

Байесовский выбор моделей: экспоненциальное семейство распределений

Александр Адуенко

19е сентября 2018

- Формула Байеса: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$;
- Формула полной вероятности: $P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A})$;
- Определение априорных вероятностей и selection bias;
- Тестирование гипотез
 - Ошибка первого рода и мощность критерия;
 - Критическая область и как ее определить;
- Проблема множественного тестирования гипотез
 - Проблема ложных открытий при независимом одновременном тестировании множества гипотез;
 - FWER и FDR как обобщения вероятности ошибки первого рода;
 - Поправка Бонферрони как консервативное средство контроля FWER;
 - Поправка Бенджамини-Хохберга для контроля FDR для положительно регрессионно зависимых гипотез.

Экспоненциальное семейство распределений

Распределение $p(\mathbf{x})$ в экспоненциальном семействе, если плотность вероятности (функция вероятности) представима в виде

$$p(\mathbf{x}|\Theta) = \frac{1}{Z(\Theta)} h(\mathbf{x}) \exp(\Theta^\top \mathbf{u}(\mathbf{x})).$$

Распределение	Плотность	$\mathbf{u}(\mathbf{x})$	Θ	$Z(\Theta)$
Be(p)	$p^x (1-p)^{1-x}$	x	$\log \frac{p}{1-p}$	$\frac{1}{1-p}$
Poisson(λ)	$\frac{\lambda^x}{x!} e^{-\lambda}$	x	$\log \lambda$	e^λ
$\Gamma(\alpha, \beta)$	$\frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}$	$[\log x, x]$	$[\alpha, -\beta]$	$\frac{\Gamma(\alpha)}{\beta^\alpha}$
$B(\alpha, \beta)$	$\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$	$[\log x, \log(1-x)]$	$[\alpha, \beta]$	$\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$
Dir(α)	$\frac{\Gamma(\sum \alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_i p_i^{\alpha_i - 1}$	$[\log p_i]$	α	$\frac{\prod_j \Gamma(\alpha_j)}{\Gamma(\sum \alpha_j)}$
$N(\mathbf{m}, \Sigma^{-1})$	$\frac{\sqrt{\det \Sigma}}{(2\pi)^{n/2}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^\top \Sigma (\mathbf{x}-\mathbf{m})}$	$[\mathbf{x}, \mathbf{x}\mathbf{x}^\top]$	$[\Sigma \mathbf{m}, -\frac{1}{2}\Sigma]$	$\frac{(2\pi)^{n/2} e^{-\frac{1}{2}\mathbf{m}^\top \Sigma \mathbf{m}}}{\sqrt{\det \Sigma}}$

Пример:
$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-m)^2} = \underbrace{\frac{1}{\sqrt{2\pi\sigma e \frac{m^2}{2\sigma^2}}}_{Z(\Theta)}} e^{\underbrace{\frac{u_1(x)}{x}}_{\theta_1} \cdot \underbrace{\frac{m}{\sigma^2}}_{\theta_2} + \underbrace{\frac{u_2(x)}{x^2}}_{\theta_2} \cdot \underbrace{\frac{-1}{2\sigma^2}}_{\theta_2}},$$

$$Z(\Theta) = \sqrt{-\pi/\theta_2} e^{-\frac{\theta_1^2}{4\theta_2}}.$$

Экспоненциальное семейство распределений.

Достаточные статистики.

Статистика $T(\mathbf{x})$ называется **достаточной** относительно параметра Θ , если $p(\mathbf{x}|T(\mathbf{x}) = t, \Theta) = p(\mathbf{x}|T(\mathbf{x}) = t)$.

Теорема Фишера-Неймана о факторизации. $T(\mathbf{x})$ достаточна относительно параметра $\Theta \iff p(\mathbf{x}|\Theta) = h(\mathbf{x})g(\Theta, T(\mathbf{x}))$.

Экспоненциальное семейство: $p(\mathbf{x}|\Theta) = \frac{1}{Z(\Theta)}h(\mathbf{x})\exp(\Theta^\top \mathbf{u}(\mathbf{x}))$.

Свойство: $E\mathbf{u}(\mathbf{x}) = \nabla \log Z(\Theta)$, $E\dot{\mathbf{u}}\dot{\mathbf{u}}^\top = \nabla\nabla \log Z(\Theta)$.

Пример (нормальное распределение): $Z(\Theta) = \sqrt{-\pi/\theta_2}e^{-\frac{\theta_1^2}{4\theta_2}}$.

$Eu_1(x) = Ex = -\frac{\theta_1}{2\theta_2} = m$, $Ex^2 = \frac{\theta_1^2}{4\theta_2^2} - \frac{1}{2\theta_2} = m^2 + \sigma^2$;

$E\dot{u}_1^2 = Dx^2 = \frac{1}{2\theta_2^2} - \frac{\theta_1^2}{2\theta_2^3} = 2\sigma^4 + 4m^2\sigma^2$.

Пример (гамма-распределение): $p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$.

$\log Z(\Theta) = \log \frac{\Gamma(\alpha)}{\beta^\alpha} = \log \Gamma(\alpha) - \alpha \log \beta$;

$E \log x = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - \log \beta = \psi(\alpha) - \log \beta$; $Ex = -\frac{\partial}{\partial \beta} \log Z(\Theta) = \frac{\alpha}{\beta}$.

