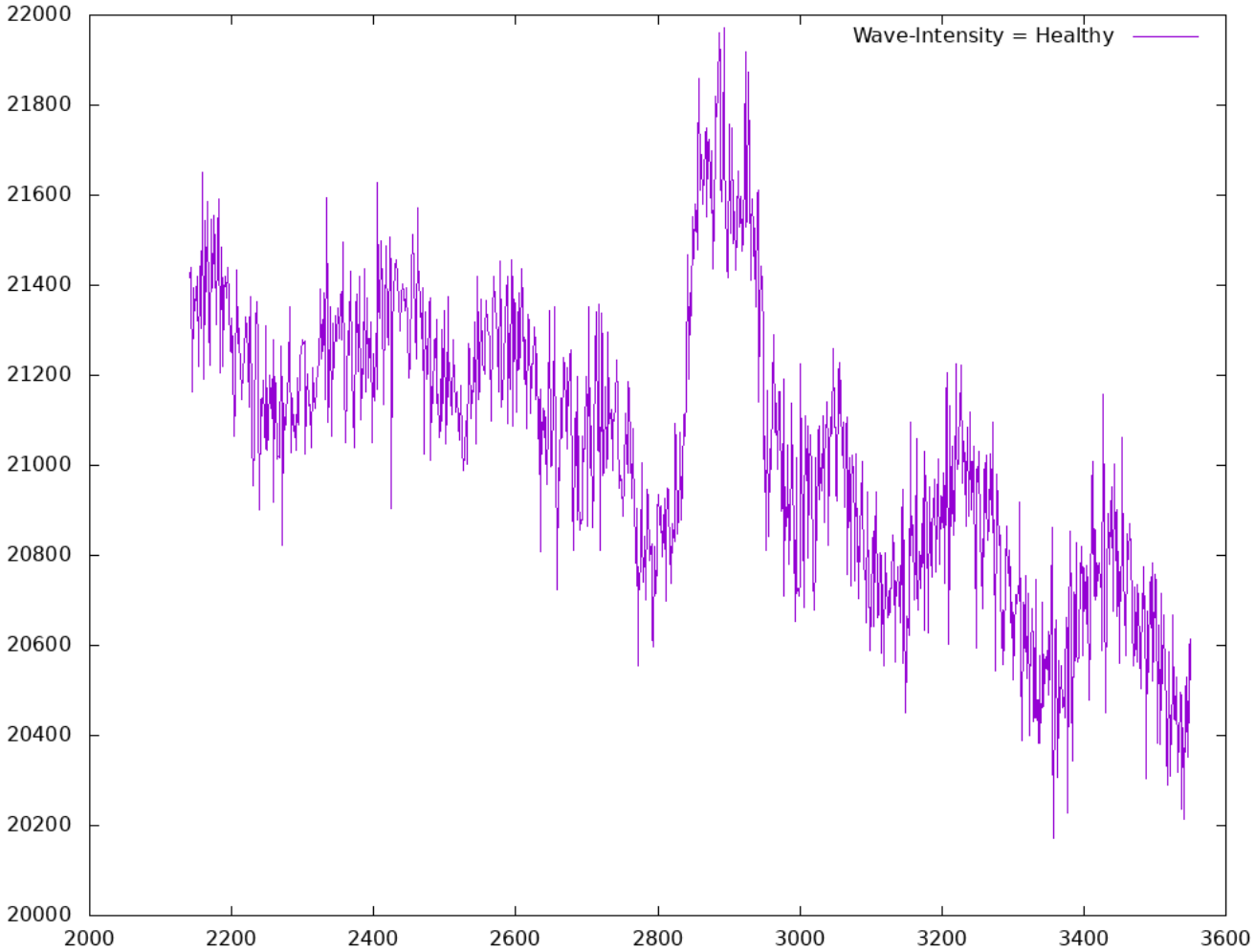
Классификация опухолевых клеток с использованием моделей машинного обучения в среде Альт Линукс

Александр Крагин, Москва (НИУ ВШЭ)
Илья Обрубов,, Москва (НИУ ВШЭ)
Воронин Игорь Шатура (ИПЛИТ РАН)

