

Байесовский выбор моделей: введение

Александр Адуенко

9е сентября 2020

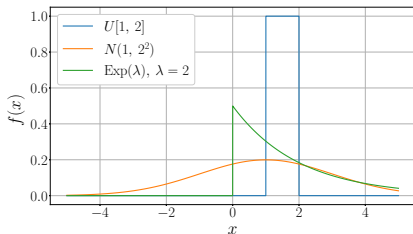
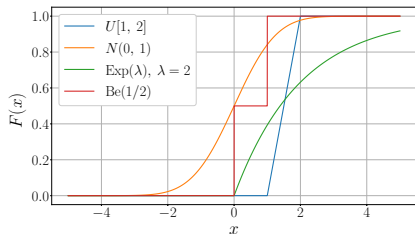
- Формула Байеса: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$;
- Формула полной вероятности: $P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A})$;
- Определение априорных вероятностей и selection bias;
- Тестирование гипотез
 - Ошибка первого рода и мощность критерия;
 - Критическая область и как ее определить;
- Проблема 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;
- Зависимость формы классификатора от функции полезности.

Случайные величины и их характеристики

Случайная величина – измеримая функция, заданная на некотором вероятностном пространстве.

Функция распределения: $F_{\xi}(\mathbf{x}) = P(\xi < \mathbf{x})$, $\mathbf{x} \in \mathbb{R}^n$.

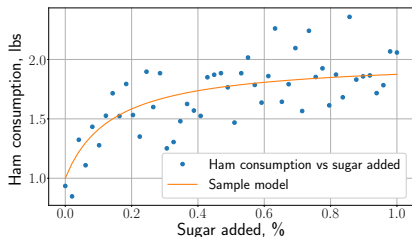
Функция плотности распределения: $f_{\xi}(\mathbf{x}) = \frac{\partial F_{\xi}(\mathbf{x})}{\partial \mathbf{x}}$.



10%-ная квантиль (для одномерной с.в.) – $x : F(x) = 0.1$.

Важные характеристики: мат. ожидание, дисперсия, стандартное отклонение, медиана, мода, коэффициент асимметрии (skewness), коэффициент эксцесса (kurtosis).

Статистика – измеримая функция выборки (тоже случайная величина). Пусть требуется проверить утверждение: «чем больше сахара добавлено в продукт, тем больше его душевое потребление».



Пусть даны НОР пары $\mathbf{z}_i = (x_i, y_i)$, $i = \overline{1, n}$, показывающие для ветчины, сколько сахара добавлено, и сколько её продано на одного человека.

Гипотеза H_0 : монотонной зависимости нет.

Требуется: построить статистику $T(\mathbf{Z})$ и на уровне значимости $\alpha = 0.05$ проверить гипотезу.

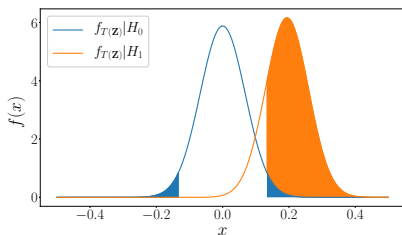
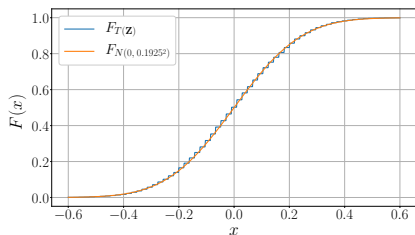
Идеальная положительная монотонная зависимость:

$$x_{i_1} > x_{i_2} \implies y_{i_1} > y_{i_2}.$$

Идея: введем $\xi_i = F_x(x_i)$, $\eta_i = F_y(y_i)$, $\xi_i, \eta_i \sim U[0, 1]$. Скажем, что монотонной зависимости нет, если $F_{\xi\eta}(a, b) = F_\xi(a)F_\eta(b)$.

$$T(\mathbf{Z}) = \frac{2}{n(n-1)} \sum_{i < j} \text{sign}(x_i - x_j) \text{sign}(y_i - y_j).$$

$$ET(\mathbf{Z})|H_0 = 0, \text{DT}(\mathbf{Z})|H_0 = \frac{2(2n+5)}{9n(n-1)}.$$



Гипотеза H_0 : монотонной зависимости нет.

