Universidade Federal do Ceará
Departamento de Computação

Aprendizagem de Máquina Probabilística

César Lincoln Cavalcante Mattos

2024

Agenda

1. Modelo beta-binomial
2. Modelo Dirichlet-multinomial
3. Classificador naive Bayes
4. Tópicos adicionais
5. Referências

Modelo beta-binomial

- Considere N lançamentos de uma moeda com probabilidade θ de ser cara $(X = 1)$ e $1 - \theta$ de ser coroa $(X = 0)$.
- A verossimilhança desse experimento pode ser escrita por:

$$x_i \sim \mathrm{Ber}(\theta),$$

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{N_1}(1-\theta)^{N-N_1},$$

$$N_1 = \sum_{i=1}^{N} \mathbb{I}(x_i = 1).$$

- Podemos escrever a probabilidade do número de caras:

$$p(N_1 = k) = \mathrm{Bin}(k|N, \theta) = \binom{N}{k}\theta^k(1-\theta)^{N-k}.$$

- Note que o termo $\binom{N}{k}$ não depende de θ.
- θ pode ser inferido com $\mathcal{D} = (N_1, N)$ ou $\mathcal{D} = (x_1, \dots, x_N)$.

Modelo beta-binomial

- Podemos escolher uma **priori conjugada** para $\theta \in [0, 1]$, com mesmo formato da verossimilhança, como a **distribuição beta**:

$$p(\theta) = \mathrm{Beta}(\theta|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1},$$

em que $a, b > 0$ são hiperparâmetros.

- Nesse caso a posteriori é analítica:

$$\begin{aligned}
p(\theta|\mathcal{D}) &\propto p(\mathcal{D}|\theta)p(\theta) \\
&\propto \theta^{N_1}(1-\theta)^{N-N_1}\theta^{a-1}(1-\theta)^{b-1} \\
&\propto \theta^{N_1+a-1}(1-\theta)^{N-N_1+b-1} \\
&= \mathrm{Beta}(\theta|N_1+a, N-N_1+b).
\end{aligned}$$

Modelo beta-binomial

- A soluções MAP, ML, média e variância da posteriori beta são:

$$\theta_{\text{MAP}} = \frac{a + N_1 - 1}{a + b + N - 2}, \quad \theta_{\text{ML}} = \frac{N_1}{N},$$
$$\mathbb{E}[\theta|\mathcal{D}] = \frac{a + N_1}{a + b + N},$$
$$\mathbb{V}[\theta|\mathcal{D}] = \frac{(a + N_1)(b + N - N_1)}{(a + b + N)^2(a + b + N + 1)}.$$

- A distribuição preditiva, i.e., probabilidade da próxima observação ser cara $(x_* = 1)$, também é analítica:

$$p(x_* = 1|\mathcal{D}) = \int_0^1 p(x_* = 1|\theta)p(\theta|\mathcal{D})\mathrm{d}\theta$$
$$= \int_0^1 \theta \mathrm{Beta}(\theta|N_1 + a, N - N_1 + b)\mathrm{d}\theta = \mathbb{E}[\theta|\mathcal{D}].$$

- Alternativamente, temos $p(x_*|\mathcal{D}) = \mathrm{Ber}(x_*|\mathbb{E}[\theta|\mathcal{D}])$.

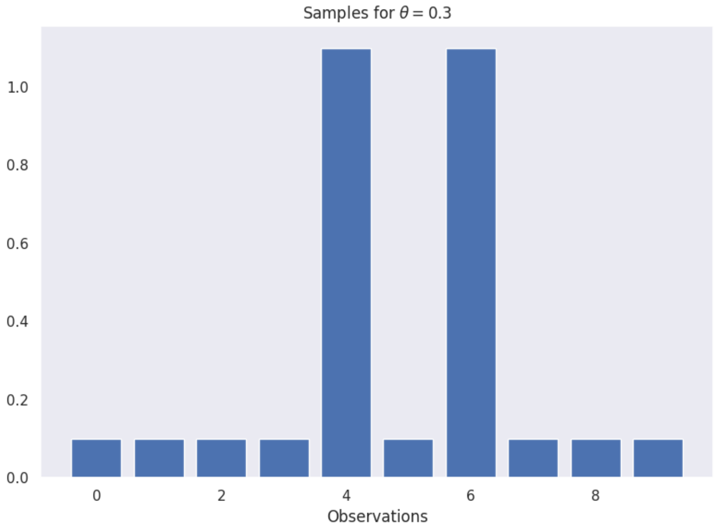
Modelo beta-binomial

- A distribuição preditiva do número de caras obtidas nas próximas M observações ($\hat{a} = N_1 + a$ e $\hat{b} = N - N_1 + b$) é dada por:

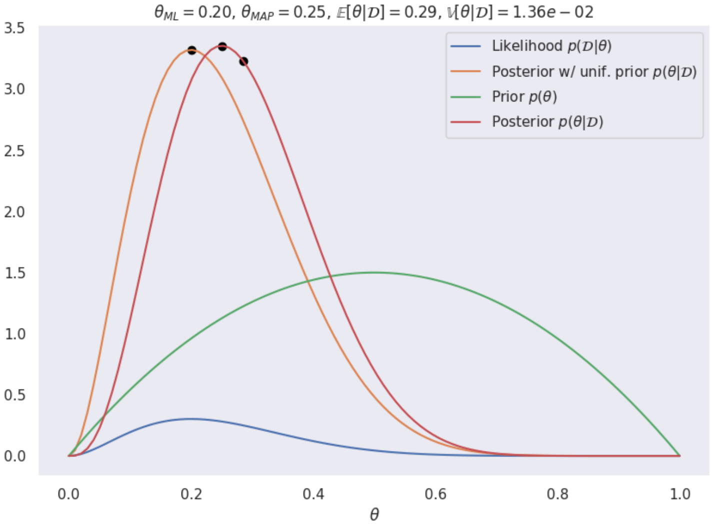
$$
\begin{aligned}
p(k|\mathcal{D}, M) &= \int_0^1 \mathrm{Bin}(k|M,\theta)\mathrm{Beta}(\theta|\hat{a},\hat{b})\mathrm{d}\theta \\
&= \binom{M}{k} \frac{\Gamma(\hat{a}+\hat{b})}{\Gamma(\hat{a})\Gamma(\hat{b})} \int_0^1 \theta^k (1-\theta)^{M-k} \theta^{\hat{a}-1} (1-\theta)^{\hat{b}-1} \mathrm{d}\theta \\
&= \binom{M}{k} \frac{\Gamma(\hat{a}+\hat{b})}{\Gamma(\hat{a})\Gamma(\hat{b})} \int_0^1 \theta^{k+\hat{a}-1} (1-\theta)^{M-k+\hat{b}-1} \mathrm{d}\theta \\
&= \binom{M}{k} \frac{\Gamma(\hat{a}+\hat{b})}{\Gamma(\hat{a})\Gamma(\hat{b})} \frac{\Gamma(k+\hat{a})\Gamma(M-k+\hat{b})\Gamma(M+1)}{\Gamma(k+1)\Gamma(M-k+1)\Gamma(M+\hat{a}+\hat{b})} \\
&= \mathrm{Bb}(k|\hat{a},\hat{b},M) \quad \text{(distribuição beta-binomial).}
\end{aligned}
$$

- Temos ainda que $\mathbb{E}[k|\mathcal{D}] = M\frac{\hat{a}}{\hat{a}+\hat{b}}$ e $\mathbb{V}[k|\mathcal{D}] = \frac{M\hat{a}\hat{b}}{(\hat{a}+\hat{b})^2}\frac{\hat{a}+\hat{b}+M}{\hat{a}+\hat{b}+1}$.

