Санкт-Петербургский Политехнический Университет им. Петра Великого

Институт прикладной математики и механики

Кафедра прикладной математики

**Отчёт по курсовой работе по дисциплине “Математическая статистика”**

Выполнил студент:

Мишутин Д. В.

Группа:

3630102/70301

Проверил:

К.ф.-м.н., доцент

Баженов Александр Николаевич

Санкт-Петербург

2020 г.

Оглавление

[1 Постановка задачи 3](#_Toc51970852)

[2 Теория 3](#_Toc51970853)

[3 Реализация 4](#_Toc51970854)

[4 Результаты 4](#_Toc51970855)

[5 Выводы 7](#_Toc51970856)

[6 Литература 7](#_Toc51970857)

[7 Приложения 7](#_Toc51970858)

Список иллюстраций и таблиц

[Таблица 1 Интенсивности………………………………………………………………………………………………………………..…](#таблица_1)4

[Рис. 1 Пример работы на файле 1.1\_70..txt………………………………………………………………………………………..](#рис_1)5

[Таблица 2 Значения K для Африки……………………………………………………………………………………………….……...](#таблица_2)5

[Таблица 3 Значения K для Русского Севера………………………………………………………………………………………….](#таблица_3)6

[Таблица 4 Коэффициенты корреляции………………………………………………………………………………………….……](#таблица_4)6

[*Рис. 2 Боксплот для значений m2………………………………………………………………………………………………………..*](#рис_2)*7*

# 1 Постановка задачи

Есть набор 2D данных в текстовом формате – следы жизни в геологических объектах. Образцы взяты с двух разных регионов:

* Русского Севера
* Центральной Африки

На объект подавалось излучение от ближнего ультрафиолетового до видимого. Длина волны – первая переменная . Когда свет с заданной попадал на объект, его поглощали молекулы и в свою очередь, излучали свет с длинами волны примерно в том же диапазоне. То, что они излучали записывается в виде графика .

Известна область для каждой аминокислоты в координатах . Пики на графике можно идентифицировать с излучением протеиногенных аминокислот, т. е. это остатки органической жизни.

Задача: для классификации двух типов данных предложен параметр , который позволяет достаточно уверенно проводить разделение этих типов. При этом не используются данные по переменной . Необходимо ввести параметры и , построить корреляции с и и проанализировать полученные результаты.

# **2 Теория**

* Дополнительные параметры:

|  |  |
| --- | --- |
|  | (2.1) |

|  |  |
| --- | --- |
|  | (2.2) |

* Параметр для разделения регионов:

|  |  |
| --- | --- |
|  | (2.3) |

* Коэффициент корреляции Пирсона:

|  |  |
| --- | --- |
|  | (2.4) |

Аминокислоты и их области интенсивности:

Таблица 1 Интенсивности

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Тип компонента | Буквенное обозначение | Цвет зоны |
| 320-350 | 420-480 | Humic-like | C | Красный |
| 250-260 | 380-480 | Humic-like | A | Зелёный |
| 310-320 | 380-420 | Marine Humic-like | M | Синий |
| 270-280 | 300-320 | Tyrosine-like, Protein-like | B | Жёлтый |
| 270-280 | 320-350 | Tryptophane-like, Protein-like or phenol-like | T | Белый |

# 3 Реализация

Был использован язык *Python 3.8.2*: модуль *numpy* для работы с массивами, функция *pearsonr* из модуля *scipy.stats* для вычисления коэффициента корреляции Пирсона, модуль *matplotlib* для построения графиков, модуль *pandas* для оптимального хранения статистических данных и функция *display* из модуля *IPython.display* для их корректного отображения в таблицах.

# 4 Результаты

По данным из каждого файла строится график. Чтобы убрать лучи рэлеевского рассеяния, строится график функции нормы градиента и обрабатывается фильтром Савицкого – Голея. Затем на исходном графике затемняются точки наивысшей интенсивности с графика градиента, выделяются области аминокислот, а также отмечаются точки локальных максимумов для наглядности.

Рис. 1 Пример работы на файле 1.1\_70..txt

Изображение выглядит как темный, сидит, монитор, экран

Автоматически созданное описание

Далее для каждого файла определённого региона высчитывались значения параметра и коэффициенты корреляции с и для регионов.

