Naive Bayes

25 de Abril de 2015

Formulas de probabilidade básicas

$$
\begin{aligned}
P(A \wedge B) &= P(A \mid B)\, P(B) = P(B \mid A)\, P(A) && (1)\\
P(A \vee B) &= P(A) + P(B) - P(A \wedge B) && (2)\\
P(h \mid D) &= \frac{P(D \mid h)\, P(h)}{P(D)} && (3)\\
P(B) &= \sum_{i=1}^{n} P(B \mid A_i)\, P(A_i) \text{ se} \sum_{i=1}^{n} P(A_i) = 1 && (4)
\end{aligned}
$$

Classificador Naive Bayes

Seja $a_1, \ldots, a_n$ um conjunto de atributos e $v_i \in V$ um conjunto de valores de classificação possíveis. Então, de acordo com o teorema de *Bayes*, a probabilidade desse conjunto de atributos pertencer à classe v_i é

$$P(v_i \mid a_1, \ldots, a_n) = \frac{P(a_1, \ldots, a_n \mid v_i)\, P(v_i)}{P(a_1, \ldots, a_n)}$$

O nosso objetivo é descobrir qual é o v_i que maximiza $P(v_i \mid a_1, \ldots, a_n)$, isto é, pretendemos descobrir

$$
\begin{aligned}
\underset{v_i \in V}{\arg\max}\, P(v_i \mid a_1, \ldots, a_n) &= \underset{v_i \in V}{\arg\max} \frac{P(a_1, \ldots, a_n \mid v_i)\, P(v_i)}{P(a_1, \ldots, a_n)}\\
&= \underset{v_i \in V}{\arg\max}\, P(a_1, \ldots, a_n \mid v_i)\, P(v_i)
\end{aligned}
$$

Reparem que não precisamos de dividir por $P(a_1, \ldots, a_n)$ visto que o termo é o mesmo para todos os v_i e por isso não é necessário. O problema é que o custo de estimação de $P(a_1, \ldots, a_n \mid v_i)$ é muito elevado, precisaríamos de muitos dados para o conseguir. Assim, este algoritmo assume que os valores dos atributos são condicionalmente independentes, e assim:

$$P(a_1, \ldots, a_n \mid v_i) \approx \prod_{k=1}^{n} P(a_k \mid v_i)$$

Temos assim o Classificador Naive Bayes (NB):

$$v_{\mathrm{NB}} = \underset{v_i \in V}{\arg\max}\, P(v_i) \prod_{k=1}^{n} P(a_k \mid v_i)$$

Outlook	**Temperature**	**Humidity**	**Wind**	**Play Tennis?**
Overcast	Hot	Normal	Weak	Yes
Overcast	Mild	High	Strong	Yes
Sunny	Mild	Normal	Strong	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Cool	Normal	Weak	Yes
Overcast	Cool	Normal	Strong	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Mild	High	Weak	Yes
Overcast	Hot	High	Weak	Yes
Rain	Cool	Normal	Strong	No
Sunny	Hot	High	Strong	No
Sunny	Hot	High	Weak	No
Rain	Mild	High	Strong	No
Sunny	Mild	High	Weak	No

Tabela 1: Exemplo de um dataset

Exemplo

O conjunto de dados apresentado na tabela 1 refere-se à decisão de jogar ténis de acordo com vários atributos, Outlook, Temperature, Humidity e Wind. Se estivermos a utilizar um classificador Naive Bayes qual é a classe que ele sugere para o caso dado a seguir:

Outlook	**Temperature**	**Humidity**	**Wind**
Sunny	Cool	High	Strong

Neste caso, o que pretendemos é descobrir v_{NB} de acordo com a formula dada acima:

$$v_{\mathrm{NB}} = \underset{v_i \in \{Yes,No\}}{\arg\max}\; P(v_i) \times P(Outlook = Sunny \,|\, v_i) \times P(Temperature = Cool \,|\, v_i) \times P(Humidity = High \,|\, v_i) \times P(Wind = Strong \,|\, v_i)$$

Para isso vamos calcular as frequências dos valores de cada atributo para cada um dos v_i tal como se vê na tabela 2. A seguir calculamos as frequências relativas (tabela 3) e podemos finalmente calcular os valores esperados:

Outlook	**Yes**	**No**	**Temperature**	**Yes**	**No**	**Humidity**	**Yes**	**No**	**Wind**	**Yes**	**No**
Sunny	2	3	Hot	2	2	High	3	4	Strong	3	3
Overcast	4	0	Mild	4	2	Normal	6	1	Weak	6	2
Rain	3	2	Cool	3	1						
Total	9	5	**Total**	9	5	**Total**	9	5	**Total**	9	5