Контроль вероятности ошибки первого рода:

$$P(H_0 \text{ отвергнута} | H_0) \leq \alpha.$$

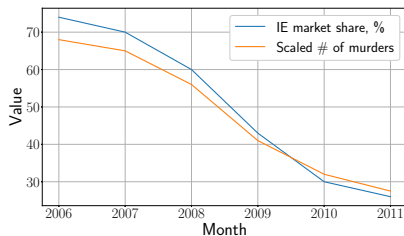
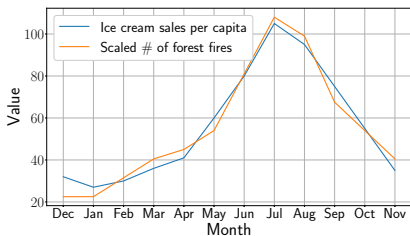
Мощность критерия: $P(H_0 \text{ отвергнута} | \overline{H_0}) \rightarrow \max.$

Критическая область: $|T(\mathbf{Z})| > t_\alpha.$

Множественное тестирование гипотез

$H_0 = \cup_{i \in M} H_0^i$, $M = \{1, \dots, m\}$, $M_0 = \{i : H_0^i \text{ — верна}\}$,
 $R = \{i : H_0^i \text{ — отвергнута}\}$.

	# верных	# неверных	Всего
# принятых H_0	U	T	$m - R$
# отвергнутых H_0	V	S	R
Всего	m_0	$m - m_0$	m



Меры качества:

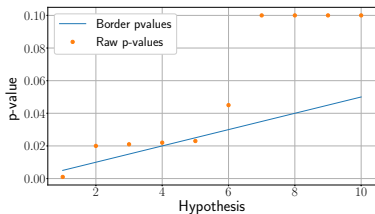
$$\text{FWER} = P(V \geq 1) \leq \alpha, \text{ FDR} = E \left(\frac{V}{R} I(R > 0) \right).$$

Поправка Бонферрони. Заменяем достигаемые уровни значимости p_1, \dots, p_m на поправленные (adjusted) уровни значимости $\tilde{p}_1, \dots, \tilde{p}_m$, где $\tilde{p}_i = \min(1, mp_i)$.

Теорема. Поправка Бонферрони обеспечивает $\text{FWER} \leq \frac{m_0\alpha}{m} \leq \alpha$.

Доказательство. $\text{FWER} = P(V \geq 1) = P\left(\bigcup_{j=1}^{m_0} \{p_{i_j} \leq \alpha/m\}\right) \leq \sum_{j=1}^{m_0} P(p_{i_j} \leq \alpha/m) \leq \frac{m_0\alpha}{m} \leq \alpha$.

Поправка Бенджамини-Хохберга.



Пусть $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$, тогда при положительной регрессионной зависимости для $p(p_1, \dots, p_m)$ при $\tilde{p}_{(m)} = \min(1, p_{(m)})$, $\tilde{p}_{(m-i)} = \min(1, \frac{m}{m-i} p_{(m-i)})$, $\tilde{p}_{(m-i+1)}$ обеспечивается $\text{FDR} \leq \frac{m_0}{m} \alpha$.

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

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

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

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?

