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

Список иллюстраций и таблиц

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 x_1. Когда свет с заданной x_1 попадал на объект, его поглощали молекулы и в свою очередь, излучали свет с длинами волны x_2 примерно в том же диапазоне. То, что они излучали записывается в виде графика $I(x_1, x_2)$.

Известна область для каждой аминокислоты в координатах (x_1, x_2). Пики на графике I можно идентифицировать с излучением протеиногенных аминокислот, т. е. это остатки органической жизни.

Задача: для классификации двух типов данных предложен параметр K, который позволяет достаточно уверенно проводить разделение этих типов. При этом не используются данные по переменной M. Необходимо ввести параметры m_1 и m_2, построить корреляции K с m_1 и m_2 и проанализировать полученные результаты.

2 Теория

- Дополнительные параметры:

$$m_1 = \frac{M}{C + A} \tag{2.1}$$

$$m_2 = \frac{M}{B + T} \tag{2.2}$$

- Параметр для разделения регионов:

$$K = \frac{A + C}{B + T} = \frac{m_2}{m_1} \tag{2.3}$$

- Коэффициент корреляции Пирсона:

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})\,(y_i - \overline{y})}{n\sigma_X\sigma_Y} \tag{2.4}$$

Аминокислоты и их области интенсивности:

Таблица 5 Интенсивности

$E_{x_{max}}(nm)$	$E_{m_{max}}(nm)$	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

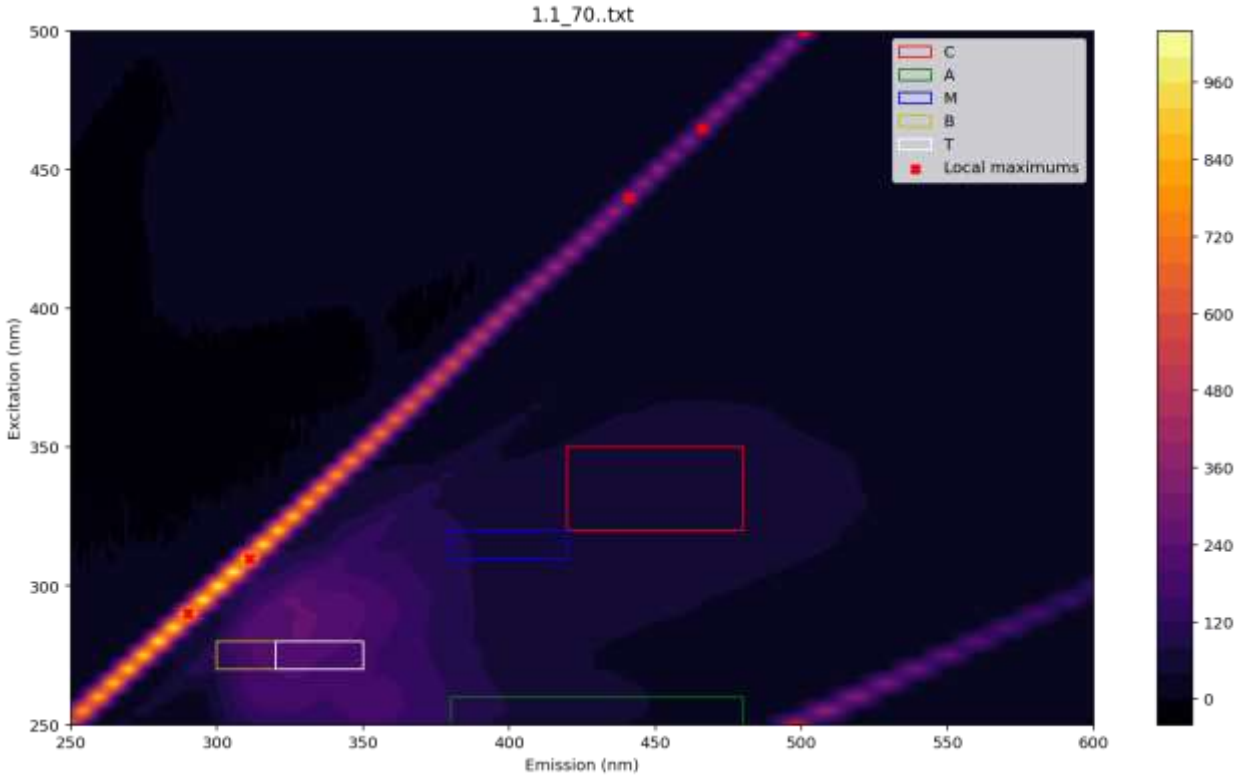
3 Реализация

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. 3 Пример работы на файле 1.1_70..txt



Далее для каждого файла определённого региона высчитывались значения параметра K и коэффициенты корреляции с m_1 и m_2 для регионов.

Таблица 6 Значения K для Африки

Африка	K	m_2
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

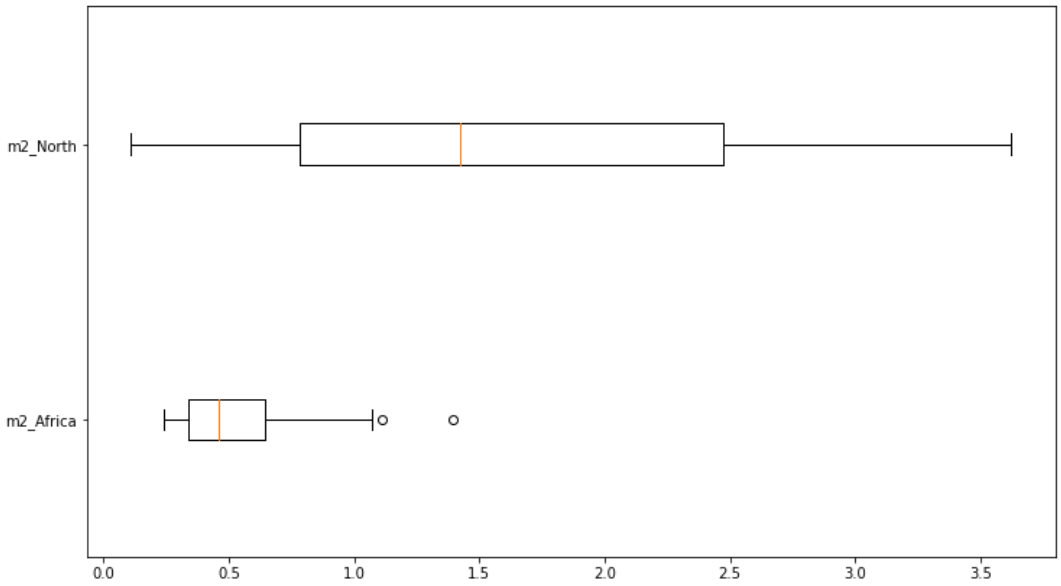
Таблица 7 Значения K для Русского Севера

Русский Север	K	m_2
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 8 Коэффициенты корреляции

Корреляция	K, m_1	K, m_2
Африка	0.321127	0.827296
Русский Север	-0.337818	0.955279

Рис. 4 Боксплот для значений m2



5 Выводы

По таблице 4 видно, что для Африки и Русского Севера корреляция между K и m_1 слабая. Зато, при корреляции с m_2, у обоих регионов коэффициенты близки к 1. Т. е. параметр m_2 имеет сильную линейную связь с K (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 m_2, имея точности 76.66% и 88.88% соответственно (самый оптимальный вариант).

6 Литература

Основы работы с *numpy* (отдельная глава курса)

Pandas обзор

Документация по *scipy*

7 Приложения

Код лабораторной