Modelo beta-binomial

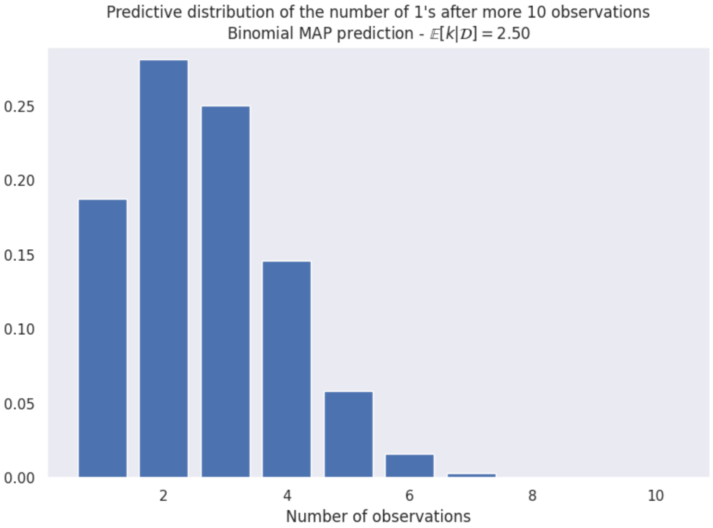


Modelo beta-binomial

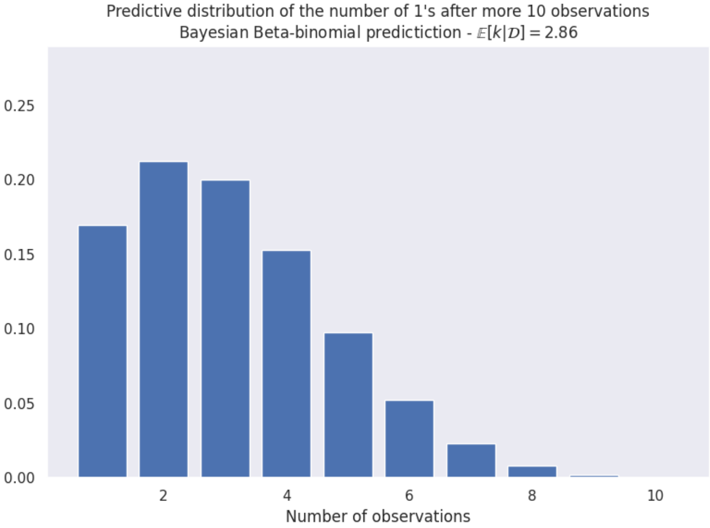


$$p(\theta) = \mathrm{Beta}(\theta|a = 2, b = 2),$$
$$p(N_1 = k) = \mathrm{Bin}(k|N, \theta), \quad k = 2, N = 10,$$
$$p(\theta|\mathcal{D}) = \mathrm{Beta}(\theta|a = 4, b = 10).$$

Modelo beta-binomial - solução MAP



Modelo beta-binomial - solução Bayesiana



Agenda

Modelo Dirichlet-multinomial

- Considere N lançamentos de um dado com K faces.
- A k-ésima face tem probabilidade θ_k de ser observada.
- A verossimilhança desse experimento pode ser escrita por:

$$p(x_1, x_2, \dots, x_N|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{N_k},$$

$$N_k = \sum_{i=1}^{N} \mathbb{I}(x_i = k), \quad \sum_{k=1}^{K} \theta_k = 1.$$

- Alternativamente poderíamos usar a **distribuição multinomial**:

$$\begin{aligned} p(N_1, N_2, \dots, N_k|\boldsymbol{\theta}) &= \binom{N}{N_1, \dots, N_k} \prod_{k=1}^{K} \theta_k^{N_k} \\ &= \frac{N!}{N_1!N_2!\cdots N_k!} \prod_{k=1}^{K} \theta_k^{N_k}. \end{aligned}$$

Modelo Dirichlet-multinomial

- Escolhemos uma **priori conjugada** para $\boldsymbol{\theta} \mid \theta_k \in [0, 1]$, a **distribuição de Dirichlet**:

$$p(\boldsymbol{\theta}) = \mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_1 + \ldots + \alpha_K)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1},$$

em que $\alpha_1, \ldots, \alpha_k > 0$ são hiperparâmetros.

- Nesse caso a posteriori é analítica:

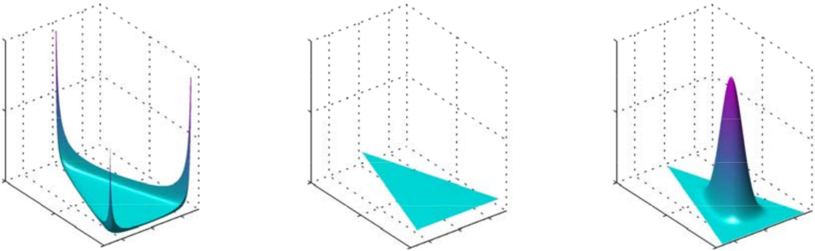
$$p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \propto \prod_{k=1}^{K} \theta_k^{N_k} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} = \prod_{k=1}^{K} \theta_k^{N_k + \alpha_k - 1},$$

$$p(\boldsymbol{\theta}|\mathcal{D}) = \mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha} + \boldsymbol{N}), \quad \boldsymbol{N} = [N_1, \ldots, N_K]^\top.$$

