Lecture 07

Naive Bayes

https://github.com/dalcimar/MA28CP-Intro-to-Machine-Learning

Eventos dependentes

Seja uma caixa contendo **2 bolas brancas** e **3 pretas**. Considere o experimento da retirada aleatória e observância da cor das bolas, com representação de **Bernoulli** (X_1) de ocorrência

- $X_1=0$ caso a bola retirada seja branca
- $X_1=1$ caso a bola retirada seja preta

Pela abordagem clássica, a probabilidade da bola ser branca é:

$$P(X_1 = 0) = \frac{2}{5} \qquad\qquad P(X_1 = 1) = \frac{3}{5}$$

Eventos dependentes

Após a primeira retirada, e **sem reposição**, você retira **mais uma bola**. Qual a probabilidade da segunda bola ser branca?

- A resposta mais intuitiva é **DEPENDE**. De fato, a probabilidade da segunda retirada (X_2) está **condicionada** ao resultado da primeira.
  - Caso a primeira seja branca, restam mais 1 branca e as demais 3 pretas, e, nesse caso, a probabilidade de uma segunda retirada branca é ¼
  - Caso a primeira tenha sido preta, seria ¾

De maneira geral, a probabilidade do evento B condicionada ao evento A é representada por P(B/A). Nesse caso, temos que:

$$P(X_2 = 0 | X_1 = 0) = \frac{1}{4} \qquad P(X_2 = 0 | X_1 = 1) = \frac{2}{4}$$

Dessa forma, dizemos que o evento X_2 é dependente do evento X_1.

Probabilidade conjunta

A **probabilidade conjunta** dos eventos A e B é expressa pela regra do produto: $P(B \cap A) = P(B|A)P(A)$

- também pode ser expressa por $P(B \cap A) = P(B,A)$

Eventos independentes

Considere um experimento, o lançamento de uma moeda honesta e a observância da face virada para cima. Considere a representação **Bernoulli** (X_1) de ocorrência:

- $X_1=0$ caso resultado do lançamento seja coroa
- $X_1=1$ caso o resultado do lançamento seja cara

Por ser justa, a probabilidade do resultado cara e o de coroa é

$$P(X_1 = 1) = \frac{1}{2} \qquad P(X_1 = 0) = \frac{1}{2}$$

Eventos independentes

Em seguida, a moeda é lançada novamente e o resultado representado por X2. Perceba que nesse caso, a probabilidade da ocorrência de cara no segundo lançamento também é igual a $P(X_2=1)=½$ por não ser afetado pelo resultado do primeiro lançamento.

Dizemos que X_2 é independente de X_1, e nesse caso, temos que a regra do produto é:

$$P(X_2, X_1) = P(X_2|X_1)P(X_1)$$

$$P(X_2, X_1) = P(X_2)P(X_1)$$

Permutabilidade

Permutabilidade é a propriedade da **alteração no ordenamento de realizações** em uma **sequência de eventos sem que a probabilidade conjunta seja alterada**.

Considere o lançamento de 5 moedas não viciadas, com representação semelhante à apresentada anteriormente, sendo observado o seguinte resultado (1,0,1,1,0), isto é, (cara,coroa,cara,cara,coroa). Sabemos que os lançamentos são independentes e que a probabilidade conjunta desse evento é:

$$P(1,0,1,1,0) = P(X_5 = 1) \times P(X_4 = 0) \times P(X_3 = 1) \times P(X_2 = 1) \times P(X_1 = 0) = \frac{1}{2^5}$$

Nessa situação, caso fosse observado o evento (1,1,1,0,0), a probabilidade não seria alterada, pois $P(1,0,1,1,0)=P(1,1,1,0,0)=1/2^5$

Dessa forma, **a ordem da ocorrências dos resultados cara não altera a probabilidade conjunta**, desde que seja a mesma quantidade de resultados de “sucesso”. Esse é uma versão não rigorosa do Teorema de De Finetti.

De maneira geral, a **independência** de uma sequência de eventos **garante a permutabilidade** da mesma.

Permutabilidade

No entanto, a **independência não é condição necessária** para a permutabilidade, **apenas suficiente**.

- Isto é, **ainda que uma sequência não seja formada por eventos independente, é possível que sejam permutáveis**.

