Глава 1

Дебильник

1.1 Многомерное нормальное распределение

<u>def</u>. Стандартный гауссовский вектор — случайный n-мерный вектор $Z = (Z_1, Z_2, \dots Z_n)$, координаты которого независимы и имеют распределение $\mathcal{N}(0, 1)$.

<u>def</u>. Гауссовский вектор (Нормальный вектор) — вектор, для которого существует матрица $\mathbf{A} \in \mathbb{R}^{n \times m}$, стандартный гауссовский вектор $Z \in \mathbb{R}^m$, и вектор $b \in \mathbb{R}^n$ такие, что $X = \mathbf{A}Z + b$.

<u>def</u>. Распределение нормального вектора $X \in \mathbb{R}^n - \mathcal{N}(\mu, \mathbf{\Sigma})$ или $\mathcal{N}_n(\mu, \mathbf{\Sigma})$, где $\mu = \mathbb{E}X$ и $\mathbf{\Sigma} = \operatorname{cov}(X)$.

<u>def</u>. Распределение хи-квадрат с n степенями свободы — распределение $\chi^2(n)$ величины $\chi^2 = Z_1^2 + Z_2^2 + \ldots + Z_n^2$, где $Z_1, Z_2, \dots Z_n$ — независимы $\mathcal{N}(0, 1)$ величины.

<u>def</u>. Распределение Стьюдента с n степенями свободы — распределение $T(n)$ величины $\frac{\sqrt{n}X}{\sqrt{Y}}$, где $X \sim \mathcal{N}(0, 1)$, $Y \sim \chi^2(n)$ и независимы.

<u>def</u>. Распределение Фишера со степенями свободы n и m — распределение $F(n, m)$ величины $\frac{X/n}{Y/m}$, где $X \sim \chi^2(n)$, $Y \sim \chi^2(m)$ и независимы.

1.2 Условное матожидание

def. Условное матожидание $\mathbb{E}(Y \mid X)$ случайной величины Y при условии случайной величины X — такая измеримая функция g_0 величины X, при которой $\mathbb{E}(Y - g(X))^2$ минимально для всех измеримых функций g.

Условное матожидание — ортогональная проекция Y на линейное пространство всех измеримых функций X. То есть УМО — единственная измеримая функция, которая удовлетворяет условию ортогональности:

$$\forall g \colon \mathbb{E}(Y - \mathbb{E}(Y \mid X))g(X) = 0.$$

1.3 Статистическая модель, выборка

def. Статистическая модель — множество распределений $\mathfrak{P}$, которое, по нашему мнению, адекватно приближает $\mathcal{P}_D$.

def. Данные d — реализация случайного элемента D, имеющего распределение $\mathcal{P}_D$.

Статистические модели делят на:

- параметрические, если $\mathfrak{P} = \{\mathcal{P}_0 \mid \theta \in \Theta \subset \mathbb{R}^k\}$.

  Пример: $\mathfrak{P} = \{\mathcal{N}(\mu, \sigma^2) \mid \mu \in \mathbb{R}, \sigma^2 \geqslant 0\}$.

- непараметрические, если $\mathfrak{P} = \{\mathcal{P}_0 \mid \theta \in \Theta \subset V\}$, где V не обязательно конечномерное.

  Пример: $\mathfrak{P} = \{\mathcal{P}^{\otimes n} \mid \int_{\mathfrak{X}} x\mathcal{P}(dx) = 0\}$

- семипараметрические, если $\mathfrak{P} = \{\mathcal{P}_0 \mid \theta \in \Theta \subset \mathbb{R}^k \times V\}$.

  Пример: линейная регрессия $Y = X\beta + \varepsilon$, $\beta \in \mathbb{R}^k$, $\mathbb{E}\varepsilon = 0$, $\mathbb{D}\varepsilon = \sigma^2$.

