Санкт-Петербургский политехнический университет
Петра Великого

Институт прикладной математики и механики
Кафедра «Прикладная математика»

Отчет по лабораторной работе № 1
по дисциплине: теория принятия экономических решений.

Выполнила студентка:
Заболотских Екатерина Дмитриевна
группа: 3630102/70301

Проверил:
Павлова Людмила Владимировна

Санкт-Петербург
2020 г.

Оглавление

Постановка задачи 2

Решение 3

Результаты и выводы 5

Приложение 10

Список иллюстраций

Рисунок 1: OB size = 300 5

Рисунок 2: TB size = 50 6

Рисунок 3: OB size = 100 7

Рисунок 4: TB size = 50 8

Список таблиц

Таблица 1: OB size = 300 5

Таблица 2: TB size = 50 6

Таблица 3: OB size = 50 7

Таблица 4: TB size = 50 8

Таблица 5: OB size = 50 9

Таблица 6: TB size = 50 9

Постановка задачи

Реализовать метод дискриминантной классификации, а именно Байесовскую процедуру классификации с заменой на состоятельные оценки.

Смоделировать обучающие выборки и тестовую выборку заданных объемов из нормального трехмерного распределения. Найти оценки параметров распределения, построить дискриминантную функцию. Оценить константу C и расстояние Махаланобиса.

Произвести вероятностный анализ ошибочной классификации через построение четырехпольной таблицы сопряженности (матрицы соответствий). Оценить вероятности ошибочной классификации.

Провести анализ данных из репозитория: https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/

Решение

Построим две обучающие выборки – матрицы размером $(n_1, 3)$ и $(n_2, 3)$ – выбранные три признака из таблицы «объект-свойство» из репозитория. ($n_1 = 300$ и $n_2 = 300$)
Будем рассматривать признаки: 2, 3, 8 (номера столбцов). (p = 3)

1. Найдем оценки параметров распределений:
   - Выборочное среднее: $\hat{\mu}_j^{(k)} = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ij}^{(k)}, k = 1, 2$
   - Выборочную матрицу коварияций:

$$S = \frac{1}{n_1 + n_2 - 2}\left[(n_1 - 1)S^{(1)} + (n_2 - 1)S^{(2)}\right]$$

$$S^{(k)} = \left(s_{lj}^{(k)}\right), l, j = \overline{1, p}, k = 1,2$$

$$s_{lj}^{(k)} = \frac{1}{n_k - 1}\sum_{i=1}^{n_k}(x_{il}^{(k)} - \hat{\mu}_l^{(k)})(x_{ij}^{(k)} - \hat{\mu}_j^{(k)}), k = 1,2$$

Заменим вектор параметром дискриминантной функции, $\boldsymbol{\alpha}$, оценкой **a** :

- $\boldsymbol{\alpha} = \Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) \to \hat{\alpha} = \boldsymbol{a} = S^{-1}(\hat{\mu}^{(1)} - \hat{\mu}^{(2)})$

Строим оценки ξ_1 и ξ_2:

- $\xi_1 \to \bar{z}_1 = \frac{1}{n_1}\sum_{i=1}^{n} z_i^{(1)} \qquad z_i^{(1)} = a_1 x_{i1}^{(1)} + \dots + a_p x_{ip}^{(1)}$
- $\xi_2 \to \bar{z}_2 = \frac{1}{n_2}\sum_{i=1}^{n} z_i^{(2)} \qquad z_i^{(2)} = a_1 x_{i1}^{(2)} + \dots + a_p x_{ip}^{(2)}$

Находим константу С:

- $\mathrm{c} = \frac{z1 + z2}{2}$

Таким образом можем произвести соотношение элементов к классам:

- $x \cdot a < c \to x \in D_1$
- $x \cdot a \geq c \to x \in D_2$

2. Оцениваем расстояние Махаланобиса (смешанная и несмешанная):
   - $D^2 = \frac{(\bar{z}_1 - \bar{z}_2)^2}{s_z^2}$
   - $D_H^2 = \frac{n_1 + n_2 - p - 3}{n_1 + n_2 - 2} D^2 - p\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$

3. Строим тестовые выборки объемом по 50.

