdot 0.2644^2}\right)$$

Probabilités conditionnelles (Qualitatifs)

Ensuite, le modèle va calculer les probabilités conditionnelles des variables qualitatives grâce aux fréquences du jeux d'entrainement

$P(\text{salary} = \text{medium}|left = 0) = 45.65\%$

$P(salary = \text{mediu, m}|left = 1) = 36.38\%$

Probabilité jointes

Le modèle utilisera ensuite ses calculs pour déterminer les probabilités jointes $P(satisfaction_level = 0.83, salary = medium)$ lorsque

Pour left = 0

$$P(\text{ satisfaction\_level } = 0.83, \text{salary=medium } | \text{ left } = 0\ ) = P(\text{ satisfaction\_level } = 0.83 \mid \text{left } = 0\ ) \cdot P(\text{ salary= medium } | \text{ left } = 0\ ) \cdot P(\text{left } = 0)$$

Pour left = 1

$$P(\text{ satisfaction\_level } = 0.83, \text{salary=medium } | \text{ left } = 1\ ) = P(\text{ satisfaction\_level } = 0.83 \mid \text{left } = 1\ ) \cdot P(\text{ salary= medium } | \text{ left } = 1\ ) \cdot P(\text{left } = 1)$$

Et enfin

$$P(\text{satisfaction\_level } = 0.83, \text{salary=medium}) = P(\text{ satisfaction\_level } = 0.83, \text{salary=medium } | \text{ left } = 0\ ) + P(\text{ satisfaction\_level } = 0.83, \text{salary=medium } | \text{ left } = 1\ )$$

Sélection de la classe

Avec tous les éléments calculés, le modèle est maintenant capable de calculer et comparé les probabilités à posteriori. (Les valeurs ci-dessous prennent ont été calculé via R en utilisant la fonction predict et en utilisant la fonction predict)

$P(\text{ left } = 0 \mid \text{satisfaction\_level } = 0.83, \text{salary } = \text{ medium }) = 96.40\%$

$P(\text{ left } = 1 \mid \text{satisfaction\_level } = 0.83, \text{salary } = \text{ medium }) = 3.59\%$

Étant donné que $P(\text{left } = 0 \mid x) > P(\text{left } = 1 \mid x)$ alors le modèle prédit que cet employé restera dans l'entreprise.