Байесовская регрессия и классификация

Зинина Анастасия

Постановка задачи

- Задача: найти классификатор $a : X \to Y$ с минимальной вероятностью ошибки
- Пусть известна совместная плотность $p(x, y) = p(x)P(y \mid x) = p(y)P(x \mid y)$
- Принцип максимума апостериорной вероятности:

$$a(x) = \underset{y \in Y}{argmax} P(y \mid x)$$

Оптимальный байесовский классификатор/регрессор

- $L(y, a(x))$-функция потерь
- Функционал риска

$$R(a, x) = E(L(y, a(x)) \mid x)$$

- Минимизируем ожидаемые потери:
  - задача классификации:

$$a(x) = \underset{s}{argmin} R(s, x) = \underset{s}{argmin} \sum_{y \in Y} L(y, s) P(y \mid x) = \underset{s}{argmin} \sum_{y \in Y} L(y, s) P(y) P(x \mid y)$$

  - задача регрессии:

$$a(x) = \underset{s}{argmin} R(s, x) = \underset{s}{argmin} \int_{y \in Y} L(y, s) P(y \mid x) dy$$

Функционал среднего риска

- Оценим не риск для конкретного x, а в среднем

$$R(a) = E_x R(a(x), x)$$

- Для задачи классификации с дискретными признаками:

$$R(a) = \sum_{x \in X} R(a(x), x) P(x) \leqslant \sum_{x \in X} \min_s R(s, x) P(x)$$

- Оптимальный байесовский классификатор минимизирует и средний риск

Восстановление распределений

- Знаем, как построить классификатор, минимизирующий ожидаемые потери, если известны априорные вероятности классов $P(y)$ и функции правдоподобия $p(x \mid y)$
- Найдём эмпирические оценки $\hat{P}(y)$ и $\hat{p}(x \mid y)$

Эмпирическое оценивание

- **Оценка априорных вероятностей частотами:**

$$\hat{P}(y) = \frac{l_y}{l}, l_y - \text{количество объектов с данным значением y}$$

- **Оценка функций правдоподобия:**
  - параметрическое оценивание $\hat{p}(x) = \phi(x, \theta)$;
  - восстановление смеси распределений $\hat{p}(x) = \sum_{j=1}^{k} w_j \phi(x, \theta)$;
  - непараметрическое оценивание $\hat{p}(x) = \sum_{i=1}^{m} \frac{1}{mV(h)} K(\frac{\rho(x, x_i)}{h})$;

Наивный байесовский классификатор

- "Наивная"гипотеза: признаки $f_j : X \to D_j$ -независимые случайные величины с плотностями распределения $p_{y,j}(\xi)$

$$p_y(x) = p_{y,1}(\xi_1)...p_{y,n}(\xi_n), x = (\xi_1, ..., \xi_n)$$

- Получим классификатор:

$$a(x) = \max_{y \in Y}(\ln \lambda_y \hat{P}_y + \sum_{j=1}^{n} \ln \hat{p}_{yj}(\xi_j))$$

- Предположим, что распределение имеет определенный вид:

$$p(x) = \varphi(x; \theta)$$

- Определим параметр θ согласно принципу максимума правдогподобия:

$$L(\theta; X^m; G^m) = \sum_{j=1}^{n} g_i \ln \varphi(x_i; \theta) \to \max_{\theta},$$

где $(g_1, ..., g_m)$-вектор весов объектов

- Необходимое условие экстремума:

$$\frac{\partial}{\partial \theta} L(\theta; X^m, G^m) = \sum_{i=1}^{m} g_i \frac{\partial}{\partial \theta} \ln \varphi(x_i; \theta) = 0$$

Квадратичный дискриминант Фишера

- Пусть классы имеют n-мерные гауссовские плотности:

$$p_y(x) = N(x; \mu_y, \Sigma_y) = \frac{e^{-0.5(x-\mu_y)^T \Sigma_y^{-1}(x-\mu_y)}}{\sqrt{(2\pi)^n det \Sigma_y}}$$

- Разделяющая поверхность

$$\{x \in X \mid \lambda_t P_t p_t(x) = \lambda_s P_s p_s(x)\}$$

квадратична для всех $y, s \in Y,\ y \neq s$

- Если $\Sigma_y = \Sigma_s$, то она вырождается в линейную

Квадратичный дискриминант Фишера

- ОМП в нашем случае:

$$\hat{\mu}_y = \frac{1}{G_y} \sum_{i:y_i=y} g_i x_i;$$

$$\hat{\Sigma}_y = \frac{1}{G_y} \sum_{i:y_i=y} g_i (x_i - \hat{\mu}_y)(x_i - \hat{\mu}_y)^T,$$