- A média e a variância são dadas por:

$$\mathbb{E}[\theta_k|\mathcal{D}] = \frac{\alpha_k + N_k}{N + \sum_{k=1}^{K} \alpha_k}, \quad \mathbb{V}[\theta_k|\mathcal{D}] = \frac{\frac{\alpha_k}{\sum_{k=1}^{K} \alpha_k}\left(1 - \frac{\alpha_k}{\sum_{k=1}^{K} \alpha_k}\right)}{1 + \sum_{k=1}^{K} \alpha_k}.$$

Modelo Dirichlet-multinomial



Distribuição de Dirichlet sobre 3 variáveis. Da esquerda para direita: $\alpha_k = 0.1$, $\alpha_k = 1$ e $\alpha_k = 10$

Modelo Dirichlet-multinomial

- A solução MAP $\boldsymbol{\theta}_{\text{MAP}}$ deve ser obtida maximizando $\log p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$.
- Usamos um **Lagrangiano** com a restrição de $\sum_{k=1}^{K} \theta_k = 1$:

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}, \lambda) &= \log \prod_{k=1}^{K} \theta_k^{N_k + \alpha_k - 1} + \lambda \left(1 - \sum_{k=1}^{K} \theta_k\right) \\
&= \sum_{k=1}^{K} (N_k + \alpha_k - 1) \log \theta_k + \lambda \left(1 - \sum_{k=1}^{K} \theta_k\right),
\end{aligned}
$$

em que $\lambda \in \mathbb{R}$ é um **multiplicador de Lagrange**.

Modelo Dirichlet-multinomial

- Calculamos $\frac{\partial \mathcal{L}(\boldsymbol{\theta},\lambda)}{\partial \theta_k} = 0$:

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta},\lambda)}{\partial \theta_k} = \frac{N_k + \alpha_k - 1}{\theta_k} - \lambda = 0.$$

$$\lambda \theta_k = N_k + \alpha_k - 1.$$

- Como $\sum_{k=1}^{K} \theta_k = 1$, temos:

$$\lambda = \sum_{k=1}^{K} (N_k + \alpha_k - 1) = N + \alpha_0 - K, \qquad \text{em que } \alpha_0 = \sum_{k=1}^{K} \alpha_k.$$

- O máximo do Lagrangiano então será:

$$[\theta_{\mathsf{MAP}}]_k = \frac{N_k + \alpha_k - 1}{N + \alpha_0 - K}.$$

- Para $\alpha_k = 1$ (priori uniforme), obtemos a solução de máxima verossimilhança:

$$[\theta_{\mathsf{ML}}]_k = N_k / N.$$

Modelo Dirichlet-multinomial

- A distribuição preditiva, i.e., a probabilidade da próxima observação ser $x_* = k$, também é analítica:

$$
\begin{aligned}
p(x_* = k|\mathcal{D}) &= \int p(x_* = k|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})\mathrm{d}\boldsymbol{\theta} \\
&= \int p(x_* = k|\theta_k) \left[\int p(\boldsymbol{\theta}_{-k}, \theta_k|\mathcal{D})\mathrm{d}\boldsymbol{\theta}_{-k}\right] \mathrm{d}\theta_k \\
&= \int \theta_k p(\theta_k|\mathcal{D})\mathrm{d}\theta_k \\
&= \mathbb{E}[\theta_k|\mathcal{D}] = \frac{\alpha_k + N_k}{N + \sum_{k=1}^{K} \alpha_k}.
\end{aligned}
$$

Agenda

1. Modelo beta-binomial
2. Modelo Dirichlet-multinomial
3. Classificador naive Bayes
4. Tópicos adicionais
5. Referências

Classificador naive Bayes

- Em uma tarefa de classificação com C classes, podemos considerar uma distribuição $p(\boldsymbol{x}|y = c)$ para os padrões da classe c.
- O **classificador naive Bayes** parte da suposição de independência dos atributos do padrão $\boldsymbol{x}$ condicionados à classe:

$$p(\boldsymbol{x}|y = c, \boldsymbol{\theta}) = \prod_{d=1}^{D} p(x_d|y = c, \boldsymbol{\theta}_{dc}),$$

em que $\boldsymbol{\theta}$ coleciona os parâmetros do modelo e $\boldsymbol{\theta}_{dc}$ indica os parâmetros referentes ao d-ésimo atributo da classe c.

