5. Условные математические ожидания и условные распределения

1. Пусть $\xi = (\xi_1, \xi_2) \sim N(a, \Sigma)$, где $a = \begin{pmatrix} 1 \\ 4 \end{pmatrix}$ и $\Sigma = \begin{pmatrix} 10 & 8 \\ 8 & 10 \end{pmatrix}$. Постройте график плотности этого случайного вектора. Для $y \in \{-3, 0, 1, 5\}$ постройте графики $f_{\xi_1|\xi_2}(x|y)$. Построить график $\mathsf{E}(\xi_1|\xi_2 = y)$ в зависимости от y и проведите на этом графике прямую $x = \mathsf{E}\xi_1$.

2. Рассмотрим модель смеси гауссовских распределений

$$p(x) = \sum_{k=1}^{K} p_k(x)\, \mathsf{P}(T = k),$$

где T — номер компоненты смеси, а $p_k(x)$ — плотность распределения $\mathcal{N}(a_k, \Sigma_k)$. Загрузите данные из набора Ирисы Фишера. В этих данных представлена выборка из распределения случайного 4-мерного вектора X, являющегося смесью трех гауссовских векторов. Иными словами, выборка состоит из трех частей ($K = 3$), каждой соответствует номер от 1 до 3 (номер компоненты смеси, в сооответствии с которым выборка представлена в смеси). Сама выборка находится в `load_iris()["data"]`, а в `load_iris()["target"]` записаны номера компонент смеси, соответствующие каждому элементу выборки. В предположении, что каждая из трех выборок имеет гауссовское распределение, оцените их параметры $(a_1, a_2, a_3, \Sigma_1, \Sigma_2, \Sigma_3)$.

Параметры распределения $\mathcal{N}(a, \Sigma)$ по выборке $X_1, \ldots, X_n$ оцениваются следующим образом: $\hat{a} = \overline{X}$, оценка ковариация i-ой и j-ой компонент равна $\overline{X^i X^j} - \overline{X^i} \cdot \overline{X^j}$ (или, что то же самое, оценка матрицы Σ равна $\frac{1}{n} X^T X$, где $X_{i,j} = X_i^j - \overline{X^j}$).

Занумеруем координаты данных векторов числами $0, 1, 2, 3$. Для пар координат $(0, 1)$, $(1, 3)$ и $(2, 3)$ вычислите плотность каждой компоненты смеси (три плотности для каждой пары координат), оценив параметры распределений по проекциям трех выборок (каждая выборка соответствует одной компоненте смеси) на соответствующие плоскости. Нарисуйте графики (рисовать нужно линии уровня) этих плотностей ($3 \times 3 = 9$ штук), на которые нанесите также соответствующие проекции точек выборки.

Будем считать, что случайные векторы T и X независимы. Оцените вероятности $\mathsf{P}(T = k)$ частотами вхождений k-ых компонент смеси в данную выборку. На основе полученных оценок вычислите условное математическое ожидание $\mathsf{E}(X|T \neq k)$ для всех $k = 1, 2, 3$ (три числа). Для пар координат $(0, 1)$, $(1, 3)$ и $(2, 3)$ получите новые оценки (распределения T) и постройте графики условной плотности $p_{(X|I\{T \neq k\})}(x|1)$ (9 штук).

Классифицируйте все пространство 4-мерных векторов по принципу

$$k = \arg\max_k p_{(X|I\{T \neq k\})}(x|1)$$

(здесь условная плотность считается на основе оценок для 4-мерных векторов). Посчитайте долю ошибок на выборке. Нарисуйте классификацию всего пространства в проекции на пары координат $(0, 1)$, $(1, 3)$ и $(2, 3)$, где закрасьте разными цветами области, которые образовались в результате классификации.