где $G_y = \sum_{i:y_i=y} g_i$

- Получаем алгоритм-**квадратичный дискриминант:**

$$a(x) = \underset{y \in Y}{argmax}(\ln \lambda_y P_y - \frac{1}{2}(x - \hat{\mu}_y)^T \hat{\Sigma}_y^{-1}(x - \hat{\mu}_y) - 0.5 \ln det \Sigma_y)$$

Линейный дискриминант Фишера

- Допустим, что ковариационные матрицы классов равны $\Sigma_y = \Sigma$
- Получаем алгоритм-**линейный дискриминант:**

$$a(x) = \underset{y \in Y}{argmax}(\ln \lambda_y P_y - \frac{1}{2}\hat{\mu}_y^T \hat{\Sigma}_y^{-1} \hat{\mu}_y + x_T \hat{\Sigma}^{-1} \hat{\mu}_y) = \underset{y \in Y}{argmax}(x^T \alpha_y + \beta_y)$$

Усложнение модели - смесь распределений

- Модель плотности:

$$\text{p(x)}=\sum_{j=1}^{k} w_j p_j(x), \quad \sum_{j=1}^{k} w_j = 1, \quad w_j \geqslant 0, \quad p_j(x) = \varphi(x;\theta_j)$$

- Задачи:
  - оценить параметры w_j и θ_j
  - оценить k

Латентные переменные

- Латентные переменные нельзя выделить из данных напрямую, однако они сильно влияют вид наших данных.
- Два типа скрытых переменных:
  -дополнительные: нужны в дополнение к уже доступным данным для объяснения определённого вида данных в рамках выбранной модели. Примеры: пропущенные данные, смеси распределений

  -скрытые: можно извлечь из доступных данных и понять причины, влияющие на наблюдаемые данные. Пример: понижение размерности.
- Модели с латентными переменными часто имеют меньше параметров, чем модели, в которых корреляции представлены в "видимом"пространстве признаков

Идея EM-алгоритма на примере

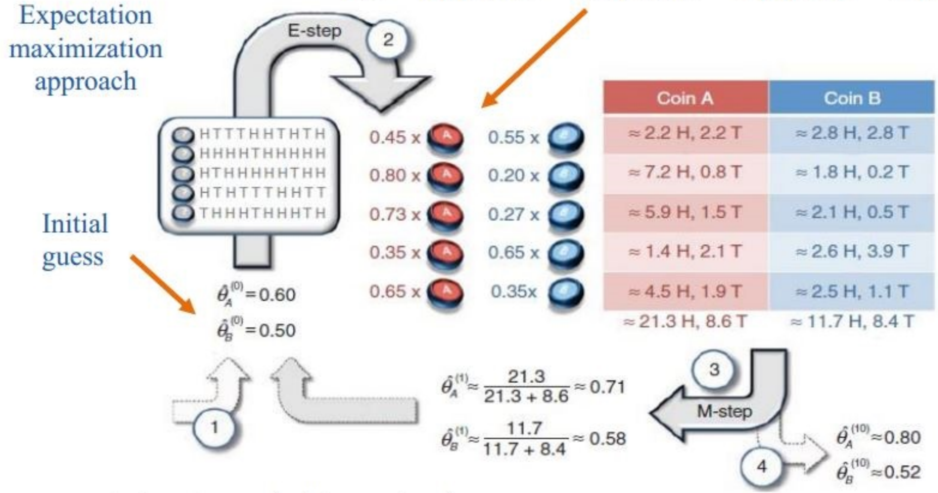
Простой эксперимент. Две монет (нет гарантии, что честные). Нужно оценить их смещение, повторяя 5 раз следующую процедуру: случайно выбираем одну из двух монет (с равной вероятностью) и делаем 10 независимых подбрасываний выбранной монеты.

$$Bin(k|n, \theta) = C_n^k \theta^k (1-\theta)^{n-k}$$

Идея EM-алгоритма на примере

Теперь предположим, что мы не знаем, какая из монет подбрасывалась.