Classificador naive Bayes

- Temos as seguintes opções para $p(x_d|y = c, \boldsymbol{\theta}_{dc})$:
  - $\rightarrow$ No caso de atributos reais, podemos usar uma distribuição Gaussiana $p(x_d|y = c, \boldsymbol{\theta}_{dc}) = \mathcal{N}(x_d|\mu_{dc}, \sigma^2_{dc})$;
  - $\rightarrow$ No caso de atributos binários, podemos usar uma distribuição de Bernoulli $p(x_d|y = c, \theta_{dc}) = \mathrm{Ber}(x_d|\theta_{dc})$.
  - $\rightarrow$ No caso de atributos categóricos no formato 1-of-K, podemos usar uma distribuição multinoulli/categórica $p(\boldsymbol{x}_d|y = c, \boldsymbol{\theta}_{dc}) = \mathrm{Cat}(\boldsymbol{x}_d|\boldsymbol{\theta}_{dc}) = \prod_{k=1}^{K} \theta_{dck}^{x_{dk}}$.
- Note que podemos escolher outras distribuições e/ou combinar diferentes distribuições para diferentes atributos.

Classificador naive Bayes

- Sendo $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_c]^\top$ o vetor de probabilidades a priori das classes, podemos escrever a verossimilhança do modelo:

$$p(\boldsymbol{X}, \boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(y_i|\boldsymbol{\pi}) \prod_{d=1}^{D} p(x_{id}|y_i, \boldsymbol{\theta}_d)$$

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{c=1}^{C} \pi_c^{\mathbb{I}(y_i=c)} \prod_{d=1}^{D} \prod_{c=1}^{C} p(x_{id}|\boldsymbol{\theta}_{dc})^{\mathbb{I}(y_i=c)}$$

$$\log p(\mathcal{D}|\boldsymbol{\theta}) = \sum_{c=1}^{C} N_c \log \pi_c + \sum_{i|y_i=c} \sum_{d=1}^{D} \sum_{c=1}^{C} \log p(x_{id}|\boldsymbol{\theta}_{dc}),$$

em que N_c é o número de exemplos da classe c.

Classificador naive Bayes

- A solução ML para π_c é dada por $\hat{\pi}_c = \frac{N_c}{N}$.
- A solução ML para os demais parâmetros depende da distribuição condicional escolhida para os atributos:
  - $\rightarrow$ No caso de Gaussianas $p(x_d|y = c, \boldsymbol{\theta}_{dc}) = \mathcal{N}(x_d|\mu_{dc}, \sigma^2_{dc})$:

$$\hat{\mu}_{dc} = \frac{1}{N_c} \sum_{i|y_i=c} x_{id},$$

$$\hat{\sigma}^2_{dc} = \frac{1}{N_c - 1} \sum_{i|y_i=c} (x_{id} - \hat{\mu}_{dc})^2.$$

  - $\rightarrow$ No caso Bernoulli $p(x_d|y = c, \theta_{dc}) = \mathrm{Ber}(x_d|\theta_{dc})$:

$$\hat{\theta}_{dc} = \frac{\sum_{i|y_i=c} \mathbb{I}(x_{id} = 1)}{N_c} = \frac{N_{dc}}{N_c}.$$

  - $\rightarrow$ No caso categórico $p(\boldsymbol{x}_d|y = c, \theta_{dc}) = \mathrm{Cat}(\boldsymbol{x}_d|\boldsymbol{\theta}_{dc})$:

$$\hat{\theta}_{dck} = \frac{\sum_{i|y_i=c} \mathbb{I}(x_{idk} = 1)}{N_c} = \frac{N_{dck}}{N_c}.$$

Classificador naive Bayes - inferência Bayesiana

- Alternativamente, podemos seguir uma abordagem Bayesiana e escolher uma priori fatorada para os parâmetros $p(\boldsymbol{\theta})$:

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}) \prod_{d=1}^{D} \prod_{c=1}^{C} p(\theta_{dc}).$$

- Para atributos binários, i.e., verossimilhanças de Bernoulli, escolhemos as priori conjugadas abaixo:

$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}), \quad p(\theta_{dc}) = \mathrm{Beta}(\theta_{dc}|a, b).$$

- As posteriori seguem os modelos beta-binomial e Dirichlet-multinomial:

$$p(\boldsymbol{\theta}|\mathcal{D}) = p(\boldsymbol{\pi}|\mathcal{D}) \prod_{d=1}^{D} \prod_{c=1}^{C} p(\theta_{dc}|\mathcal{D}),$$

$$p(\boldsymbol{\pi}|\mathcal{D}) = \mathrm{Dir}(N_1 + \alpha_1, \ldots, N_C + \alpha_C),$$

$$p(\theta_{dc}|\mathcal{D}) = \mathrm{Beta}(N_{dc} + a, N_c - N_{dc} + b).$$

Classificador naive Bayes - predições

- Dado um novo padrão $\boldsymbol{x}_*$, devemos computar a distribuição preditiva para cada classe (omitindo as dependências a $\boldsymbol{\theta}$):

$$p(y_* = c|\boldsymbol{x}_*, \mathcal{D}) \propto p(y_* = c|\mathcal{D}) \prod_{d=1}^{D} p(x_{*d}|y_* = c, \mathcal{D}).$$

