Прикладная статистика в машинном обучении

Семинар 7

---

26 октября 2020 г.

Задача 1. Теорема Хершелла-Максвелла.

Рассмотрим замкнутую плоскую фигуру внутри которой случайно летают частицы. Обозначим $V = (V_x, V_y)'$ – вектор скоростей случайно выбранной частицы.

Предположим, что

1. Распределение вектора V не должно меняться при его повороте на любой угол (то есть не зависит от направления вектора).
2. V_x и V_y независимы.
3. $\mathrm{Var}(V_x) = \mathrm{Var}(V_y) = 1$.
4. $f(v_x, v_y)$ существует и непрерывна.

a) Найдите координаты вектора V', который получается поворотом вектора V на $90°$ против часовой стрелки.

b) Докажите, что распределения V_x, V_y и $-V_y$ совпадают.

c) По предположению 3 понятно, что $\mathrm{Var}(V_x) = \mathrm{Var}(V_y) = 1$. Покажите, что $\mathbb{E}(V_x) = \mathbb{E}(V_y) = 0$.

d) Докажите, что совместная функция плотности $f_V(v_x, v_y)$ представима в виде

$$f_V(v_x, v_y) = h(v_x^2 + v_y^2),$$

где h – некоторая функция. Сделайте вывод из этого утверждения.

e) Покажите, что совместная функция плотности $f_V(v_x, v_y)$ представима в виде

$$f_V(v_x, v_y) = g(v_x^2)g(v_y^2),$$

где g – некоторая функция.

f) Докажите, что выражение $h'(t)/h(t)$ равно константе.

g) Найдите функцию $h(t)$ с точностью до константы.

h) Выпишите $f_V(v_x, v_y)$ с точностью до константы.

Задача 2. Независимость длин проекций.

Пусть вектор $u \in \mathbb{R}^3$ имеет многомерное стандартное нормальное распределение.

a) Найдите проекцию вектора u на подпространство $V = \{(x_1, x_2, x_3) \mid x_1 = x - 2 = x_3\}$.

b) Найдите проекцию вектора u на подпространство $W = \{(x_1, x_2, x_3) \mid x_2 = 0, x_1 = -x_3\}$.

c) Являются ли V и W ортогональными?

d) Убедитесь, что $\|\hat{u}_V\|$ и $\|\hat{u}_W\|$ независимы.

Задача 3. Распределение хи-квадрат.

Теорема 1. *Пусть вектор $u \in \mathbb{R}^n$ имеет многомерное стандартное нормальное распределение. Пусть $V \subset \mathbb{R}^n$, $\dim V = k$. Тогда $\|\hat{u}_V\|^2 \sim \chi^2_k$.*

a) Пусть вектор $u \in \mathbb{R}^3$ имеет многомерное стандартное нормальное распределение. Найдите проекцию и распределение квадрата длины проекции вектора u на подпространство $V = \{(x_1, x_2, x_3 \mid x_3 = 2x_1 + x_2)\}$.

b) Найдите распределение квадрата длины проекции вектора u на подпространство $V^\perp$.

Задача 4. F-распределение.

Определение 1. *Пусть случайные величины $X \sim \chi^2_a$, $Y \sim \chi^2_b$, и X и Y независимы. Тогда случайная величина Z имеет распределение Фишера с a и b степенями свободы:*

$$Z = \frac{X/a}{Y/b} \sim F_{a,b}.$$

Пусть вектор $u \in \mathbb{R}^n$ имеет многомерное стандартное нормальное распределение. Пусть V и W – ортогональные подпространства в $\mathbb{R}^n$, $\dim V = k$. $\dim W = m$.

a) Постройте проекции вектора u на V и W. Как распределены квадраты длин этих проекций?

b) Постройте проекцию вектора u на $V \cup W$.

c) Покажите на рисунке угол, квадрат тангенса которого равен отношению квадратов длин проекций u на V и W. Обозначим этот угол как α.

d) Какое распределение имеет величина
$$\frac{m}{k}\tan^2\alpha?$$

e) Поясните идею сравнения прогнозов моделей при помощи F-распределения.

Задача 5. Применение F-распределения.

Теорема 2. *Пусть имеются UR-модель и R-модель и тестируется гипотеза*

$$H_0 : \text{Верны и UR- и R-модели}$$

против

$$H_A : \text{UR-модель верна, а R-модель неверна.}$$

Тогда H_0 отвергается на уровне значимости α, если F-статистика

$$F = \frac{(RSS_R - RSS_{UR})/(df_R - df_{UR})}{RSS_{UR}/df_{UR}}$$

превышает критическое значение F_α. Здесь df – число степеней свободы в соответствующей модели, $df = n - k$, где n – число наблюдений, k – число регрессоров.

Пусть UR-модель задаётся следующим образом:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + u_i.$$

Тестируется следующая R-модель:

$$y_i = \beta_0 + \beta_1 (x_i + z_i) + u_i.$$

Предположим, что $u_i \sim \mathcal{N}(0, 1)$ и независимы.

a) Постройте подпространства V_{UR} и V_R и проекции вектора y на них.

b) Покажите RSS_{UR} и RSS_R. Обозначим угол между ними как α.

c) Покажите, что $RSS_R - RSS_{UR} = \|\hat{y}_{UR} - \hat{y}_R\|^2$.

d) Выразите $\tan^2 \alpha$ через y, $\hat{y}_{UR}$, $\hat{y}_R$.

e) Рассмотрим векторы $y - \hat{y}_{UR}$ и $\hat{y}_{UR} - \hat{y}_R$. Найдите размерности подпространств, в которых они лежат.

f) Выпишите F-статистику в геометрическом и в классическом смыслах.