Если $D = [X_1, \dots X_n]$ и X_i независимы и имеют одинаковое распределение $\mathcal{P}_X$, D называется выборкой объема n и обозначается $X_{[n]}$, $\mathcal{P}_X$ — генеральная совокупность. В этом случае модель приобретает вид $\mathfrak{P} = \{\mathcal{P}^{\otimes n} \mid \mathcal{P} \in \mathfrak{P}_X\}$, где $\mathfrak{P}_X$ — модель для $\mathcal{P}_X$.

1.4 Формула Байеса, априорное, апостериорное распределение

- Априорное распределение — наше ощущение относительно значения параметра до проведения эксперимента.
- Апостериорное распределение — ощущение после получения данных эксперимента.

def (Формула Байеса)**.** Здесь p — вероятность, d — данные, θ — параметры.

$$p(\theta \mid d) = \frac{p(d \mid \theta) \cdot p(\theta)}{p(d)}.$$

- $p(\theta \mid d)$ — апостериорное распределение,
- $p(d \mid \theta)$ — правдоподобие,
- $p(\theta)$ — априорное распределение,
- $p(d)$ — вероятность данных.

1.5 Расстояние Кульбака-Лейблера, энтропия

Пусть мы принимаем случайные символы $x_1, \dots x_k$, вероятность появления x_i равна p_i, записываем с помощью битовой строки длины l_i. Тогда средняя длина символа равна

$$l = \sum_{i=1}^{k} p_i \cdot l_i.$$

Чтобы минимизировать l, необходимо подобрать следующие $l_i = -\log_2 p_i$. И тогда средняя длина будет равна $H(x) \coloneqq -\sum_{i=1}^{k} p_i \cdot \log_2 p_i$, эта величина называется двоичной энтропией сообщения. Аналогично можно брать любой другой логарифм, мы будем использовать натуральный.

Для непрерывной величины можно завести дифференциальную энтропию:

$$H(X) = -\int p(x) \log p(x) dx.$$

Пусть случайная величина X имеет функцию вероятности p, но мы кодируем символы, как-будто она имеет функцию вероятности q. Тогда средняя длина сообщения будет равна $-\sum_{i=1}^{k} p_i \cdot \log q_i$, эта величина называется кросс-энтропией $H(p \mid q)$ распределений p и q.

$H(p \mid q)$ всегда будет больше $H(p)$, так как $H(p)$ минимально.

<u>def</u>. Величина потери информации из-за использования q вместо p называется расстоянием Кульбака-Лейблера между p и q:

$$D_{KL}(p, q) = H(p \mid q) - H(p) = -\sum_{i=1}^{k} p_i \cdot \log \frac{q_i}{p_i}.$$

Для непрерывных величин все обобщается следующим образом

$$D_{KL} = -\int p_i \cdot \log \frac{q_i}{p_i}.$$

1.6 Статистика...

1.6.1 Статистика

Параметр или характеристика распределения — функционал от этого распределения.

<u>def</u>. Статистика — функция θ^* от данных d.

Пусть модель $\mathfrak{P}_{[n]} = \{\mathcal{P}^{\otimes n} \mid \mathcal{P} \in \mathfrak{P}\}$, искомая характеристика $\theta \colon \mathfrak{P} \to \mathbb{R}^k$.

1.6.2 Несмещенность

Чему равна оценка как случайная величина в среднем, если она равна характеристике?

<u>def</u>. Оценка Θ^* называется

- несмещенной, если $\forall \mathcal{P} \in \mathfrak{P} \colon \mathbb{E}\theta^*(X_{[n]}) = \theta(\mathcal{P})$, где $X_{[n]} \sim \mathcal{P}^{\otimes n}$,

- асимптотически несмещенной, если $\forall \mathcal{P} \in \mathfrak{P}\colon \mathbb{E}\theta^*(X_{[n]}) \to \theta(\mathcal{P})$.

Смещение — величина $b(\theta^*) = \mathbb{E}(\theta^*(X_{[n]})) - \theta(\mathcal{P})$.

Среднеквадратичная ошибка — величина $\mathrm{MSE}(\theta^*) = \mathbb{E}\left(\theta^*(X_{[n]}) - \theta(\mathcal{P})\right)^2$.