$$p(z=A|\mathbf{X},\theta) = \frac{p(z=A|\theta)p(\mathbf{X}|z=A,\theta)}{p(z=A|\theta)p(\mathbf{X}|z=A,\theta)+p(z=B|\theta)p(\mathbf{X}|z=B,\theta)}$$



$$Bin(k|n,\theta) = C_n^k \theta^k (1-\theta)^{n-k}$$

EM-алгоритм

- Не применяем принцип максимума правдоподобия к модели смеси распределений, а вводим скрытые переменные.
- 2 повторяющихся до стабилизации параметров шага - E и M
- **E-шаг** - вычисляем вероятность принадлежности x_i к определенному классу с параметрами w_j, θ_j, вычисленными на предыдущем шаге.
    - Плотность вероятности того, что объет получен из j-ой компоненты смеси:
    $$p(x, \theta_j) = p(x)P(\theta_j \mid x) = w_j p_j(x)$$
    - Обозначим
    $$g_{ij} \equiv P(\theta_j \mid x_i) - \text{скрытые переменные}$$
    - $\sum\limits_{j=1}^{k} g_{ij} = 1$ для всех $i = 1, ..., l$
    - Зная параметры, можем вычислить скрытые переменные:
    $$g_{ij} = \frac{w_j p_j(x)}{p(x_i)} = \frac{w_j p_j(x)}{\sum\limits_{s=1}^{k} w_s p_s(x_i)}$$

KASPERSKY lab

- **M-шаг** - обновляем параметры, максимизируя логарифм правдоподобия:

$$Q(\theta) = \ln \prod_{i=1}^{m} p(x_i) = \sum_{i=1}^{m} \ln \sum_{j=1}^{k} w_j p_j(x_i) \to \max_{\theta}, \ \sum_{j=1}^{k} w_j = 1, w_j \geqslant 0$$

- Запишем функцию Лагранжа:

$$L(\Theta, X) = \sum_{i=1}^{m} \ln(\sum_{j=1}^{k} w_j p_j(x_i)) - \lambda(\sum_{j=1}^{k} w_j - 1)$$

- Производная равна нулю:

$$\frac{\partial L}{\partial w_j} = \sum_{i=1}^{m} \frac{p_j(x_i)}{\sum_{s=1}^{k} w_s p_s(x_i)} - \lambda = 0, \quad j = 1, ..., k. \quad (1)$$

$$\sum_{i=1}^{m} \sum_{j=1}^{k} \frac{w_j p_j(x_i)}{\sum_{s=1}^{k} w_s p_s(x_i)} = \lambda \sum_{j=1}^{k} w_j$$

$$\lambda = m$$

- Умножим (1) на w_j, подставим $\lambda = m$:

$$w_j = \frac{1}{m} \sum_{i=1}^{m} \frac{w_j p_j(x_i)}{\sum_{s=1}^{k} w_s p_s(x_i)} = \frac{1}{m} \sum_{i=1}^{m} g_{ij}, \quad j = 1, ..., k$$

- Теперь приравняем нулю производную функции Лагранжа по θ_j:

$$\frac{\partial L}{\partial \theta_j} = \sum_{i=1}^{m} \frac{w_j}{\sum_{s=1}^{k} w_s p_s(x_i)} \frac{\partial}{\partial \theta_j} p_j(x_i) = \sum_{i=1}^{m} \frac{w_j p_j}{\sum_{s=1}^{k} w_s p_s(x_i)} \frac{\partial}{\partial \theta_j} \ln p_j(x_i)$$

$$= \sum_{i=1}^{m} g_{ij} \frac{\partial}{\partial \theta_j} \ln p_j(x_i) = \frac{\partial}{\partial \theta_j} \sum_{i=1}^{m} g_{ij} \ln p_j(x_i) = 0, \quad j = 1, ..., k.$$

EM-RBF

Допущения:

- функции правдоподобия классов представимы в виде смесей k_y компонент, $y \in Y = \{1, ..., M\}$
- компоненты имеют n-мерные гауссовские плотности с некоррелированными признаками:

$$\mu_{yj} = (\mu_{yj1}, ..., \mu_{yjn}), \quad \Sigma_{yj} = diag(\sigma^2_{yj1}, ..., \sigma^2_{yjn}), \quad j = 1, ..., k_y :$$

$$p_y(x) = \sum_{j=1}^{k_y} w_{yj} p_{yj}(x), \quad p_{yj}(x) = N(x; \mu_{yj}, \quad \Sigma_{yj}), \quad \sum_{j=1}^{k_y} w_{yj} = 1, \quad w_{yj} \geqslant 0$$

