

Байесовский выбор моделей: введение

Александр Адуенко

5е сентября 2018

Байесовский выбор моделей: план курса

- Вводная лекция. Вспоминание результатов из теории вероятностей и статистики.
- Введение в байесовские методы. Базовые результаты и обозначения. Априорное распределение и неинформативные распределения (Jeffreys prior). Экспоненциальное семейство распределений.
- Байесовские модели классификации, регрессии, кластеризации, сокращения размерности.
- Понятие обоснованности в байесовском выборе моделей и его интерпретация.
- Построение интерпретируемых адекватных мультимodelей для описания сложных выборок.
- Построение и выбор моделей при анализе временных рядов. Гауссовские процессы.
- EM-алгоритм и вариационный вывод.
- Введение в графические модели.

Система оценивания

- 12 лекций + 3-4 небольших теста на них (суммарно до 100 баллов);
- 9 заданий:
 - 6 небольших (скорее теоретических) по 50 баллов,
 - 3 более крупных (скорее практических) по 100 баллов;
- Экзамен:
 - Письменная часть (200 баллов),
 - Устная часть (300 баллов).

Замечания:

- На оценку k требуется набрать $100k$ баллов;
- Экзамен можно пропустить только, если набрано не менее 550 баллов до экзамена;
- Задания содержат задачи более, чем на 50 / 100 баллов, поэтому можно выбрать, что выполнять;
- В каждом задании баллы лучшей работы удваиваются, если она оценена более, чем в 50 / 100 баллов (не более 125 / 250 баллов);
- За каждую неделю опоздания балл за задание снижается в 2 раза. Задание не принимается после его разбора или объявления об этом.

Задача. Пусть проводится эксперимент по угадыванию стороны выпадания честной монеты. Известно, что оракул прав с вероятностью $p_1 = 0.9$, а обычный человек с вероятностью $p_2 = 0.5$. Известно, что человек P оказался прав во всех $n = 10$ бросаниях. С какой вероятностью P является оракулом?

Совместное вероятности: $P(A \cdot B) = P(A)P(B|A) = P(B)P(A|B)$.

Формула Байеса: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.

$A = [P - \text{оракул}]$, $B = [n \text{ из } n]$.

Формула полной вероятности: $P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A})$,

$P(B|A) = p_1^n$, $P(B|\bar{A}) = p_2^n$.

$P(A|B) = \frac{P(A)p_1^n}{P(A)p_1^n + (1 - P(A))p_2^n}$.

Вопрос: Как определить $P(A)$?

Определение априорного распределения

Идея: из предыдущего опыта и разумных соображений выбрать $P(A)$.

Пример 1: отсутствие опыта (оракул и обычный неразличимы)

$$\rightarrow P(A) = 0.5$$

Пример 2: оракулов не бывает ($P(A) = 0$) или "я ни одного за свою жизнь не видел, но может, бывают" ($P(A) = 0.0001$).

Вопрос: только ли нашим опытом определяется априорное распределение? Может ли постановка эксперимента повлиять на априорное распределение?

Пример 3: Пусть человек P хочет выглядеть оракулом в прогнозе результатов двухпартийных выборов между партиями "прелестных" и "замечательных". На первых выборах P выбирает 1024 человека (вероятно, известных и уважаемых) и рассылает 512 из них прогноз «выиграют прелестные», а 512 оставшихся - «выиграют замечательные». Пусть выиграли "замечательные". Тогда 512 людей знают, что P верно предсказал исход выборов. Далее история повторяется 9 раз. Тогда в конце есть 1 человек, который знает, что P угадал результат 10 выборов из 10.

Случай 1 (честный эксперимент, нет selection bias)

Пусть $P(A) = 0.0001$ (основано на предыдущем опыте), тогда

$$P_n(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{0.0001 \cdot 0.9^n}{0.0001 \cdot 0.9^n + 0.9999 \cdot 0.5^n}.$$

$$P_{10}(A|B) = 0.0345; P_{20}(A|B) = 0.9273, P_{30}(A|B) = 0.9998.$$

Замечание: для $P(A) = 0.5$ $P_{10}(A|B) = 0.9972$; $P_{20}(A|B) = 0.999992$.

Случай 2 (предварительно выбран лучший из 100 случайно взятых людей по $k = 100$ попыткам)

Вопрос: сколько оракулов среди этих 100 случайно выбранных людей?

а) $P(\tilde{A}) = 0.5$; б) $P(\tilde{A}) = 0.0001$.

Эффективно при таком эксперименте меняется $P(A)$:

а) $P(A) \approx 1$; б) $P(A) = 0.01$.

$$P_n(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{0.01 \cdot 0.9^n}{0.01 \cdot 0.9^n + 0.99 \cdot 0.5^n}.$$

$$P_{10}(A|B) = 0.7829; P_{20}(A|B) = 0.9992, P_{30}(A|B) = 0.999998.$$

Тестирование гипотез

Пусть имеется выборка $\{x_1, \dots, x_n\}$.

