Universidade Federal de Viçosa
Centro de Ciências Exatas e Tecnológicas
Departamento de Informática

**Classificador Bayesiano Ingênuo**
Levi H. S. Lelis
levi.lelis@ufv.br

# Revisão Probabilidade

## Probabilidade a Priori

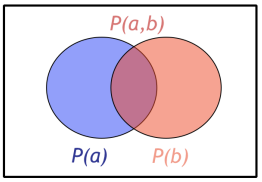
A probabilidade a priori de ocorrência de um evento (e.g., dois dados de 6 faces somar 11) é dado pela soma das probabilidades das saídas que resultam naquele evento.

$$P(1) = \frac{1}{6}$$
$$P(2 \wedge 3) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$
$$P(\text{soma igual a } 11) = P(5 \wedge 6) + P(6 \wedge 5) = \frac{1}{36} + \frac{1}{36} = \frac{1}{18}$$
$$P(\text{duplos}) = P(1 \wedge 1) + P(2 \wedge 2) + P(3 \wedge 3) + P(4 \wedge 4) + P(5 \wedge 5) + P(6 \wedge 6) = 6 \times \frac{1}{36} = \frac{1}{6}$$

## Probabilidade Condicional

Meningite faz o paciente ter uma rigidez no pescoço em 70% dos casos. A probabilidade a priori de um paciente ter meningite é 1 em 50000 e a probabilidade a priori de um paciente ter rigidez no pescoço é de 1%. Qual é a probabilidade de o paciente ter meningite dado que ele tem rigidez no pescoço?

Para responder a pergunta acima, considere o diagrama de eventos abaixo onde $a$ represente meningite e $b$ rigidez no pescoço. Se o paciente tem rigidez no pescoço, então estamos lidando com os eventos na parte direita do diagrama, em vermelho. A probabilidade de o paciente possuir meningite dado que possui pescoço rígido, $P(a|b)$, é dada por $\frac{P(a,b)}{P(b)}$.




Além que $P(a|b) = \frac{P(a,b)}{P(b)}$, do diagrama acima podemos extrair a seguinte equação:

$$P(b|a) = \frac{P(a,b)}{P(a)}$$
$$P(a,b) = P(b|a)P(a)$$

Igualando os valores de $P(a, b)$ obtemos:

$$P(b|a)P(a) = P(a|b)P(b)$$
$$P(a|b) = \frac{P(b|a)P(a)}{P(b)}$$

A equação acima é conhecida como o Teorema de Bayes. Essa talvez seja a equação mais utilizada em IA.

Voltando ao exemplo de diagnóstico da meningite, temos que $P(b|a) = 0.7$, $P(a) = \frac{1}{50000}$ e $P(b) = 0.01$, resultando em:

$$P(a|b) = \frac{\frac{1}{50000} \times 0.7}{0.01} = 0.0014$$

### Dependência e Independência de Eventos

Um estudo revelou que a probabilidade de selecionar pessoas proficientes em leitura dado que as pessoas possuem braços compridos é maior do que pessoas selecionadas aleatoriamente da população. Esse estudo sugere inicialmente que a proficiência de leitura é dependente do comprimento do braço das pessoas.

O que acontece é que muitas pessoas com braços curtos ainda não foram alfabetizadas (crianças muito novas, por exemplo) e por isso ainda não são proficientes em leitura. Uma vez que passamos a condicionar na idade das pessoas, a relação entre proficiência em leitura e comprimento de braço desaparece.

De uma forma geral, dadas variáveis aleatórias $Y$ (e.g., proficiência em leitura), $X$ (comprimento do braço) e $Z$ (idade), dizemos que $Y$ é condicionalmente independente de $X$ dado $Z$:

$$P(Y|X, Z) = P(Y|Z)$$

Utilizaremos a independência condicional para construir um algoritmo de classificação.

## Classificador Bayesiano Ingênuo (CBI)

No problema da meningite nós tínhamos os *efeitos* e calculamos a probabilidade da *causa*. O CBI utiliza a mesma ideia para classificar objetos em classes distintas. Diferente do problema da meningite, muitas vezes teremos várias evidências (características) e não apenas uma. Por exemplo, podemos utilizar a ocorrência de palavras em um email como evidência de que o email é SPAM ou HAM.

Considere os emails da tabela abaixo onde cada linha representa um email.

| SPAM | HAM |
|---|---|
| offer is secret | play sports today |
| click secret link | went sports today |
| secret sports link | secret sports event |
| | sports is today |
| | sports costs money |

O CBI usa a fórmula de Bayes e as probabilidades aprendidas a partir de um conjunto de treinamento para, dado um novo objeto com um conjunto de evidências (características), classificar o objeto.