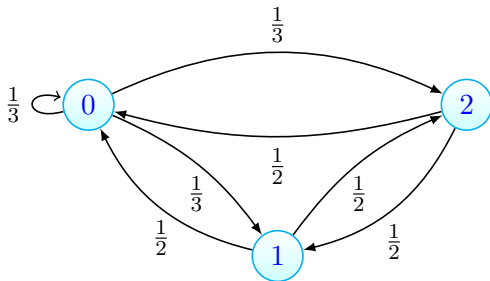
Пример: $K = 2$,

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

Наивный байесовский классификатор: продолжение

Пример. Классификация пользователей по интересующему атрибуту (например, полу, возрасту, достатку, интересу к некоторому товару) по истории 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. $p_{11} = -1, p_{22} = -100, p_{12} = 1, p_{21} = 1.$

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	$Z(\Theta)$
$\text{Be}(p)$	$p^x (1-p)^{1-x}$	x	$\log \frac{p}{1-p}$	$\frac{1}{1-p}$
$\text{Poisson}(\lambda)$	$\frac{\lambda^x}{x!} e^{-\lambda}$	x	$\log \lambda$	e^λ
$\Gamma(\alpha, \beta)$	$\frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}$	$[\log x, x]$	$[\alpha, -\beta]$	$\frac{\Gamma(\alpha)}{\beta^\alpha}$
$B(\alpha, \beta)$	$\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$	$[\log x, \log(1-x)]$	$[\alpha, \beta]$	$\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$
$\text{Dir}(\alpha)$	$\frac{\Gamma(\sum \alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_i p_i^{\alpha_i - 1}$	$[\log p_i]$	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:
$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-m)^2} = \underbrace{\frac{1}{\sqrt{2\pi\sigma e \frac{m^2}{2\sigma^2}}}}_{Z(\Theta)} e^{\underbrace{x}_{u_1(x)} \cdot \underbrace{\frac{m}{\sigma^2}}_{\theta_1} + \underbrace{x^2}_{u_2(x)} \cdot \underbrace{\frac{-1}{2\sigma^2}}_{\theta_2}},$$

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

Экспоненциальное семейство распределений.

Достаточные статистики.

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

Пример:
$$p(\mathbf{x}|\Theta) = \frac{1}{Z^n(\Theta)} \exp\left(\theta_1 \sum_{i=1}^n x_i + \theta_2 \sum_{i=1}^n x_i^2\right).$$

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

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

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

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

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

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

$$E \log x = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - \log \beta = \psi(\alpha) - \log \beta; \quad E x = \frac{\alpha}{\beta}.$$

Положительная регрессионная зависимость.

Пусть $\mathbf{p} = [p_1, \dots, p_m]^T$ вектор достигаемых уровней значимости в задаче множественной проверки гипотез, а $D \subseteq \mathbb{R}^m$ – возрастающее множество ($\mathbf{x} \in D, \mathbf{y} \geq \mathbf{x} \implies \mathbf{y} \in D$), тогда если $P(\mathbf{p} \in D | p_{i_1} = x_1, \dots, p_{i_j} = x_j)$ не убывает по (x_1, \dots, x_j) для любого набора (i_1, \dots, i_j) , то имеет место положительная регрессионная зависимость для совместного распределения $F(p_1, \dots, p_m)$.

Положительная регрессионная зависимость по каждому элементу из подмножества M_0 .

Пусть $\mathbf{p} = [p_1, \dots, p_m]^T$ вектор достигаемых уровней значимости в задаче множественной проверки гипотез, а $D \subseteq \mathbb{R}^m$ – возрастающее множество ($\mathbf{x} \in D, \mathbf{y} \geq \mathbf{x} \implies \mathbf{y} \in D$), тогда если $P(\mathbf{p} \in D | p_i = x_i), i \in M_0$ не убывает по x_i , то имеет место положительная регрессионная зависимость по каждому элементу подмножества M_0 для совместного распределения $F(p_1, \dots, p_m)$.

- 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 Kendall, Maurice G. "A new measure of rank correlation." *Biometrika* 30.1/2 (1938): 81-93.
- 6 Agresti, Alan. Analysis of ordinal categorical data. Vol. 656. John Wiley & Sons, 2010.
- 7 Дрейпер, Норман Р. Прикладной регрессионный анализ. Рипол Классик, 2007.
- 8 Кобзарь, Александр Иванович. Прикладная математическая статистика. Физматлит, 2006.