$H_0 : p(x_1, \dots, x_n) \in P$, где P – некоторое множество распределений.

Требуется: проверить гипотезу H_0 на уровне значимости

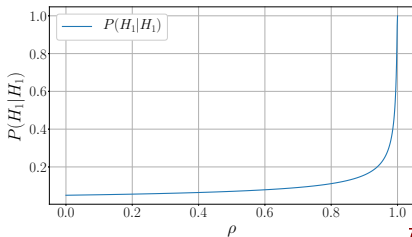
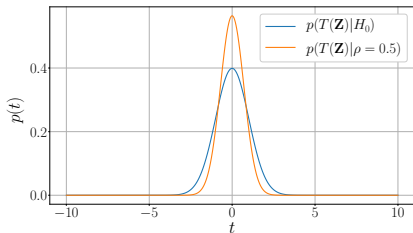
$P(H_0 \text{ отвергнута} | H_0) \leq \alpha$.

Пример: Пусть имеется выборка пар $\mathbf{z}_i = (x_i, y_i)$, $i = \overline{1, n}$,

$\mathbf{z}_i \sim N \left(\mathbf{z}_i | (0, 0)^\top, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$.

Гипотеза H_0 : $\rho = 0$

$T(\mathbf{Z}) = \frac{1}{\sqrt{2n}} \sum_{i=1}^n (x_i - y_i) \sim N(0, 1 - \rho)$.

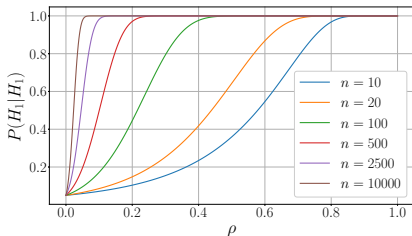
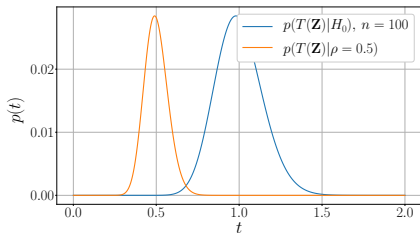


Тестирование гипотез: продолжение

$$T(\mathbf{Z}) = \frac{1}{2n} \sum_{i=1}^n (x_i - y_i)^2 = \frac{1 - \rho}{n} \xi, \quad \xi \sim \chi^2(n).$$

$$\frac{x_i - y_i}{\sqrt{2(1 - \rho)}} \sim N(0, 1) \implies \frac{(x_i - y_i)^2}{2(1 - \rho)} \sim \chi^2(1).$$

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

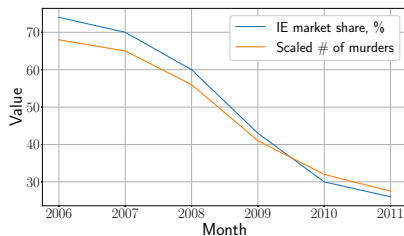
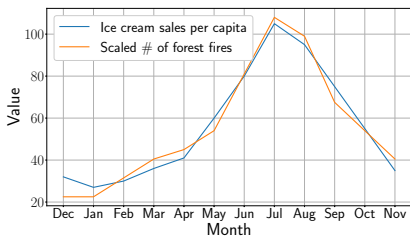


Вариант статистики: $T(\mathbf{Z}) = \frac{1}{n} \sum_{i=1}^n x_i y_i$.

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:

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

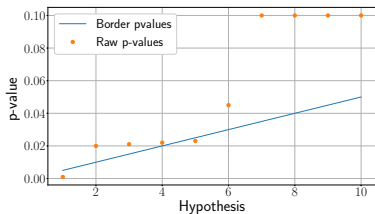
Поправки для учета эффекта множественных проверок

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?

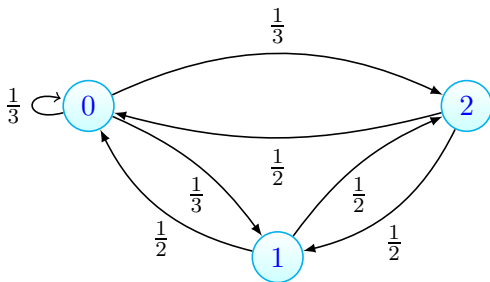
Пример: $K = 2$,

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

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

Положительная регрессионная зависимость.

Пусть $\mathbf{p} = [p_1, \dots, p_m]^\top$ 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 $\mathbf{p} = [p_1, \dots, p_m]^\top$ 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 Benjamini, Yoav, and Daniel Yekutieli. "The control of the false discovery rate in multiple testing under dependency." *Annals of statistics* (2001): 1165-1188.
- 6 Дрейпер, Норман Р. Прикладной регрессионный анализ. Рипол Классик, 2007.
- 7 Кобзарь, Александр Иванович. Прикладная математическая статистика. Физматлит, 2006.