Таблица 2 Значения K для Африки

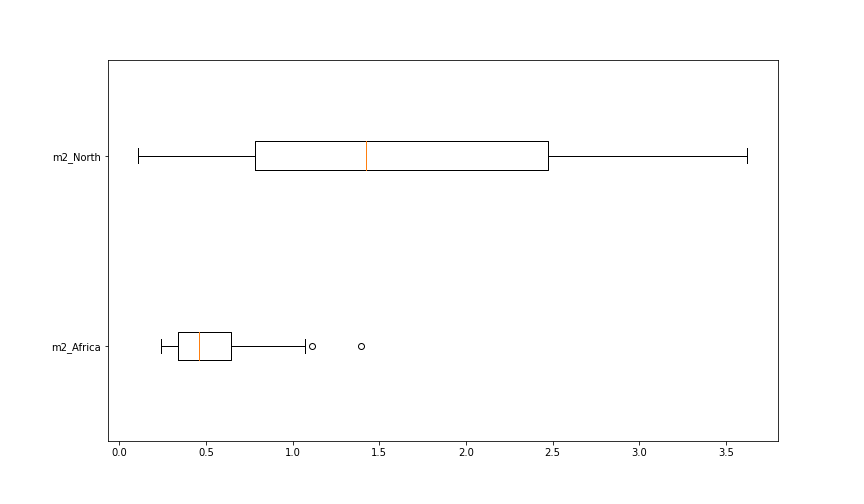
|  |  |  |
| --- | --- | --- |
| Африка |  |  |
| **1.1\_70..txt** | 1.463542 | 0.273376 |
| **1.2\_21.txt** | 1.922796 | 0.333485 |
| **1.3\_68.txt** | 2.679532 | 0.499371 |
| **1.4\_114.txt** | 1.708243 | 0.327983 |
| **1.5\_11.txt** | 2.143091 | 0.370068 |
| **1.6\_37.txt** | 1.948490 | 0.336029 |
| **2.3\_5 (400).txt** | 2.622960 | 0.643876 |
| **2.3\_5 (600).txt** | 1.708243 | 0.327983 |
| **2.3\_5.txt** | 2.462687 | 0.638975 |
| **2.4\_7.txt** | 2.274053 | 0.402768 |
| **3.1\_14.txt** | 2.069670 | 0.389189 |
| **3.2\_69.txt** | 2.590766 | 0.483579 |
| **3.3\_15 (600).txt** | 3.306511 | 0.976156 |
| **3.4\_20(800).txt** | 3.136010 | 1.109579 |
| **3.4\_20.txt** | 3.900566 | 1.396637 |
| **3.5\_43.txt** | 3.028176 | 0.488321 |
| **3.6\_49(800).txt** | 2.663941 | 0.445824 |
| **4.1\_45.txt** | 2.191567 | 0.404147 |
| **4.2\_80.txt** | 1.907667 | 0.327841 |
| **4.3\_84.txt** | 2.343400 | 0.431579 |
| **4.4\_87.txt** | 2.131170 | 0.473307 |
| **4.5\_108.txt** | 1.312571 | 0.241106 |
| **4.6\_88.txt** | 1.934876 | 0.357270 |
| **5.1\_90.txt** | 1.790424 | 0.331992 |
| **5.2\_2.txt** | 3.801942 | 0.677048 |
| **5.3\_66.txt** | 4.128717 | 0.916844 |
| **5.3\_66-1.txt** | 4.831961 | 1.073949 |
| **5.4\_92.txt** | 4.694208 | 0.849342 |
| **5.5\_28.txt** | 3.368572 | 0.576612 |
| **5.6\_95.txt** | 3.637692 | 0.631947 |

Таблица 3 Значения K для Русского Севера

|  |  |  |
| --- | --- | --- |
| Русский Север |  |  |
| **1701.txt** | 17.182651 | 2.474726 |
| **1702.txt** | 23.305749 | 3.496403 |
| **1702old.txt** | 17.945509 | 3.034029 |
| **1704.txt** | 17.523280 | 2.461407 |
| **1706.txt** | 5.399792 | 0.789786 |
| **1708\_1to10.txt** | 23.901524 | 2.366542 |
| **1708\_1to20.txt** | 4.508159 | 0.668021 |
| **1711.txt** | 11.189969 | 1.737484 |
| **1712.txt** | 0.587783 | 0.111257 |
| **1727.txt** | 5.187718 | 0.779121 |
| **1728.txt** | 21.693521 | 3.410823 |
| **1728old.txt** | 19.908014 | 3.622415 |
| **1729.txt** | 13.186196 | 1.888211 |
| **1730.txt** | 2.515725 | 0.363079 |
| **1732.txt** | 7.537229 | 1.115526 |
| **1733.txt** | 4.709272 | 0.749295 |
| **1733old.txt** | 4.621420 | 0.818236 |
| **1734.txt** | 5.311251 | 0.823655 |

Таблица 4 Коэффициенты корреляции

|  |  |  |
| --- | --- | --- |
| Корреляция |  |  |
| Африка | 0.321127 | 0.827296 |
| Русский Север | -0.337818 | 0.955279 |

Рис. 2 Боксплот для значений m2

# 5 Выводы

По таблице 4 видно, что для Африки и Русского Севера корреляция между и слабая. Зато, при корреляции с , у обоих регионов коэффициенты близки к 1. Т. е. параметр имеет сильную линейную связь с (можно линейно выразить один параметр через другой, имея малую погрешность), а значит его так же можно использовать для классификации.

Опираясь на боксплот, можно выбрать число 0.66 для разделения Африки и Русского Севера по параметру , имея точности 76.66% и 88.88% соответственно (самый оптимальный вариант).

# 6 Литература

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

[Pandas обзор](https://www.dataquest.io/blog/pandas-cheat-sheet/)

[Документация по *scipy*](https://docs.scipy.org/doc/scipy/reference/index.html)

# 7 Приложения

[Код лабораторной](https://github.com/MeShootIn/matstat/blob/master/lab_8/lab_8.ipynb)