Considere a mesma representação do exemplo da caixa com bolas apresentado anteriormente para o caso da retirada de 5 bolas sem reposição.

- Como visto, **esses eventos não são independentes**, pois a probabilidade de um determinado evento na sequência, depende do resultado observado nos eventos anteriores.

Permutabilidade

Dessa forma, vamos verificar se o evento (1,0,1,1,0) é permutável. Inicialmente, vamos calcular a probabilidade P(1,0,1,1,0):

$$P(1,0,1,1,0) = P(X_1 = 1) \times P(X_2 = 0|X_1 = 1) \times P(X_3 = 1|X_2 = 0, X_1 = 1) \times$$
$$P(X_4 = 1|X_3 = 1, X_2 = 0, X_1 = 1) \times P(X_5 = 0|X_4 = 1, X_3 = 1, X_2 = 0, X_1 = 1)$$

$$P(1,0,1,1,0) = \tfrac{3}{5} \times \tfrac{2}{4} \times \tfrac{2}{3} \times \tfrac{1}{2} \times \tfrac{1}{1} = \tfrac{1}{10}$$

Agora, vamos calcular a probabilidade P(1,1,1,0,0), alterando o resultado do segundo e o quarto evento:

$$P(1,1,1,0,0) = P(X_1 = 1) \times P(X_2 = 1|X_1 = 1) \times P(X_3 = 1|X_2 = 1, X_1 = 1) \times$$
$$P(X_4 = 0|X_3 = 1, X_2 = 1, X_1 = 1) \times P(X_5 = 0|X_4 = 0, X_3 = 1, X_2 = 1, X_1 = 1)$$

$$P(1,1,1,0,0) = \tfrac{3}{5} \times \tfrac{2}{4} \times \tfrac{1}{3} \times \tfrac{2}{2} \times \tfrac{1}{1} = \tfrac{1}{10}$$

Teorema de Bayes

Considerando os eventos A e B permutáveis, o termo P(A∩B) é igual a P(B∩A) e, dessa forma, pode ser escrita como:

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

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

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

Nesse caso, o probabilidade **P(A)** é denominada **probabilidade a priori**, isto é, a informação sobre o evento A antes que se soubesse algo sobre o evento B. Mais adiante, quando se tenha conhecimento sobre B, a probabilidade relacionada ao evento A deve ser atualizada pela probabilidade do evento B. A probabilidade **P(A/B)** é agora denominada **probabilidade a posteriori**.

Teorema de Bayes

Example

- A doctor knows that meningitis causes stiff neck 50% of the times
- Prior probability of any patient having meningitis is 1/50000
- Prior probability of any patient having stiff neck is 1/20

If a patient has stiff neck (evidence), what's the posterior probability he/she has meningitis?

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

Bayes Theorem

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

Likelihood
Probability of collecting this data when our hypothesis is true

Prior
The probability of the hypothesis being true before collecting data

$$P(H|D) = \frac{P(D|H)\,P(H)}{P(D)}$$

Posterior
The probability of our hypothesis being true given the data collected

Marginal
What is the probability of collecting this data under all possible hypotheses?

Naive Bayes classifier

There **is not a single algorithm**, but a **family of algorithms** based on **a common principle**:

- all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable
    - a fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness, and diameter features.
- In many practical applications, parameter estimation for naive Bayes models uses the method of **maximum likelihood**; in other words, one can work with the naive Bayes model without accepting Bayesian probability or using any Bayesian methods.

Naive Bayes classifier

Abstractly, naïve Bayes is a conditional probability model: given a problem instance to be classified, represented by a vector **x** representing some n features (independent variables), it assigns to this instance probabilities

$$p(C_k \mid x_1, \ldots, x_n)$$

Using Bayes' theorem, the conditional probability can be decomposed as

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

Naive Bayes classifier

The numerator is equivalent to the joint probability model

$$p(C_k, x_1, \dots, x_n)$$

which can be rewritten as follows, using the chain rule for repeated applications of the definition of conditional probability:

$$\begin{aligned}
p(C_k, x_1, \dots, x_n) &= p(x_1, \dots, x_n, C_k) \\
&= p(x_1 \mid x_2, \dots, x_n, C_k)\, p(x_2, \dots, x_n, C_k) \\
&= p(x_1 \mid x_2, \dots, x_n, C_k)\, p(x_2 \mid x_3, \dots, x_n, C_k)\, p(x_3, \dots, x_n, C_k) \\
&= \cdots \\
&= p(x_1 \mid x_2, \dots, x_n, C_k)\, p(x_2 \mid x_3, \dots, x_n, C_k) \cdots p(x_{n-1} \mid x_n, C_k)\, p(x_n \mid C_k)\, p(C_k)
\end{aligned}$$