Tabela 2: Frequências absolutas do dataset da tabela 1

Outlook	**Yes**	**No**	**Temperature**	**Yes**	**No**	**Humidity**	**Yes**	**No**	**Wind**	**Yes**	**No**
Sunny	$\frac{2}{9}$	$\frac{3}{5}$	Hot	$\frac{2}{9}$	$\frac{2}{5}$	High	$\frac{3}{9}$	$\frac{4}{5}$	Strong	$\frac{3}{9}$	$\frac{3}{5}$
Overcast	$\frac{4}{9}$	$\frac{0}{5}$	Mild	$\frac{4}{9}$	$\frac{2}{5}$	Normal	$\frac{6}{9}$	$\frac{1}{5}$	Weak	$\frac{6}{9}$	$\frac{2}{5}$
Rain	$\frac{3}{9}$	$\frac{2}{5}$	Cool	$\frac{3}{9}$	$\frac{1}{5}$						

Tabela 3: Frequências relativas do dataset da tabela 1

$$\begin{aligned}
P(PlayTennis = Yes) \quad \times \quad & P(Outlook = Sunny \,|\, PlayTennis = Yes) \times \\
& P(Temperature = Cool \,|\, PlayTennis = Yes) \times \\
& P(Humidity = High \,|\, PlayTennis = Yes) \times \\
& P(Wind = Strong \,|\, PlayTennis = Yes) = \\
= \quad & \frac{9}{14} \times \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} = 0.00529
\end{aligned}$$

$$\begin{aligned}
P(PlayTennis = No) \quad \times \quad & P(Outlook = Sunny \,|\, PlayTennis = No) \times \\
& P(Temperature = Cool \,|\, PlayTennis = No) \times \\
& P(Humidity = High \,|\, PlayTennis = No) \times \\
& P(Wind = Strong \,|\, PlayTennis = No) = \\
= \quad & \frac{5}{14} \times \frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} = 0.02057
\end{aligned}$$

Sendo assim, neste caso decidiríamos não ir jogar ténis. Para além disso, podemos estimar a probabilidade condicional do valor ser *não* dados os atributos observados:

$$\frac{0.02057}{0.02057 + 0.00529} = 79.54\%$$

Estimar probabilidades

Até agora estimámos as probabilidades pelas frequências relativas simples. Isto é, para estimar a probabilidade $P(Humidity = Normal \,|\, PlayTennis = No)$ dividimos o número de ocorrências de $Humidity = Normal$ quando $PlayTennis = No$ pelo número de ocorrências de $PlayTennis = No$ o que nos dá $P(Humidity = Normal \,|\, PlayTennis = No) = \frac{1}{5}$.

O problema desta estimação é quando a probabilidade real é muito baixa, por exemplo 0.02. Neste caso, é muito provável que a frequência relativa seja 0, e ao multiplicar 0 por outros valores obtém-se sempre 0. Nestes casos, costuma-se utilizar a **estimativa m da probabilidade**:

$$\frac{n_c + m \times p}{n + m}$$

em que n_c é o número de ocorrências que pretendemos (no nosso caso, quando $Humidity = Normal$ se $PlayTennis = No$) e n é o número total de exemplos para os quais $PlayTennis = No$, p é a *estimativa à priori da probabilidade* que pretendemos determinar e m é uma constante chamada *tamanho equivalente da amostra* que determina o nosso grau de confiança em p relativamente à nossa amostra. Tipicamente escolhemos o valor de p assumindo a distribuição uniforme. Assim, caso existam k valores possíveis, assumimos que $p = \frac{1}{k}$. Por exemplo, ao estimar o valor de $P(Humidity = Normal \,|\, PlayTennis = No)$ sabemos que $Humidity$ tem dois valores possíveis ($High$ e $Normal$) e por isso $p = \frac{1}{2}$. Se m for zero então a estimativa m da probabilidade dá simplesmente $\frac{n_c}{n}$. Se m for 2 e assumindo para p o valor à priori segundo a distribuição uniforme, a estimativa m daria

$$\frac{n_c + 1}{n + 2}$$

A razão porque se chama a m o tamanho equivalente da amostra é porque aos n valores observados na amostra se juntam m valores virtuais segundo a probabilidade p.

