Métodos estatísticos em aprendizagem

Celso Vital Crivelaro

Thiago Francisco de Almeida

Pierre Marie Yves Lévêque

Aprendizagem estatísticas

A aprendizagem estatística, como o próprio nome diz, se baseia em dados previamente armazenados e suas estatísticas.

Em um exemplo, temos um pacote de balas de cereja e de lima, e essas balas são embaladas de uma forma que não é possível distinguir seu sabor antes de abrir.

O pacote de balas só pode assumir um desses valores:

- h1: 100% cereja
- h2: 75% cereja + 25% lima
- h3: 50% cereja + 50% lima
- h4: 25% cereja + 75% lima
- h5: 100% lima

Assim, o aprendizado de bayes tem o papel de achar a qual o tipo de saco de bala é o mais provável a partir das balas serem tiradas do saco.

A partir de um valor observado d e uma hipótese h têm por Bayes:

$$P(h_i \mid d) = a\ P(d \mid h_i)P(h_i)$$

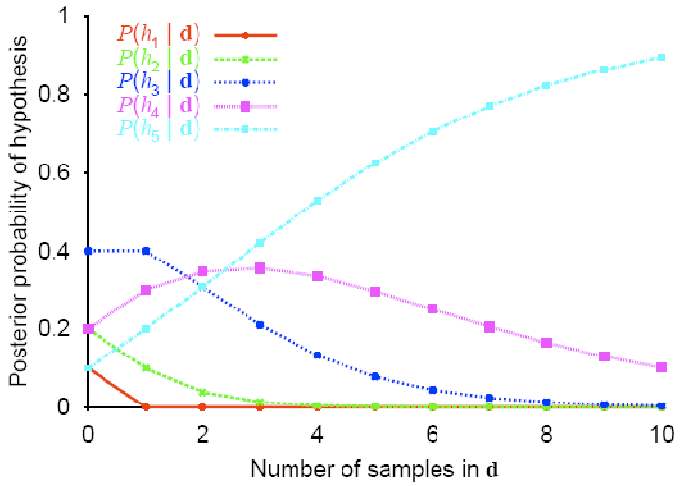
Para uma quantidade desconhecida X, temos:

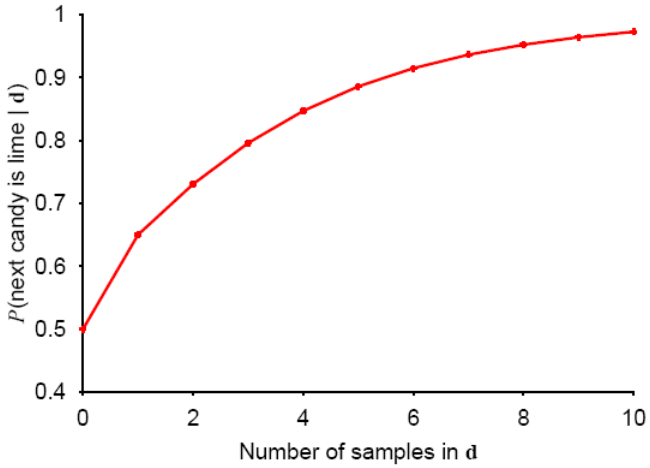
$$P(X \mid d) = \sum P(X \mid d,hi)\ P(hi \mid d) = \sum P(X \mid hi)\ P(hi \mid d)$$

Supomos agora que a probabilidade do saco ser h1,....,h5 é dado por {0.1,0.2,0.4,0.2,0.1}

Por exemplo, suponha que o saco de balas é realmente da configuração h5, e que as 10 primeiras balas fosse de lima, então a probabilidade $P(d \mid h3) = 0.5^{10}$, pois metade dos doces em h3 é lima.

Calculando dessa forma a chance do saco de balas assumirem cada possibilidade se dá no gráfico abaixo:





A previsão é ótima quer o conjunto de dados seja pequeno ou grande, porém é necessário aumentar a o espaço de hipóteses. E para problemas reais de aprendizagem o espaço de hipóteses é em geral muito grande ou infinito.