Naive Bayes classifier

Now the "naïve" conditional independence assumptions come into play: assume that all features in **x** are mutually independent, conditional on the category C_k. Under this assumption,

$$
\begin{aligned}
p(C_k \mid x_1, \dots, x_n) &\propto p(C_k, x_1, \dots, x_n) \\
&\propto p(C_k)\, p(x_1 \mid C_k)\, p(x_2 \mid C_k)\, p(x_3 \mid C_k)\, \cdots \\
&\propto p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)\,,
\end{aligned}
$$

$$\boxed{P(B \mid A_1, \dots, A_n) = \frac{P(B) \cdot \prod_{i=1}^{n} P(A_i \mid B)}{\prod_{i=1}^{n} P(A_i)}}$$

$$\hat{y} = \underset{k \in \{1, \dots, K\}}{\operatorname{argmax}}\ p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k).$$

Naive Bayes classifier

Para várias variáveis aleatórias A_1, A_2, ..., A_n, e B: raciocínio análogo

- $P(B|A_1,A_2,...,A_n) = P(A_1,A_2,...,A_n|B)*P(B) / P(A_1,A_2,...,A_n)$

Se as vars aleatórias A1 ... An forem independentes entre si tem-se:

- $P(A_1,A_2,...,A_n) = P(A_1)*P(A_2)* \ldots *P(A_n)$ [independência]
- $P(A_1,A_2,...,A_n|B) = P(A_1|B)*P(A_2|B)* \ldots * P(A_n|B)$ [independência condicional]

Naive Bayes classifier

Naive Bayes é um algoritmo que utiliza o Teorema de Bayes com a hipótese de independência entre atributos

- Porque assumir independência entre atributos $A_1,\ldots,A_n$?
  - estimar probabilidades conjuntas $P(A_1,A_2,...,A_n)$ e $P(A_1,A_2,...,A_n|B)$ demandaria uma quantidade mínima de exemplos de cada combinação possível de valores de $A_1,A_2,...,A_n$
  - impraticável, especialmente para quantidades elevadas de atributos !
- Apesar da hipótese ser quase sempre violada, o método (Naive Bayes) se mostra bastante competitivo na prática!

Outlook (A_1)			Temperature (A_2)			Humidity (A_3)			Windy (A_4)			Play (B)	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

$$P(B \mid A_1, \ldots, A_n) = \frac{P(B) \cdot \prod_{i=1}^{n} P(A_i \mid B)}{\prod_{i=1}^{n} P(A_i)}$$

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Outlook (A_1)			Temperature (A_2)			Humidity (A_3)			Windy (A_4)			Play (B)	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	???

P(Yes|Sunny, Cool, High, True) = (2/9 × 3/9 × 3/9 × 3/9 × 9/14) / P(Sunny, Cool, High, True)

P(No|Sunny, Cool, High, True) = (3/5 × 1/5 × 4/5 × 3/5 × 5/14) / P(Sunny, Cool, High, True)

P(Yes|Sunny, Cool, High, True) = **0.0053** / P(Sunny, Cool, High, True)

P(No|Sunny, Cool, High, True) = **0.0206** / P(Sunny, Cool, High, True)

Play = No

$$\hat{y} = \underset{k \in \{1,\dots,K\}}{\operatorname{argmax}} \ p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k).$$

GAUSSIAN NAIVE BAYES CLASSIFIER

"Gaussian" because this is a normal distribution

This is our prior belief

$$p(\text{class} \mid \text{data}) = \frac{p(\text{data} \mid \text{class}) \times p(\text{class})}{p(\text{data})}$$

We don't calculate this in naive bayes classifiers

ChrisAlbon