Санкт-Петербургский Политехнический Университет им. Петра Великого

Высшая школа прикладной математики и вычислительной физики

Отчёт по лабораторной работе №2 по дисциплине “Теория принятия экономических решений”

**Сокращение размерности выборки на основе метода PCA**

Выполнил студент:

Мишутин Д. В.

Группа:

3630102/70301

Проверил:

К.ф.-м.н., доцент

Павлова Людмила Владимировна

Санкт-Петербург

2020 г.

Оглавление

[1 Постановка задачи 3](#_Toc56550142)

[2 Численные эксперименты 3](#_Toc56550143)

[2.1 Отладка метода с использованием модельных данных 3](#_Toc56550144)

[2.2 Построение и тестирование классификатора с использованием данных из репозитория 6](#_Toc56550145)

[3 Реализация 8](#_Toc56550146)

[4 Литература 8](#_Toc56550147)

[5 Приложения 8](#_Toc56550148)

# 1 Постановка задачи

фыв

# **2 Численные эксперименты**

## 2.1 Отладка метода с использованием модельных данных

Пусть 1-ый класс представлен обучающей выборкой объёма , размерности , из нормального распределения , а так же вектором класса (состоит из единиц, обозначающих 1-ый класс).

Пусть 2-ой класс представлен обучающей выборкой объёма , размерности , из нормального распределения , а так же вектором класса (состоит из двоек, обозначающих 2-ой класс).

Тогда итоговая выборка будет представлена как “склейка” двух выборок:

- векторы средних, - матрица ковариаций. В нашем случае .

1. После генерации были получены следующие выборочные характеристики:
   1. Для исходных данных:

* Выборочная матрица ковариаций:

Расчётные формулы:

2. Анализ полученных результатов:

Чем дальше друг от друга отстоят выборки (чем меньше расстояния между их средними и меньше дисперсии признаков), тем лучше бинарный классификатор справляется с их разделением, т. е. меньше вероятность ошибочной классификации.

С ростом мощностей выборок эмпирические оценки ошибочной классификации стремятся к теоретическим (вычисленным через функцию Лапласа).

## 2.2 Построение и тестирование классификатора с использованием данных из репозитория

Рассмотрим данные из репозитория (файл *german.data-numeric*). Данные представляют 1000 объектов (1000 резюме различных людей), относящихся к одному из двух классов. Каждый объект характеризует набор из 24 численных признаков, являющихся характеристиками материального, финансового, семейного положения и его трудоустройства.

К первому классу относятся объектов, ко второму - . Проведём разделение каждого класса на ОВ и ТВ в соотношении 75%:25% - 527:223 для ОВ и 173:77 для ТВ.

1. После разделения были получены следующие выборочные характеристики для **ОВ**:

* Выборочная несмещённая дисперсия:

Расчётная формула:

* Смещённая оценка расстояния Махаланобиса:

Расчётная формула:

* Несмещённая оценка расстояния Махаланобиса:

Расчётная формула:

1. Матрица соответствий для **ОВ**:

|  |  |  |
| --- | --- | --- |
| Pred→  True ↓ | 1 | 2 |
| 1 | 473 | 54 |
| 2 | 113 | 110 |

Табличные оценки вероятности ошибочной классификации:

Теоретические оценки вероятности ошибочной классификации:

1. Матрица соответствий для **ТВ**:

|  |  |  |
| --- | --- | --- |
| Pred→  True ↓ | 1 | 2 |
| 1 | 157 | 16 |
| 2 | 35 | 42 |

Табличные оценки вероятности ошибочной классификации:

1. Анализ полученных результатов:

Из оценок видно, что две популяции объектов выборки имеют значительную зону пересечения, из-за чего мы не можем качественно разделить два данных класса (доля ошибок на объектах второго класса близка к 50%).

Возможно лучше применить другой метод классификации (нелинейный) или повысить репрезентативность выборки.

# 3 Реализация

Была использована среда *IPython Notebook* (язык *Python 3.8.2*): модули *numpy* для работы с массивами и математических расчётов, *pandas* - для хранения данных в таблицах, *matplotlib.pyplot* – для построения графиков; *PCA* из модуля *sklearn.decomposition* для стандартизации матрицы наблюдений и *display* из модуля *IPython.display* для отображения pandas’ских таблиц в браузере.

# 4 Литература

[Основы работы с *numpy* (отдельная глава курса)](https://stepik.org/course/401)

Лекции по ТПЭР, “Метод PCA”, Павлова Л. В., 2020 г.

# 5 Приложения

[Код лабораторной](https://github.com/MeShootIn/economics-decision-theory/blob/main/lab_2/main.ipynb)