Aprendendo a classificar texto

O NB é um método bastante bem sucedido para classificar textos. Neste caso, o que se faz mais frequentemente é contar as frequências de cada palavra p_k pertencente ao texto que se pretende classificar e calcular a probabilidade de pertencer à classe c_i mediante a fórmula:

$$P(c_i \,|\, p_1, \ldots, p_n) = P(c_i) \times \prod_{k=1}^{n} P(p_k \,|\, c_i)$$

em que a probabilidade de pertencer à classe c_i é dada pela frequência relativa dos n_i textos pertencentes à classe c_i sobre todos os n textos usados no treino

$$P(c_i) = \frac{n_i}{n}$$

e a probabilidade da palavra p_k aparecer nos textos da classe c_i é dada pela estimativa m em que m é o número de palavras diferentes que ocorrem nos textos que foram usados no treino sendo dada por

$$P(p_k \,|\, c_i) = \frac{n_k + 1}{n + |Vocabulario|}$$

em que n_k é o número total de vezes em que a palavra p_k aparece nos textos da classe c_i, n é o número total de palavras que ocorre nos textos da classe c_i e $|Vocabulario|$ é o número total de palavras diferentes que ocorre nos textos.

Agora aplicamos o algoritmo para descobrir a classe:

$$c_{NB} = \arg\max_{i \in classes} P(c_i) \times \prod_{k=1}^{n} P(p_k \,|\, c_i)$$

Para evitar as probabilidades muito baixas, que podem tornar a multiplicação dos valores zero e por isso impedir o algoritmo de funcionar corretamente, utilizam-se logaritmos:

$$c_{NB} = \underset{i \in classes}{\arg\max} \log P(c_i) + \sum_{k=1}^{n} \log P(p_k \mid c_i)$$

Vamos ver um exemplo para nos ajudar a perceber.

texto	**classe**
a baixa do porto	Porto
o mercado do bolhão é no porto	Porto
a câmara do porto fica no centro do porto	Porto
a baixa de lisboa	Lisboa
o porto de lisboa	Lisboa

Vocabulário	14

	Porto	Lisboa
Exemplos	3	2
Palavras	20	8
baixa	1	1
porto	4	1
mercado	1	0
bolhão	1	0
câmara	1	0
baixa	1	1
lisboa	0	2

$$P(Porto) = \frac{3}{5}$$
$$P(Lisboa) = \frac{2}{5}$$
$$P(porto|Porto) = \frac{4+1}{20+14} = 0.1471$$
$$P(porto|Lisboa) = \frac{1+1}{8+14} = 0.0909$$
$$p(mercado|Porto) = \frac{1+1}{20+14} = 0.0588$$
$$p(mercado|Lisboa) = \frac{0+1}{8+14} = 0.0455$$

E assim, ao tentar classificar o texto *porto porto porto mercado* teríamos:

$$\begin{array}{ll} Porto & P(Porto) \times p(porto|Porto)^3 \times P(mercado|Porto) = 0.6 \times 0.1471^3 \times 0.0588 = 0.0001123 \\ Lisboa & P(Lisboa) \times p(porto|Lisboa)^3 \times P(mercado|Lisboa) = 0.4 \times 0.0909^3 \times 0.0455 = 0.0000137 \end{array}$$

O que nos permitiria concluir que estavamos a falar sobre o Porto com probabilidade

$$\frac{0.0001123}{0.0001123 + 0.0000137} = 89.13\%$$

Técnicas específicas para deteção de spam

Quando se pretende detetar spam calcula-se a probabilidade de uma dada palavra P pertencer a um texto de spam da seguinte forma:

$$P(S \mid P) = \frac{P(P \mid S) \times P(S)}{P(P \mid S) \times P(S) + P(P \mid N) \times P(N)}$$

em que S corresponde a spam, N a não spam, e P a uma dada palavra. Caso se assuma que $P(S) = P(N) = 0.5$ (na verdade as estimativas indicam que $P(S) \approx 0.8$), a fórmula pode ser simplificada da seguinte forma:

$$P(S \mid P) = \frac{P(P \mid S)}{P(P \mid S) + P(P \mid N)}$$

A probabilidade de um texto que contém as palavras $P_1, \ldots, P_n$ ser spam é dada por:

$$P(S \mid P_1, \ldots, P_n) = \frac{\prod_{i=1}^{n} P(S \mid P_i)}{\prod_{i=1}^{n} P(S \mid P_i) + \prod_{i=1}^{n} P(N \mid P_i)}$$

Repare que como só existem as classes S e N, a probabilidade $P(N \mid P_i) = 1 - P(S \mid P_i)$.