EM-RBF

- $\phi(x;\theta_j) = \prod_{d=1}^{n} \frac{1}{\sigma_{jd}\sqrt{2\pi}} exp(-0.5(\frac{x_{id}-\mu_{jd}}{\sigma_{jd}})^2)$
- M-шаг:

$$\hat{\mu}_{jd} = \frac{1}{mw_j}\sum_{i=1}^{m} g_{ij}x_{id},$$

$$\hat{\sigma}^2_{jd} = \frac{1}{mw_j}\sum_{i=1}^{m} g_{ij}(x_{id}) - \hat{\mu}_{jd})^2$$

EM-RBF

- Подставим гауссовскую смесь в байесовский классификатор:

$$a(x) = \underset{y \in Y}{argmax}\ \lambda_y P_y \sum_{j=1}^{k_y} w_{yj} N_{yj} exp(-0.5\rho_{yj}^2(x, \mu_{yj})),$$

$N_{yj} = (2\pi)^{-\frac{n}{2}}(\sigma_{yj1}...\sigma_{yjn})^{-1}$ – нормировочные множители;

$\rho_{yj}(x, \mu_{yj})$ – взвешенная евклидова метрика в $X = R^n$:

Непараметрический подход

Определим плотность и её оценку:

- Дискретный случай:

$$\hat{p}(x) = \frac{1}{m}\sum_{i=1}^{m}[x_i = x]$$

- Одномерный непрерывный случай:

$$p(x) = \lim_{h\to 0}\frac{1}{2h}P[x-h, x+h]$$

Эмпирическая оценка плотности по окну ширины h:

$$\hat{p}_h(x) = \frac{1}{2mh}\sum_{i=1}^{m}[\mid x - x_i \mid < h]$$

KASPERSKY lab

Оценка Парзена-Розенблатта

- $\hat{p}_h(x) = \frac{1}{mh}\sum_{i=1}^{m}\frac{1}{2}[\frac{|x-x_i|}{h} < 1]$

- Обобщение: оценка Парзена-Розенблатта по окну ширины h:

$$\hat{p}_h(x) = \frac{1}{mh}\sum_{i=1}^{m} K(\frac{x-x_i}{h}),$$

  где $K(r)$ -ядро, удовлетворяющее требованиям:
  - чётная функция;
  - нормированная функция;
  - невозрастающая, неотрицательная функция

Оценка Парзена-Розенблатта

Теорема
При выполнении следующих условий:

1. X^m-простая выборка из распределения $p(x)$;
2. ядро $K(z)$ непрерывно и ограничено: $\int_X K^2(z)dz < \infty$
3. $\lim\limits_{m\to\infty} h_m = 0$ и $\lim\limits_{m\to\infty} mh_m = \infty$

имеет место утверждение:

$$\hat{p}_{h_m} \to p(x) \text{ при } m \to \infty \text{ для почти всех } x \in X$$

Обобщение на многомерный случай

1. Если объекты описываются n числовыми признаками $f_j : X \to R, j = 1, ..., n$

$$\hat{p}_h(x) = \frac{1}{m}\sum_{j=1}^{m}\prod_{i=1}^{n}\frac{1}{h_j}K(\frac{f_j(x) - f_j(x_i)}{h_j})$$

2. Если на X задана функция расстояния $\rho(x, x^{'})$:

$$\hat{p}_h(x) = \frac{1}{mV(h)}\sum_{j=1}^{m}\frac{1}{h_j}K(\frac{\rho(x, x_i)}{h}),$$

где $V(h) = \int_X K(\frac{\rho(x,x_i)}{h})dx$ -нормирующий множитель, не зависящий от x_i

Метод парзеновского окна

$$a(x; X^l, h) = \underset{y \in Y}{argmax} \Gamma_y(x),$$

$$\Gamma_y(x) = \lambda_y \frac{P_y}{l_y} \sum_{j=1}^{m} \frac{1}{h_j} K(\frac{\rho(x, x_i)}{h})$$

Варианты ядер:

- $Q(r) = \frac{15}{16}(1 - r^2)^2[\mid r \mid \leqslant 1]$ — квартическое;
- $T(r) = (1 - \mid r \mid)[\mid r \mid \leqslant 1]$ — треугольное;
- $G(r) = (2\pi)^{-1/2} exp(-\frac{1}{2} r^2)$ — гауссовское;
- $(r) = \frac{1}{2}[\mid r \mid \leqslant 1]$ — прямоугольное