Assim é interessante fazer previsões com uma única hipótese, a mais provável.

Dessa forma surgiu o MAP (Maximum a posteriori), que parte do seguinte precedente:

$$P(X \mid d) \approx P(X \mid h_{map})$$

Após 3 doces de lima seguidos $h_{MAP}= h_5$, o 4º doce será previsto de lima com 100% de certeza, como podemos ver no gráfico.

Pegando o logaritmo de:

$P(h_i \mid \mathbf{d}) = cP(\mathbf{d} \mid h_i)P(h_i)$

Temos:

$- \log_2 P(d \mid h_i) - \log_2 P(h_i)$

Logo, para podermos maximizar $P(h_i \mid \mathbf{d})$ temos que minimizar $- \log_2 P(d \mid h_i) - \log_2 P(h_i)$

Aprendizagem com dados completos

Esse modelo de aprendizagem envolve a descoberta de parâmetros numéricos para um modelo de probabilidade cuja estrutura é fixa. Os dados são completos quando cada ponto de dados contém valores para toda a variável no modelo de probabilidade que está sendo aprendido.

Aprendizagem de parâmetros de máxima probabilidade: modelos discretos

Utilizando proporções desconhecidas

Supondo as proporções desconhecidas, sendo θ a fração entre 0 e 1 a proporção de cereja. A hipótese sobre essa proporção é h_θ.

Ao desembrulhar N doces, sendo 'c' doces de cereja e 'l' doces de lima , temos l = N –c e assim a probabilidade desse conjunto é :

$$P(d|h_\theta) = \prod_{j=1}^{N} P(d_j|h_\theta) = \theta^c \cdot (1-\theta)^l$$

Para termos a máxima probabilidade, temos que ter o máximo valor de θ. Para isso vamos diferenciar probabilidade logarítmica de $P(d|h_\theta)$:

$$L(d|h_\theta) - \log P(d|h_\theta) - \sum_{j=1}^{N} log P(d_j|h_\theta) - c\, log\theta + l \log(1-\theta)$$

Diferenciando e igualando a 0 para obter o máximo valor de θ:

$$\frac{dL(d \mid h_\theta)}{\quad} = \frac{c}{\quad} - \frac{\ell}{\quad} = 0 \Rightarrow \theta = \frac{c}{\quad} = \frac{c}{\quad}$$

Temos o resultado óbvio (o mesmo do senso comum): a quantidade de balas de cereja em relação ao todos que é tirado é a sua proporção!

Assim podemos resumir o método:

- Escrever uma expressão para a probabilidade dos dados com uma função do parâmetro
- Escrever a probabilidade logarítmica com relação a cada parâmetro
- Parâmetros para os quais a derivada é ZERO

Modelos de Bayes Ingênuos

O modelo de Bayes Ingênuos pressupõe que os atributos são condicionalmente independentes, assim podemos treiná-los em separado para obter a sua probabilidade.

Modelo de rede Bayesiana que a variável "C" é a raiz e as variáveis de atributos "Xi" são as folhas:

$$\theta = P(C = Verd)$$
$$\theta_{i1} = P(X_i = Verd \mid C = Verd)$$
$$\theta_{i2} = P(X_i = Verd \mid C = Falso)$$

Treinando essas variáveis, com os atributos observados (x1, ..., xn) a probabilidade de cada classe é dada por:

$$P(C \mid x1,...,xn) = P(C)\prod_i P(x_i \mid C)$$

Exemplo:

Temos os seguintes dias e a suas situações climática e tomou a decisão de jogar ou não tênis:

Dia	Tempo	Temperatura	Umidade	Vento	Jogar

D1	Sol	Quente	Alta	Fraco	Não
D2	Sol	Quente	Alta	Forte	Não
D3	Coberto	Quente	Alta	Fraco	Sim
D4	Chuva	Frio	Alta	Fraco	Sim
D5	Chuva	Frio	Normal	Fraco	Não
D6	Chuva	Frio	Normal	Forte	Não
D7	Coberto	Frio	Normal	Forte	Sim
D8	Sol	Normal	Alta	Fraco	Não
D9	Sol	Frio	Normal	Fraco	Sim
D10	Sol	Normal	Normal	Fraco	Sim
D11	**Chuva**	**Frio**	**Alta**	**Forte**	**??????**

Com as informações, jogaremos tênis no 11º dia?