Por exemplo, dada a mensagem "offer click link", nós gostaríamos de calcular a probabilidade dessa mensagem ser um SPAM. De acordo com a fórmula de Bayes nós temos:

$$P(\text{SPAM}|\text{offer, click, link}) = \frac{P(\text{SPAM})P(\text{offer, click, link}|\text{SPAM})}{P(\text{offer, click, link})}$$

Da mesma forma, podemos calcula a probabilidade de a mensagem ser HAM:

$$P(\text{HAM}|\text{offer, click, link}) = \frac{P(\text{HAM})P(\text{offer, click, link}|\text{HAM})}{P(\text{offer, click, link})}$$

Como estamos lidando com um problema de classificação, não precisamos calcular os valores exatos de $P(\text{HAM}|\text{offer, click, link})$ e $P(\text{SPAM}|\text{offer, click, link})$, basta saber qual dos dois é maior. Por isso, como os dois são divididos por $P(\text{offer, click, link})$, nós podemos remover o denominador sem alterar a relação dos valores $P(\text{SPAM}|\text{offer, click, link})$ e $P(\text{HAM}|\text{offer, click, link})$.

Então a nossa tarefa passa ser a de calcular $P(\text{SPAM})P(\text{offer, click, link}|\text{SPAM})$ e $P(\text{HAM})P(\text{offer, click, link}|\text{HAM})$.

O Teorema de Bayes seguido da regra da cadeia nos dá o seguinte resultado.

$$\begin{aligned}
P(\text{SPAM})P(\text{offer, click, link}|\text{SPAM}) &= P(\text{offer, click, link, SPAM}) \\
&= P(\text{offer}|\text{click, link, SPAM})P(\text{click, link, SPAM}) \\
&= P(\text{offer}|\text{click, link, SPAM})P(\text{click}|\text{link, SPAM})P(\text{link, SPAM}) \\
&= P(\text{offer}|\text{click, link, SPAM})P(\text{click}|\text{link, SPAM})P(\text{link}|\text{SPAM})P(\text{SPAM})
\end{aligned}$$

Alguns valores da equação acima são fáceis de serem estimados. Por exemplo, os seguintes valores são obtidos através da máxima verosimilhança com os dados de treinamento.

$$\begin{aligned}
P(\text{SPAM}) &= \frac{3}{8} \\
P(\text{link}|\text{SPAM}) &= \frac{2}{9}
\end{aligned}$$

No entanto, quanto mais evidências tivermos, maior a chance de termos uma probabilidade igual a zero. Por exemplo, $P(\text{offer}|\text{click, link, SPAM})$ mede a probabilidade da ocorrência da palavra "offer" dado que a mensagem é um SPAM e que contém as palavras "click" e "link". No conjunto de dados acima $P(\text{offer}|\text{click, link, SPAM}) = 0$. Uma alternativa para esse problema é coletar mais dados, de forma a diminuar as chances de probabilidades iguais a zero.

Uma outra alternativa é aproximar os valores de probabilidade. O CBI aproxima os valores ao assumir que as características são independentes dada a classe. Com isso temos que:

$$P(\text{SPAM})P(\text{offer, click, link}|\text{SPAM}) = P(\text{offer}|\text{SPAM})P(\text{click}|\text{SPAM})P(\text{link}|\text{SPAM})P(\text{SPAM})$$

O CBI é "ingênuo" pois os atributos muito provavelmente não são independentes dada a classe. Mas na prática, o algoritmo tende a funcionar muito bem, mesmo que os atributos não sejam independentes.

O algoritmo CBI pode então ser descrito com uma única equação.

$$\text{CBI}(X) = \arg\max_k P(k) \prod_{x \in X} P(x|k)$$

onde o $\arg\max$ considera todas as classes $k$ do problema.

## Underflow

Um problema comum durante a implementação do CBI é o *underflow*. Dado que o CBI realiza a multiplicação de vários números menores que 1, eventualmente, a precisão numérica do computador não será suficiente para armazenar números tão pequenos. Como resultado, as probabilidades de um objeto pertencer a cada uma das $k$ classes será igual a 0. Uma solução é transformar a multiplicação de números,

$$\text{CBI}(X) = \arg\max_k \left[ P(k) \prod_{x \in X} P(x|k) \right]$$

em um somatório de logs

$$\text{CBI}(X) = \arg\max_k \left[ \log P(k) \sum_{x \in X} \log P(x|k) \right]$$

## Exemplos

1. Considere a mensagem $m =$ *sports* e a base de dados exibida na tabela acima. CBI classificaria $m$ como SPAM ou como HAM?

$$P(\text{SPAM}|m) = \alpha(\frac{3}{8} \times \frac{1}{9}) = \alpha(\frac{1}{24})$$
$$P(\text{HAM}|m) = \alpha(\frac{5}{8} \times \frac{5}{15}) = \alpha(\frac{5}{24})\,,$$

   onde $\alpha = \frac{1}{P(\text{sports})}$. Como $\alpha(\frac{1}{24}) < \alpha(\frac{5}{24})$, CBI retorna que a mensagem é um HAM.