В общем случае
$$\mathrm{MSE}(\theta^*) = \mathbb{D}\theta^*(X_{[n]} + b^2(\theta^*).$$

- Выборочное среднее как оценка матожидания — несмещенная оценка,
- Выборочная дисперсия как оценка дисперсии — асимптотически несмещенная,
- Исправленная выборочная дисперсия как оценка дисперсии — несмещенная оценка.

1.6.3 Состоятельность

def. Оценка θ^* называется

- состоятельной, если $\forall \mathcal{P} \in \mathfrak{P}\colon \theta^*(X_{[n]}) \xrightarrow{\mathbb{P}} \theta(\mathcal{P})$, где $X_{[n]} \sim \mathcal{P}^{\otimes n}$,
- сильно состоятельной, если $\theta^*(X_{[n]}) \xrightarrow{\text{п. н.}} \theta(\mathcal{P})$.

1.6.4 Асимптотическая нормальность

def. Оценка θ^* называется асимптотически нормальной с коэффициентом рассеивания (или просто дисперсией) $\sigma^2\big(\theta(\mathcal{P})\big\rangle 0$, если
$$\sqrt{n}\left(\theta^*(X_{[n]}) - \theta(\mathcal{P})\right) \xrightarrow{d} \eta \sim \mathcal{N}(0, \sigma^2(\theta^*(\mathcal{P}))).$$
В многомерном случае рассматривается ковариационная матрица вместо дисперсии.

- Выборочная дисперсия и второй момент — асимптотически нормальная оценка.
- Из асимптотической нормальности следует состоятельность.

1.6.5 Эффективность

Рассмотрим класс оценок $K = \{\hat{\theta}\}$ параметра θ.

def. Оценка $\theta^* \in K$ называется **эффективной в классе** K, если *для любой другой оценки* $\hat{\theta} \in K$ *и для любого исследуемого параметра* $\theta \in \Theta$ *выполняется*

$$\mathrm{MSE}_\theta(\theta^*) \leqslant \mathrm{MSE}_\theta(\hat{\theta}).$$

Класс несмещенных оценок

$$K_0 = \{\hat{\theta} \mid \mathbb{E}\hat{\theta} = \theta, \forall\theta \in \Theta\}.$$

def. **Эффективная оценка** θ^*, если эффективна в классе K_0.

def. **Асимптотически эффективной в классе** K, если для любой оценки $\hat{\theta} \in K$ и для любого $\theta \in \Theta$ выполняется

$$\overline{\lim_{n\to\infty}} \frac{\mathrm{MSE}(\theta^*)}{\mathrm{MSE}(\hat{\theta}}.$$

1.6.6 Робастность

def. **Робастность** — свойство оценки быть устойчивой к хвостам распределения.

Пусть F — распределение, $\{G_n\}$ — последовательность распределений, что

$$|F - G_n| \coloneqq \sup_x |F(x) - G_n(x)| \to 0.$$

def. Характеристика θ обладает **качественной робастностью**, если $\theta(G_n) \to \theta(F)$

Пусть также δ_x — вырожденное распределение в точке x.

def. **Загрязненное распределение** — смесь $F_{x,\varepsilon} = (1-\varepsilon)F + \varepsilon\delta_x$.

def. **Функция влияния** характеристики θ — величина

$$IF(x) = \lim_{\varepsilon\to 0+} \frac{\theta(F_{x,\varepsilon}) - \theta(F)}{\varepsilon}.$$

def. Характеристика θ называется B-робастной или инфинитезимально робастной, если $IF(x)$ ограничена.

def. Асимптотическая толерантность характеристики θ —

$$\tau = \inf\bigl\{\varepsilon \mid \sup_x |\theta(F_{x,\varepsilon} - \theta(F)| = \infty\bigr\}.$$

1.6.7 Достаточность

def. Статистика $T(x) = \{T_1(x), \dots, T_m(x))\}$ называется достаточной, если для всех

- $\theta \in \Theta$,
- $B \in \mathfrak{P}(\mathbb{R}^n)$ и
- $t = (t_1, \dots, t_m)$

условная вероятность $\mathbb{P}(X_{[n]} \in B \mid T(X_{[n]}) = t)$ не зависит от θ.