Pegando as probabilidades a priori:

P(SIM) = 5/10 = 0.2 → De 10 dias com jogos, 5 foram jogados

P(NÃO) = 5/10 = 0.2 → De 10 dias com jogos, 5 não foram jogados

P(Sol/Sim) = 1/5 = 0.2 → De 5 dias jogados, 1 teve Sol

P(Sol/Não) = 3/5 = 0.6 → De 5 dias não-jogados, 3 tiveram Sol

P(Frio/Sim) = 2/5 = 0.4 → De 5 dias não-jogados, 2 fizeram Frio

P(Frio/Não) = 2/5 = 0.4 → De 5 dias não-jogados, 2 fizeram Frio

P(Alta/Sim) = 2/5 = 0.4 → De 5 dias não-jogados,2 tiveram umidade Alta

P(Alta/Não) = 3/5 = 0.6 → De 5 dias não-jogados, 3 tiveram umidade Alta

P(Forte/Sim) = 1/5 = 0.2 → De 5 dias não-jogados, 1 teve Vento Forte

P(Forte/Não) = 2/5 = 0.4 → De 5 dias não-jogados, 2 tiveram Vento Forte

Assim, supondo a decisão para Jogar:

P(Sim)P(Sol/Sim) P(Frio/Sim)P(Alta/Sim) P(Forte/Sim) = 0.0032

E para não jogar:

P(Não)P(Sol/Não)P(Frio/Não)P(Alta/Não) P(Forte/Não) = 0.0288

Como para não jogar o argumento é maior, decidiu para Não jogar no dia D11.

Aprendizagem de parâmetros de máxima probabilidade: modelos contínuos

Para modelos contínuos podemos usar o modelo de gaussiano linear:

$$P(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \mu = média \qquad \sigma = desvio$$

Para obter o máximo valor de x e observando os valores x1,...,xn , usaremos o mesmo método para modelos discretos:

Probabilidade Logarítmica:

$$L = N(-\log\sqrt{2\pi} - \log\sigma) - \sum_{j=1}^{N} \frac{(x_j - \mu)^2}{2\sigma^2}$$

Derivando e igualando a ZERO para obter a máxima probabilidade da média e a máxima probabilidade do desvio padrão:

$$\frac{\partial L}{\partial \sigma} = -\frac{1}{\sigma^2}\sum_{j=1}^{N}(x_j - \mu) = 0 \Rightarrow \mu = \frac{\sum_j x_j}{N}$$

$$\frac{\partial L}{\partial \mu} = -\frac{N}{\sigma^2} + \frac{1}{\sigma^3}\sum_{j=1}^{N}(x_j - \mu)^2 = 0 \Rightarrow \sigma = \sqrt{\frac{\sum_j (x_j - \mu)^2}{N}}$$

Confirmam a prática do senso comum --> A máxima probabilidade da média é a média das amostras e a máxima probabilidade do desvio-padrão é a raiz da variância das amostras.

Aprendizagem de parâmetros bayesiana

A aprendizagem de parâmetros impõe uma hipótese a priori sobre valores possíveis dos parâmetros e atualiza a distribuição a medida que os dados chegam.

θ (probabilidade de o doce ser de cereja) é o valor de uma variável aleatória de ϑ , a distribuição a priori $P(\vartheta = \theta)$ é a probabilidade que o saco tenha θ doces de cereja.

Se o parâmetro puder ter um valor entre 0 e 1, então P pode ser uma distribuição contínua diferente de zero e integral = 1 . Assim a densidade uniforme é uma candidata.