   Запускаем процедуру классификации данных из тестовой выборки с «обученным» классификатором.

   Строим четырехпольную таблицу сопряженности (матрицу соответствий)
   Находим вероятности ошибочной классификации:

- $\hat{P}(2|1) = \Phi\left(\frac{K - \frac{1}{2}D_{\text{н}}^2}{D_{\text{н}}}\right)$, $\hat{P}(1|2) = \Phi\left(\frac{-K - \frac{1}{2}D_{\text{н}}^2}{D_{\text{н}}}\right)$ $K = \ln\left(\frac{q_2}{q_1}\right)$

  $\Phi()$ – функция распределения Лапласа.

Результаты и выводы

Рассмотрим результаты классификации обучающих выборок (мощность = 300):

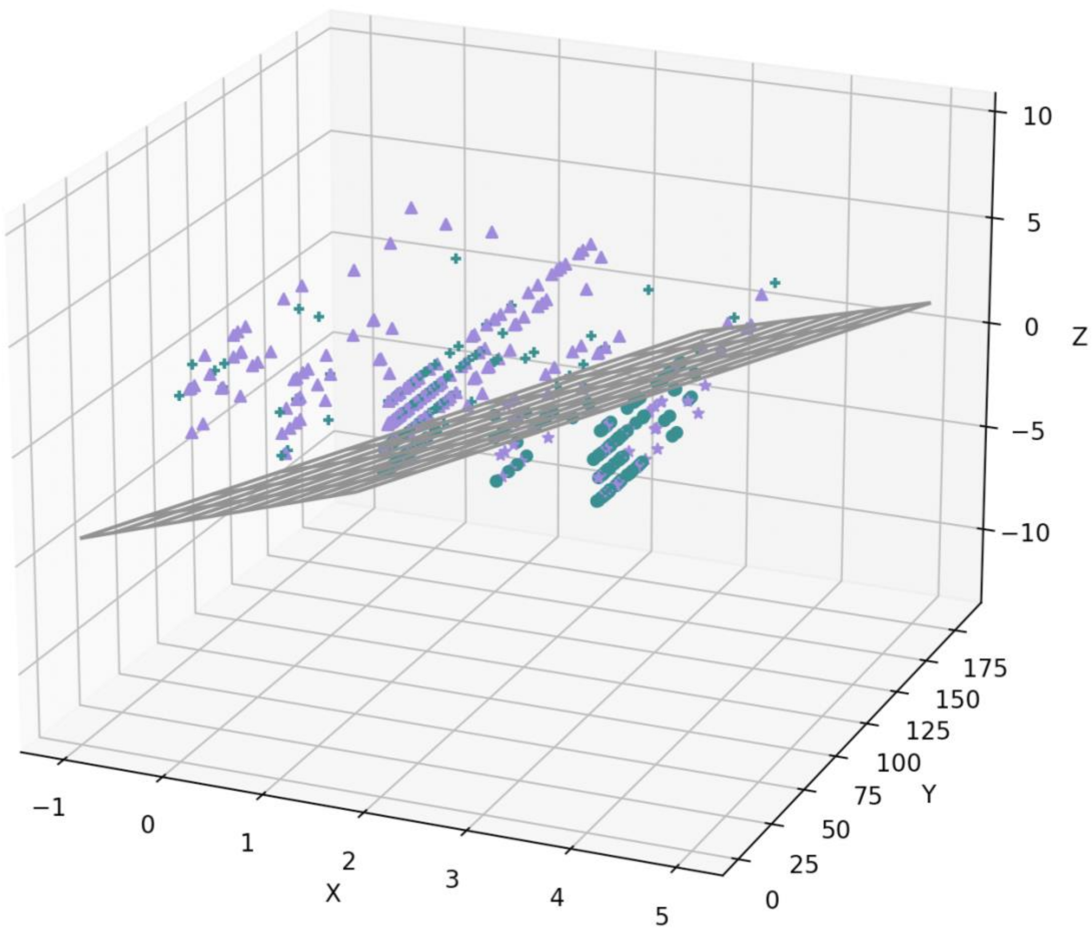

Рисунок 1: OB size = 300

Рассмотрим теперь четырехпольную таблицу сопряженности:

Pred→ True ↓	(зеленая)1	(фиолетовая)2
(зеленая)1	178	122
(фиолетовая)2	109	191

Таблица 1: OB size = 300

Из таблицы и графика видим, что классификатор ошибается, но все же больше правильных соотнесений. Посмотрим на характеристики классификатора:

- Вероятность ошибочной классификации:
  $P(1 \mid 2) = 0.35086$
  $P(2 \mid 1) = 0.308987$

Действительно, для «фиолетовой» выборки классификатор оказался более точным.

- Рассчитаны расстояния Махаланобиса:
  $D^2 = 0.827713$
  $D_H^2 = 0.802176$

Теперь рассмотрим тестовые выборки на этом же критерии (мощность = 50):

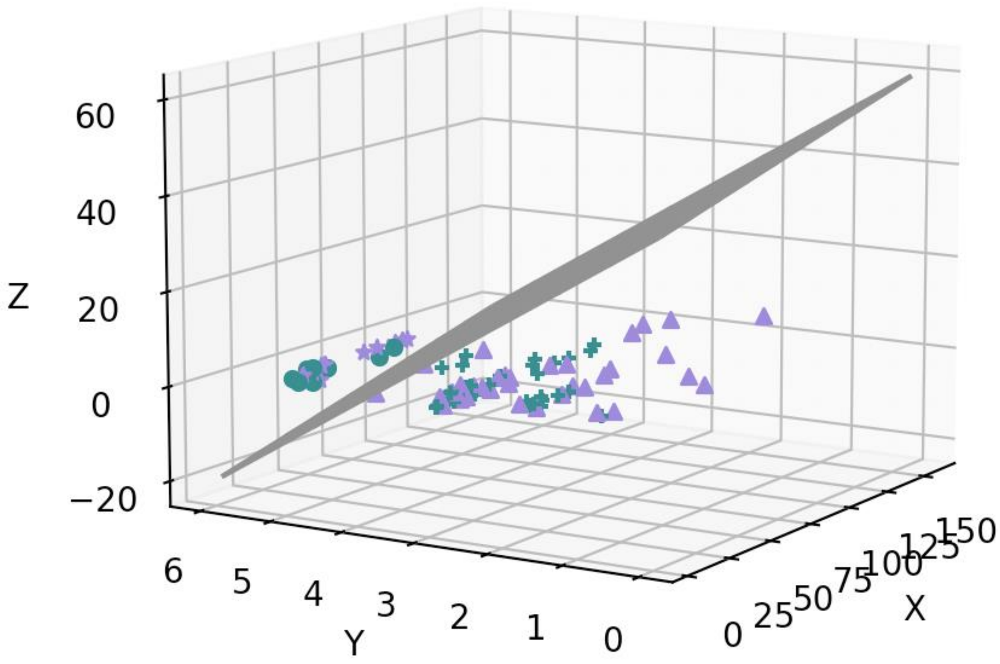

Рисунок 2: TB size = 50

Pred→ True ↓	(зеленая)1	(фиолетовая)2
(зеленая)1	28	22
(фиолетовая)2	19	31

Таблица 2: TB size = 50

Можно заметить, что для тестовой выборки результат примерно ожидаемый, и мы видим похожие соотношения попадания и промаха классификатора для обеих выборок.