   Se quisermos calcular os valores exatos das probabilidades, precisamos calcular o valor de $\alpha = \frac{1}{P(\text{sports})}$. O valor de $P(\text{sports}) = P(\text{SPAM})P(\text{sports}|\text{SPAM}) + P(\text{HAM})P(\text{sports}|\text{HAM}) = \frac{3}{8} \times \frac{1}{9} + \frac{5}{8} \times \frac{5}{15} = 0.25$ Com isso:

$$P(\text{SPAM}|m) = \frac{1}{6}$$
$$P(\text{HAM}|m) = \frac{5}{6}\,.$$

2. Como o CBI classificaria a mensagem $m =$ *secret is secret*?

$$P(\text{SPAM}|m) = \alpha(\frac{3}{8} \times \frac{1}{3} \times \frac{1}{9} \times \frac{1}{3}) = \alpha(\frac{1}{216})$$
$$P(\text{HAM}|m) = \alpha(\frac{5}{8} \times \frac{1}{15} \times \frac{1}{15} \times \frac{1}{15}) = \alpha(\frac{1}{3375})\,,$$

   Classificaria como SPAM.

3. Como o CBI classificaria a mensagem $m =$ *today is secret*?

$$P(\text{SPAM}|m) = \alpha(\frac{3}{8} \times \frac{0}{9} \times \frac{1}{9} \times \frac{1}{9}) = 0$$

   Superadaptação!

## Suavização Aditiva

Qual a probabilidade do sol nascer amanhã dado que ele nasceu por $M$ dias? Assumir que é zero seria também uma superadaptação. Uma forma de diminuir a superadaptação é utilizar a suavização aditiva.

$$P(\neg \text{sol}) = \frac{1}{M+2}\,,$$

onde o +2 indica o número de classes (nasceu e não nasceu).

De uma forma geral, a fórmula de máxima verosimilhança é dada por

$$P(X_i) = \frac{\#X_i}{N}\,,$$

onde $\#X_i$ é o número de ocorrências de $X_i$ e $N$ é o número total de ocorrências.

Com suavização aditiva para um inteiro $k > 0$, a fórmula de máxima verosimilhança passa a ser:

$$P(X_i) = \frac{\#X_i + k}{N + k|X|}\,,$$

onde $|X|$ é o número de classes (e.g., nasceu e não nasceu).

**Exemplo 1:** Para $k = 1$:

- 1 mensagem e 1 SPAM, então $P(\text{SPAM}) = \frac{2}{3}$.
- 10 mensagens e 6 SPAMs, então $P(\text{SPAM}) = \frac{7}{12}$.
- 100 mensagens e 60 SPAMs, então $P(\text{SPAM}) = \frac{61}{102}$.

**Exemplo 2:** Para $k = 1$, um dicionário de 12 palavras e a base de treinamento de emails abaixo.

| SPAM | HAM |
|---|---|
| offer is secret | play sports today |
| click secret link | went sports today |
| secret sports link | secret sports event |
| | sports is today |
| | sports costs money |

- $P(\text{SPAM}) = \frac{3+1}{8+2} = \frac{2}{5}$, $P(\text{HAM}) = \frac{5+1}{8+2} = \frac{3}{5}$
- $P(\text{today}|SPAM) = \frac{1}{9+12} = \frac{1}{21}$, $P(\text{today}|HAM) = \frac{3+1}{15+12} = \frac{4}{27}$
- $P(\text{is}|SPAM) = \frac{2}{9+12} = \frac{2}{21}$, $P(\text{is}|HAM) = \frac{1+1}{15+12} = \frac{2}{27}$
- $P(\text{secret}|SPAM) = \frac{4}{9+12} = \frac{4}{21}$, $P(\text{secret}|HAM) = \frac{1+1}{15+12} = \frac{2}{27}$

Com isso,

$$P(\text{SPAM}|\text{today, is, secret}) = \alpha(\frac{2}{5} \times \frac{1}{21} \times \frac{2}{21} \times \frac{4}{21}) = \alpha(0.00034)$$
$$P(\text{HAM}|\text{today, is, secret}) = \alpha(\frac{3}{5} \times \frac{4}{27} \times \frac{2}{27} \times \frac{2}{27}) = \alpha(0.00048)$$