## 4 Classificador Naive Bayes

Given the number of times on which an unknown event has happened and failed: Required the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named.

Thomas Bayes, 1764

O Classificador Naive Bayes é provavelmente o classificador mais utilizado em *Machine Learning* (04). O classificador é denominado ingênuo (naive) por assumir que os atributos são condicionalmente independentes, ou seja, a informação de um evento não é informativa sobre nenhum outro.

Apesar desta premissa "ingênua" e simplista, o classificador reporta o melhor desempenho em várias tarefas de classificação. Este fenômeno é discutido em (13, 04).

Até as primeiras décadas do século XVIII, problemas relacionados a probabilidade de certos eventos, dadas certas condições, estavam bem resolvidos. Por exemplo, dado um número específico de bolas negras e brancas em uma urna, qual é a probabilidade de eu sortear uma bola preta? Tais problemas são chamados de problemas de "forward probability". Porém, logo, o problema inverso começou a chamar a atenção dos matemáticos da época: Dado que uma ou mais bolas foram sorteadas, o que pode ser dito sobre o número de bolas brancas e pretas na urna? (02)

Thomas Bayes, um ministro inglês do século XVIII, foi o primeiro a formalizar uma teoria para problemas desta natureza em (01) vista como revolucionária no meio científico da época.

É exatamente este pensamento inverso que buscamos ao treinar um classificador de textos. Dado que tenho exemplos de texto de cada classe. O que posso inferir sobre o processo gerador destes textos?

Nos experimentos, examinamos dois modelos de classificação bayesiana, o modelo binário e o modelo multinomial. Para um estudo detalhado sobre raciocínio Bayesiano vide (22).

# 4.1 Fundamentos Teóricos

Ao treinar o classificador Naive Bayes calculamos uma distribuição geradora  $\Pr(d|c)$  para cada classe  $c \in \{-1,1\}$ . Na fase de classificação, simplesmente calculamos qual distribuição tem a maior probabilidade de ter gerado cada documento.

Nas máquinas Bayesianas, a criação de documentos é modelada como o seguinte processo:

- 1. Cada classe c possui uma probabilidade a priori associada Pr(c) tal que  $\sum_{c} Pr(c) = 1$ . O autor primeiro escolhe aleatoriamente se vai gerar uma avaliação positiva ou negativa seguindo estas probabilidades.
- 2. Dado que existe uma distribuição de documentos  $\Pr(d|c)$  associada a classe c escolhida, esta distribuição é usada para gerar o documento.

Sendo assim, a probabilidade de gerarmos um documento da classe c é Pr(c) Pr(d|c). Finalmente, dado um documento d a probabilidade a posteriori de que d foi gerado da classe c é:

$$Pr(c|d) = \frac{Pr(c) \ Pr(d|c)}{\sum_{\gamma} Pr(\gamma) \ Pr(d|\gamma)}$$
(4-1)

onde  $\gamma$  itera sobre ambas as classes de forma que  $\Pr(c|d)$  torna-se uma medida de probabilidade apropriada.

 $\Pr(\gamma|c)$  é estimado através da modelagem das distribuição dos termos sobre as classes gerando um conjunto de parâmetros que chamaremos de  $\Theta$ . Nossa estimativa de  $\Theta$  é baseado nos termos contidos nos documentos de treinamento

Após observarmos os dados de treinamento D, nossa distribuição a posteriori para  $\Theta$  pode ser expressa por  $\Pr(\Theta|D)$ .

Dadas estas definições e dado um documento d, a probabilidade de d pertencer a classe c é dada por:

$$Pr(c|d) = \sum_{\Theta} Pr(c|d,\Theta) Pr(\Theta|D)$$

$$= \sum_{\Theta} \frac{Pr(c|\Theta) Pr(d|c,\Theta)}{\sum_{\gamma} Pr(\gamma,\Theta) Pr(d|\gamma,\Theta)} Pr(\Theta|D)$$
(4-2)

### 4.1.1

### Modelo Binário

No modelo binário, assumimos que cada documento é representado por um vetor de atributos binários de modo que cada atributo indica a ocorrência ou não de um evento no documento.

Neste modelo,  $\phi_{c,t}$  indica a probabilidade de um documento da classe c mencionar o termo t pelo menos uma vez. Assim:

$$\Pr(d|c) = \prod_{t \in d} \phi_{c,t} \prod_{t \in W, t \notin d} (1 - \phi_{c,t})$$

$$\tag{4-3}$$

onde W é o conjunto de features. Para evitar termos que calcular  $\prod_{t \in W, t \notin d} (1 - \phi_{c,t})$  para cada documento classificado, reescrevemos a equação acima como:

$$\Pr(d|c) = \prod_{t \in d} \frac{\phi_{c,t}}{1 - \phi_{c,t}} \prod_{t \in W} (1 - \phi_{c,t})$$
 (4-4)

Precomputamos  $\prod_{t \in W} (1 - \phi_{c,t})$  para todo c, e somente computamos o primeiro produtório em tempo de classificação. Com isso, o tempo de classificação de um documento fica linear em função do número de features presentes no documento ao invés de linear em função de |W|.

Podemos imaginar este modelo como sendo dois geradores de documentos (um de textos de sentimento positivo e outro negativo). Ao gerar um documento, o gerador lança uma moeda para cada feature para decidir se a inclui ou não no documento. Obviamente que para cada termo  $w \in W$  temos um moeda diferente e é justamente as naturezas destas moedas que são determinadas na fase de treinamento.

# 4.1.2 Modelo Multinomial

No modelo multinomial, assumimos que cada documento é representado por um vetor de atributos inteiros caracterizando o número de vezes que cada feature ocorre no documento.

Podemos imaginar o modelo multinomial de forma que cada gerador controle uma "roleta" com |W| faixas. Ao gerar um documento, o gerador primeiro escolhe um comprimento l para um documento e depois gira a roleta l vezes para definir quais palavras colocará no texto. Durante a fase de treinamento calculamos as dimensões de cada faixa das roletas das classes geradoras.

Seja  $\theta_{c,t}$  a probabilidade da faixa  $t \in W$  da roleta ser sorteada num giro. Seja n(d,t) o número de vezes que t ocorre no documento d que possui comprimento  $l_-d = \sum_t n(d,t)$ . O comprimento do documento é uma variável aleatória denominada L e assumimos que segue uma distribuição apropriada para cada classe, neste modelo,

$$Pr(d|c) = Pr(L = l_d|c) Pr(l_d, c)$$

$$= Pr(L = l_d|c) {l_d \choose {n(d,t)}} \prod_{t \in d} \theta_t^{n(d,t)}$$
(4-5)

onde  $\binom{l_d}{\{n(d,t)\}} = \frac{l_d!}{n(d,t-1)!n(d,t-2)!\dots}$  é o coeficiente multinomial que pode ser desprezado uma vez que tem o mesmo valor para todo c.

Assumimos para o nosso caso, uma mesma distribuição de comprimento para ambas as classes. Apesar das avaliações positivas serem em média maiores, não achamos que seja válido considerar este critério para estimar a qualidade dos modelos. Portanto, ignoramos também o termo  $\Pr(L=l_d|c)$  nos experimentos.

Concluímos aqui a descrição do classificador Naive Bayes. Para uma referência mais completa sobre raciocínio Bayesiano vide (22).