A densidade uniforme é um membro das distribuições beta que é definida pelos hiperparâmetros a e b:

$$beta[a,b](\theta) = \alpha\theta^{a-1}(1-\theta)^{b-1}$$

Propriedade:

- Se tem uma probabilidade a priori beta[a,b], então, depois da observação de um ponto de dados, a distribuição posterior para também é uma distribuição beta.
- A família beta é chamada de conjugado a priori

Observando um doce de cereja:

$$P(\theta \mid D_1 = cereja) = P(D_1 = cereja \mid \theta)P(\theta)$$
$$= \theta \cdot beta[a,b] = \theta \cdot \theta^{a-1} \cdot (1-\theta)^{b-1}$$
$$= \theta^a \cdot (1-\theta)^{b-1} = beta[a+1,b](\theta)$$

Ou seja, podemos ver a e b como contadores virtuais. Como se um a priori beta[a,b] se comporta como tivéssemos começado com a priori beta[1,1] e visto a – 1 cerejas e b – 1 limas

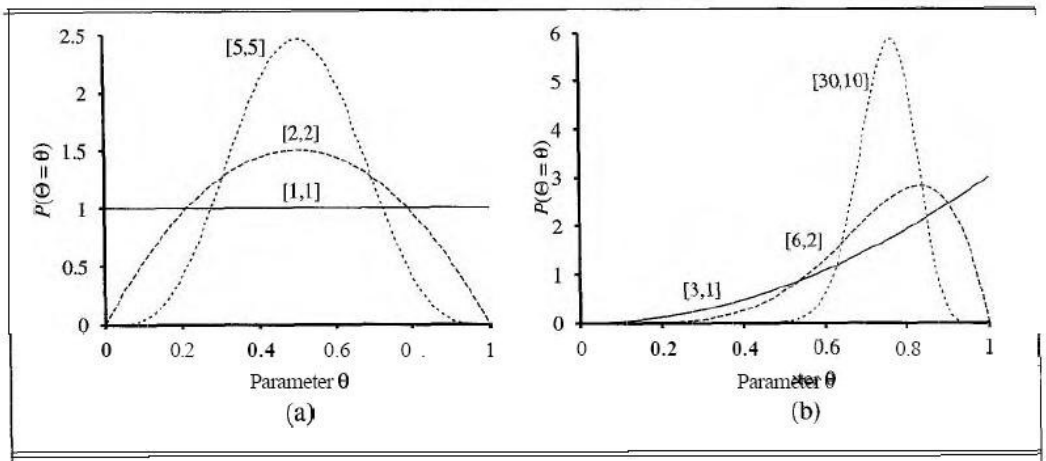


Figure 20.5 Examples of the beta[a, b] distribution for different values of [a,b].

75% de cerejas -> beta[3,1], beta[6,2], beta[30,10]

Aprendizagem de estruturas de redes bayesianas

Até agora, era conhecido a estrutura da rede, e queríamos saber as valores dos parâmetros. Agora, a estrutura da rede é desconhecida. Acontece quando tem um dúvida se tem uma relação entre dois parâmetros: por exemplo, é verdade que fumar não cria câncer ?

É óbvio que neste caso se precisa buscar a estrutura do arvore.

Tem dois principais jeitos de proceder:

1. Iniciar com um modelo que não contenha nenhum vínculo e começar a adicionar pais correspondentes a cada nó, depois ajustar os parâmetros, e medir a exatidão do modelo resultante.
2. Começar com um palpite inicial sobre a estrutura, utilizar busca por subida de encosta para fazer modificações, retornar os parâmetros após cada mudança de estrutura. As modificações possiveis são : inversão, adição ou eliminação de arcos.

Verificação da pertinência dos resultados

Precisa-se testar se as asserções de independência condicional implícitas na estrutura são realmente satisfeitas nos dados. A equação é simplesmente:

P(Sex/Sab, Bar | VaiEsperar) = P(Sex/Sab | VaiEsperar)P(Bar | VaiEsperar)

Riscos

O perigo são as flutuações estatísticas no conjunto de dados.

Elas significam que a equação nunca será satisfeita exatamente, então precisamos utilizar um teste estatístico apropriado para verificar se existe evidência estatística suficiente de que a hipótese de independência foi violada.

Quando mais rígido for este teste, mais vínculos serão adicionados: há um risque de superadaptação.