Наивный байесовский классификатор

Пусть имеется K классов $C = \{C_1, \dots, C_K\}$ и $\mathbf{x} \in \mathbb{R}^n$.

Требуется построить классификатор $f(\cdot) : \mathbb{R}^n \rightarrow C$.

$$p(C_k|\mathbf{x}) = \frac{p(C_k)p(\mathbf{x}|C_k)}{p(\mathbf{x})} \propto p(C_k)p(\mathbf{x}|C_k).$$

$$p(C_k)p(\mathbf{x}|C_k) = p(C_k)p(x_1|C_k)p(x_2|x_1, C_k) \cdot \dots \cdot p(x_n|x_1, \dots, x_{n-1}, C_k).$$

«**Наивность**»: $p(x_i|x_1, \dots, x_{i-1}, C_k) = p(x_i|C_k)$.

$$p(C_k|\mathbf{x}) = \frac{p(C_k) \prod_{i=1}^n p(x_i|C_k)}{p(\mathbf{x})}.$$

Классификатор: $f(\mathbf{x}) = \arg \max_k \left(p(C_k) \prod_{i=1}^n p(x_i|C_k) \right)$.

Вопросы:

- Как определить $p(C_k)$ и $p(x_i|C_k)$?
- Насколько плоха «наивность», и зачем она вводится?
- Почему классификатор такого вида?

Вопрос: как определить $p(C_k)$ и $p(x_i|C_k)$?

- 1 Определяем $p(C_k)$ частотно по выборке, а для $p(x_i|C_k)$ строим параметрическую модель и используем ML-оценки ее параметров по выборке;
- 2 Аналогично п.1, но используем непараметрическое оценивание плотностей;
- 3 Вводим априорное распределение на вектор вероятностей $[p(C_1), \dots, p(C_K)]^T$, параметрическую модель на $p(x_i|C_k)$ с неизвестными параметрами, и априорное распределение на параметры моделей.

Вопрос: насколько плоха «наивность», и зачем она вводится?

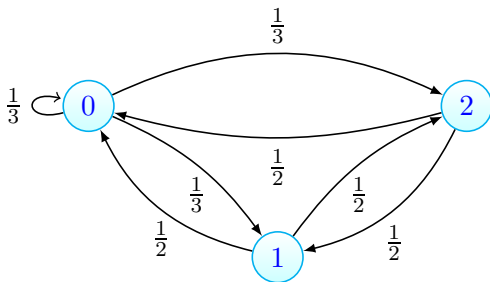
Пример: $K = 2$,

$$p(\mathbf{x}|C_1) = N\left(\mathbf{0}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right), \quad p(\mathbf{x}|C_2) = N\left(\mathbf{0}, \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\right).$$

Наивный байесовский классификатор: продолжение

Пример. Классификация пользователей по интересующему атрибуту (например, полу, возрасту, достатку, интересу к некоторому товару) по истории \mathbf{x} переходов между веб-страницами.

Предположение: переходы между страницами для каждого класса C_k описываются марковской цепью с некоторыми вероятностями перехода (разными для разных классов) между состояниями (веб-страницами).



$$p(C_k|\mathbf{x}) = \frac{p(C_k)p(\mathbf{x}|C_k)}{p(\mathbf{x})} \propto p(C_k)p(\mathbf{x}|C_k).$$

$$p(C_k)p(\mathbf{x}|C_k) = p(C_k)p(x_1|C_k)p(x_2|x_1, C_k) \cdot \dots \cdot p(x_n|x_1, \dots, x_{n-1}, C_k) = p(C_k)p(x_1|C_k)p(x_2|x_1, C_k) \cdot \dots \cdot p(x_n|x_{n-1}, C_k).$$

Вопрос: как оценить $p(x_1|C_k)$, $p(C_k)$ и $p(x_i|x_{i-1}, C_k)$?

Классификатор:

$$f(\mathbf{x}) = \arg \max_k p(C_k | \mathbf{x}) = \arg \max_k \left(p(C_k) \prod_{i=1}^n p(x_i | C_k) \right).$$

Вопрос. Пусть $p(C_k | \mathbf{x})$ известна точно. Какой классификатор оптимален?

Пусть $K = 2$ и $P = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}$ есть матрица штрафа.

Пример 1. $p_{11} = p_{22} = 0$, $p_{12} = 0$, $p_{21} = 1$;

Пример 2. $p_{11} = p_{22} = 0$, $p_{12} = 1$, $p_{21} = 1$;

Пример 3. $p_{11} = p_{22} = 0$, $p_{12} = 1$, $p_{21} = 10$;

Пример 4. $p_{11} = -1$, $p_{22} = -100$, $p_{12} = 1$, $p_{21} = 1$.

- 1 Bishop, Christopher M. "Pattern recognition and machine learning". Springer, New York (2006).
- 2 MacKay, David JC. Bayesian methods for adaptive models. Diss. California Institute of Technology, 1992.
- 3 MacKay, David JC. "The evidence framework applied to classification networks." *Neural computation* 4.5 (1992): 720-736.
- 4 Gelman, Andrew, et al. Bayesian data analysis, 3rd edition. Chapman and Hall/CRC, 2013.
- 5 Agresti, Alan. Analysis of ordinal categorical data. Vol. 656. John Wiley & Sons, 2010.
- 6 Дрейпер, Норман Р. Прикладной регрессионный анализ. Рипол Классик, 2007.
- 7 Кобзарь, Александр Иванович. Прикладная математическая статистика. Физматлит, 2006.