- A abordagem Bayesiana busca marginalizar $\boldsymbol{\theta}$ usando a distribuição a posteriori $p(\boldsymbol{\theta}_{dc}|\mathcal{D})$. Considerando atributos binários:

$$p(y_* = c|\boldsymbol{x}_*, \mathcal{D}) \propto \left[\int \mathrm{Cat}(y_* = c|\boldsymbol{\pi})p(\boldsymbol{\pi}|\mathcal{D})\mathrm{d}\boldsymbol{\pi}\right]$$
$$\prod_{d=1}^{D} \left[\int \mathrm{Ber}(x_{*d}|y_* = c, \theta_{dc})p(\theta_{dc}|\mathcal{D})\right]$$

Classificador naive Bayes - predições

- A partir do modelos conjugados anteriores, denotando $\bar{\pi}_c = \mathbb{E}[\pi_c|\mathcal{D}]$ e $\bar{\theta}_{dc} = \mathbb{E}[\theta_{dc}|\mathcal{D}]$, a preditiva também é analítica:

$$p(y_* = c|\boldsymbol{x}_*, \mathcal{D}) \propto \bar{\pi}_c \prod_{d=1}^{D} (\bar{\theta}_{dc})^{\mathbb{I}(x_{*d}=1)} (1 - \bar{\theta}_{dc})^{\mathbb{I}(x_{*d}=0)},$$

$$\bar{\theta}_{dc} = \frac{N_{dc} + b}{N_c + a + b},$$

$$\bar{\pi}_c = \frac{N_c + \alpha_c}{N + \sum_{c=1}^{C} \alpha_c}.$$

- A classe predita $\hat{y}_*$ será dada por:

$$\hat{y}_* = \arg\max_c \left[\bar{\pi}_c \prod_{d=1}^{D} (\bar{\theta}_{dc})^{\mathbb{I}(x_{*d}=1)} (1 - \bar{\theta}_{dc})^{\mathbb{I}(x_{*d}=0)} \right].$$

Classificador naive Bayes - predições

- Por conveniência numérica, aplicamos uma função logarítmica na expressão à direita:

$$
\begin{aligned}
\hat{y}_* &= \arg\max_c \log\left[\bar{\pi}_c \prod_{d=1}^{D} (\bar{\theta}_{dc})^{\mathbb{I}(x_{*d}=1)} (1-\bar{\theta}_{dc})^{\mathbb{I}(x_{*d}=0)}\right] \\
&= \arg\max_c \left[\log \bar{\pi}_c + \sum_{d|x_{*d}=1} \log \bar{\theta}_{dc} + \sum_{d|x_{*d}=0} \log(1-\bar{\theta}_{dc})\right].
\end{aligned}
$$

Agenda

Tópicos adicionais

- Família exponencial.
  - → Distribuições que podem ser escritas no formato abaixo:

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = h(\boldsymbol{x})g(\boldsymbol{\eta})\exp(\boldsymbol{\eta}^\top \boldsymbol{u}(\boldsymbol{x})),$$

    em que $\boldsymbol{\eta}$ são os chamados **parâmetros naturais** da distribuição, $h(\cdot)$, $g(\cdot)$ e $\boldsymbol{u}(\cdot)$ são funções do vetor $\boldsymbol{x}$.
  - → O termo $g(\boldsymbol{\eta})$ normaliza a distribuição e é dado por:

$$g(\boldsymbol{\eta}) \int h(\boldsymbol{x})\exp(\boldsymbol{\eta}^\top \boldsymbol{u}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} = 1,$$

    em que a integral é substituída por um somatório caso $\boldsymbol{x}$ seja discreto.
  - → Além de várias propriedades importantes, a família exponencial é a única família de distribuições que admite priori conjugada.
- Modelos não-conjugados.

Agenda

Referências bibliográficas

- **Cap. 3** - MURPHY, Kevin P. **Machine learning: a probabilistic perspective**, 2012.
- **Cap. 2** - BISHOP, Christopher M. **Pattern recognition and machine learning**, 2006.