То есть информация о θ в выборке полностью содержится в значении $T(x_{[n]})$.

thm (факторизации). *$T(x)$ достаточна, согда существуют функции g и h, что*

$$p(X_{[n]} = x_{[n]} \mid \theta) = g(T(x_{[n]}), \theta)h(x_{[n]}),$$

где p — вероятность или плотность.

1.6.8 Полнота

def. Статистика T называется полной, если для любой *измеримой g верно следствие*

$$\forall \theta \in \Theta \colon \mathbb{E}g(T(X_{[n]})) \equiv 0 \quad \Longrightarrow \quad g(T(X_{[n]})) \overset{\text{п.н.}}{=} 0.$$

1.7 Теоремы Колмогорова-Блэкуэлла-Рао и Лемана-Шеффе

thm (Колмогорова-Блэкуэлла-Рао). *Пусть θ^* — оценка параметра θ, T — достаточная статистика. Тогда*

$$\mathrm{MSE}(\theta^*) \geqslant \mathrm{MSE}(\mathbb{E}(\theta^* \mid T)).$$

thm (Лемана-Шеффе). *Пусть θ^* — оценка параметра θ, T — достаточная и полная статистика. Тогда $\mathbb{E}(\theta^* \mid T)$ — единственная эффективная оценка в классе оценок со смещением $b(\theta^*)$.*

1.8 Доверительный интервал

Пусть есть модель $\mathfrak{P}_{[n]} = \{\mathcal{P}^{\otimes n} \mid \mathcal{P} \in \mathfrak{P}\}$ и $\theta\colon \mathfrak{P} \to \mathbb{R}^k$ — искомая характеристика.

def. Доверительный интервал (точный доверительный интервал) с уровнем доверия γ — пара статистик (θ^*_L, θ^*_R), такая что для любого $\mathcal{P} \in \mathfrak{P}$ и $X_{[n]} \sim \mathcal{P}^{\otimes n}$

$$\mathbb{P}\left(\theta^*_L(X_{[n]}) \leqslant \theta(\mathcal{P}) \leqslant \theta^*_R(X_{[n]})\right) = \gamma.$$

Интервал называется

- асимптотическим, если
$$\mathbb{P}\left(\theta^*_L(X_{[n]}) \leqslant \theta(\mathcal{P}) \leqslant \theta^*_R(X_{[n]})\right) \xrightarrow{n\to\infty} \gamma.$$
- центральным, если
$$\mathbb{P}\left(\theta^*_L(X_{[n]}) > \theta(\mathcal{P})\right) = \mathbb{P}\left(\theta^*_R(X_{[n]}) < \theta(\mathcal{P})\right).$$
- левым, если
$$\mathbb{P}\left(\theta^*_L(X_{[n]}) > \theta(\mathcal{P})\right) = 0.$$
- правым, если
$$\mathbb{P}\left(\theta^*_R(X_{[n]}) < \theta(\mathcal{P})\right) = 0.$$

1.9 Бутстреп

1.9.1 Параметрический бутстреп

Если работаем с параметрической моделью, можем заменить $X = X(\theta)$ не на X^*, а на $X(\theta^*)$ и сэмплировать из этого распределения.

1.9.2 Непараметрический бутстреп

Рецепт

1. изготовим N выборок $x^*_{[n],1}, \dots, x^*_{[n],N}$ из эмпирического распределения (рандом с возвращением)
2. вычисляем $\theta^b_i = \theta^*(x^*_{[n],i}$, получаем бутстреповскую выборку $\theta^b_{[N]}$,
3. по бутстреповской выборке оцениваем, что нужно.

Ограничения

- θ^* — plug-in оценка
- θ^* — достаточно гладкая (обычно дифференцируема)
- у X достаточно много моментов (обычно конечная дисперсия)
- нужно генерировать большие выборки
- на очень больших данных трудозатратен
- на маленьких данных велика неустранимая ошибка