Построим классификатор на выборках меньшей мощности по тем же признакам (пусть 100):

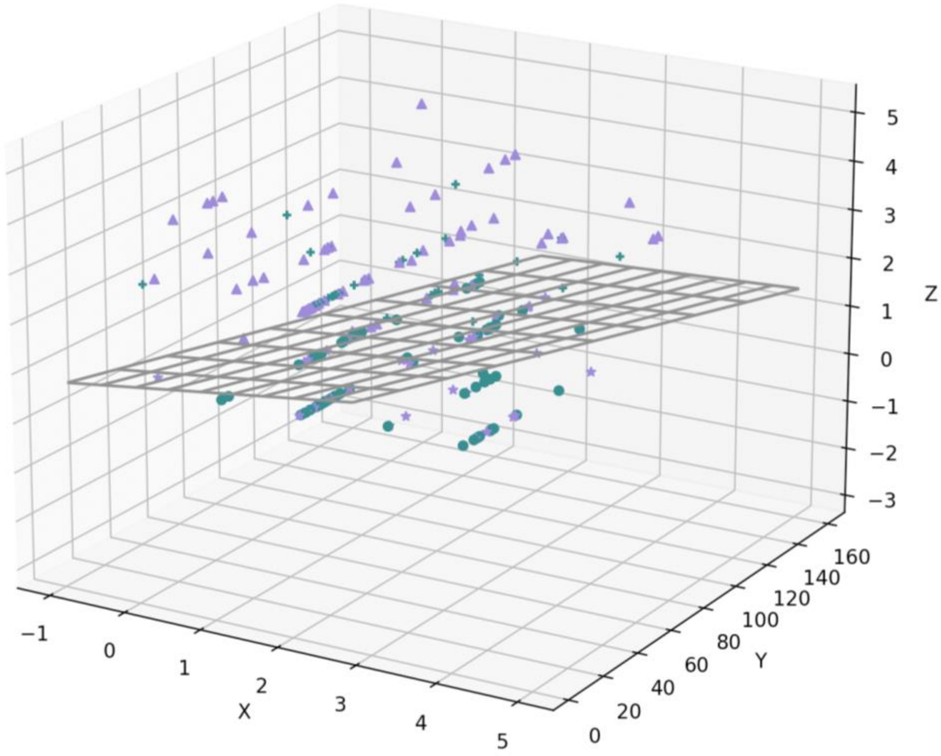

Рисунок 3: OB size = 100

Pred→ True ↓	(зеленая)1	(фиолетовая)2
(зеленая)1	65	35
(фиолетовая)2	35	65

Таблица 3: OB size = 50

Характеристики классификатора:

- Вероятность ошибочной классификации:
  $P(1 \mid 2) = 0.34145$
  $P(2 \mid 1) = 0.34145$

Апостериорная вероятность ошибочной классификации для второй выборки значительно увеличилась.

- Рассчитаны расстояния Махаланобиса:
  $D^2 = 0.813194$
  $D_H^2 = 0.736766$

Теперь посмотрим на ту же тестовую выборку:

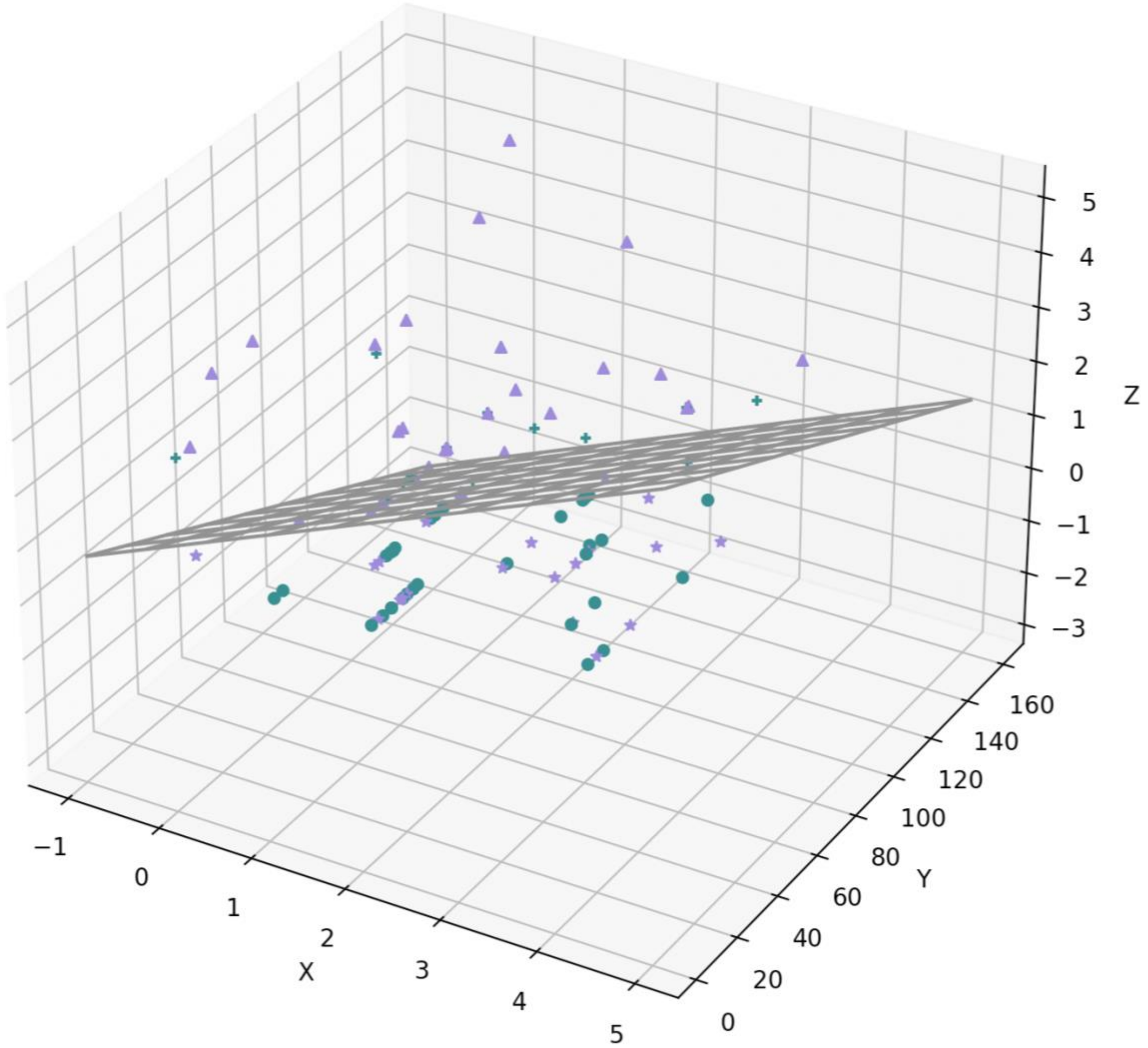

Рисунок 4: TB size = 50

Pred→ True ↓	(зеленая)1	(фиолетовая)2
(зеленая)1	32	18
(фиолетовая)2	20	30

Таблица 4: TB size = 50

Если сравнить данные с предыдущем случаем, то особых ухудшений в классификации не обнаружилось.

Тогда рассмотрим обучающие выборки мощностью 50:

Pred→ True ↓	(зеленая)1	(фиолетовая)2
(зеленая)1	19	31
(фиолетовая)2	11	39

Таблица 5: OB size = 50

Характеристики классификатора:

- Вероятность ошибочной классификации:
  $P(1\,|\,2\,) = \ 0.492568$
  $P(2\,|\,1\,) = \ 0.225548$

Апостериорная вероятность ошибочной классификации имеет большой перекос для двух выборок. Одна определяется значительно лучше другой.

- Рассчитаны расстояния Махаланобиса:
  $D^2 = 0.88058$
  $D_H^2 = 0.724644$

Для тестовой выборки:

Pred→ True ↓	(зеленая)1	(фиолетовая)2
(зеленая)1	19	31
(фиолетовая)2	12	38

Таблица 6: TB size = 50

<u>Вывод:</u> если брать разные столбцы матрицы «объект-свойство», то можно заметить, что при большем расстоянии между облаками данных, тем лучше и равномернее произойдет классификация. В других случаях при слабой разнесенности точек в пространстве, классификатор имеет большую вероятность ошибки. Так же появляются перекосы апостериорных вероятностей для выборок.

Приложение

Код работы: https://github.com/KateZabolotskih/EconomicDecisionMaking