OSEDUCONF-2023

Описание предметной области

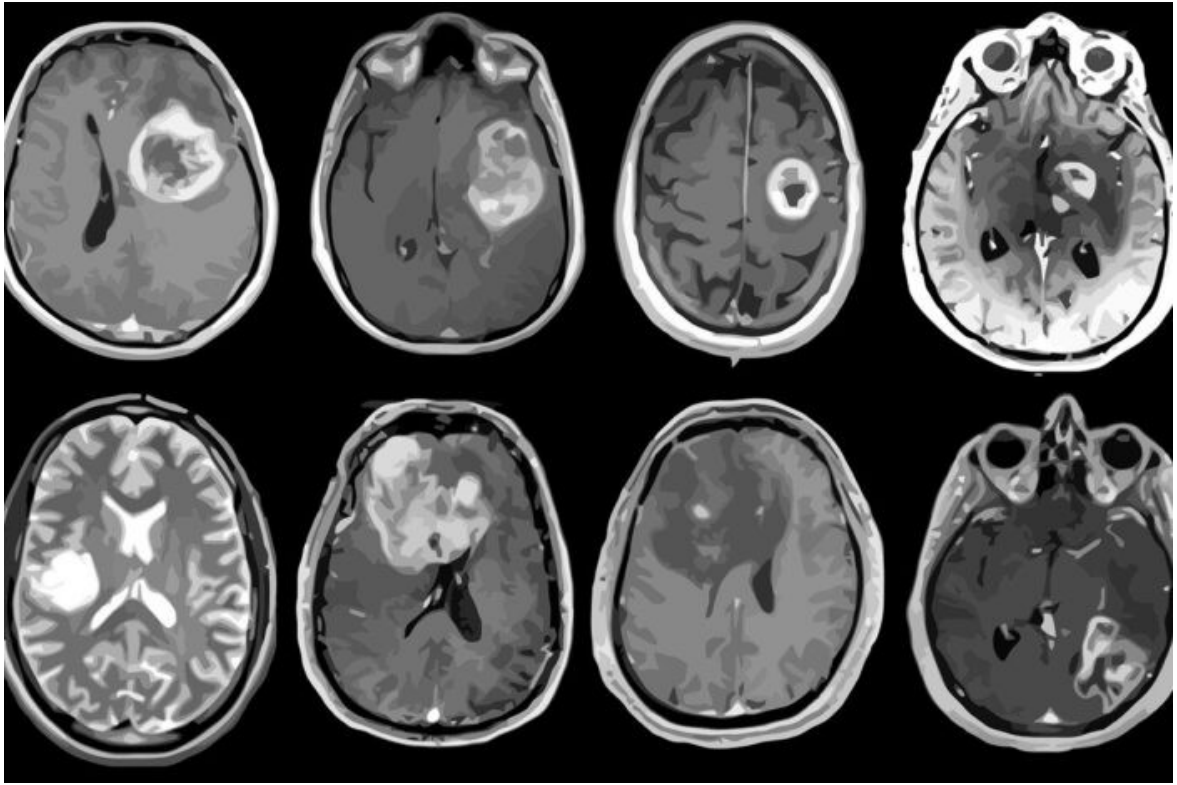
Спектроскопия комбинационного рассеяния (также известная как рамановская) - метод, который успешно используется в химии для получения структурных «отпечатков пальцев», по которым можно идентифицировать молекулы в образце. В результате получается спектрограмма, по которой можно делать выводы о состоянии ткани.


Wave-Intensity = Healthy
22000
21800
21600
21400
21200
21000
20800
20600
20400
20200
20000
2000
2200
2400
2600
2800
3000
3200
3400
3600

Актуальность

Для успешного лечения критически важно определить границы опухоли.

В связи с этим необходим точный метод классификации тканей на больные и здоровые.



Постановка задачи и данные

Будем рассматривать задачу бинарной классификации с классами “больные” и “здоровые”. На данный момент ведётся работа с 888 спектрограммами, из которых 456 принадлежат классу “больные” и 432 принадлежат классу “здоровые”.

В рамках решения задачи каждая спектрограмма будет представлена массивом, содержащим значение интенсивности отражения для каждой длины волны.

Используемые инструменты

K-Nearest Neighbors

Один из самых простых методов классификации в машинном обучении:

1) Разместим все объекты обучающей выборки в N-мерном пространстве.
2) При получении нового неизвестного объекта X, определим множество C, состоящее из K ближайших соседей объекта в пространстве.
3) Отнесём объект X к тому классу, который чаще всего встречается в множестве C.

Logistic Regression

Классическая линейная модель для классификации:

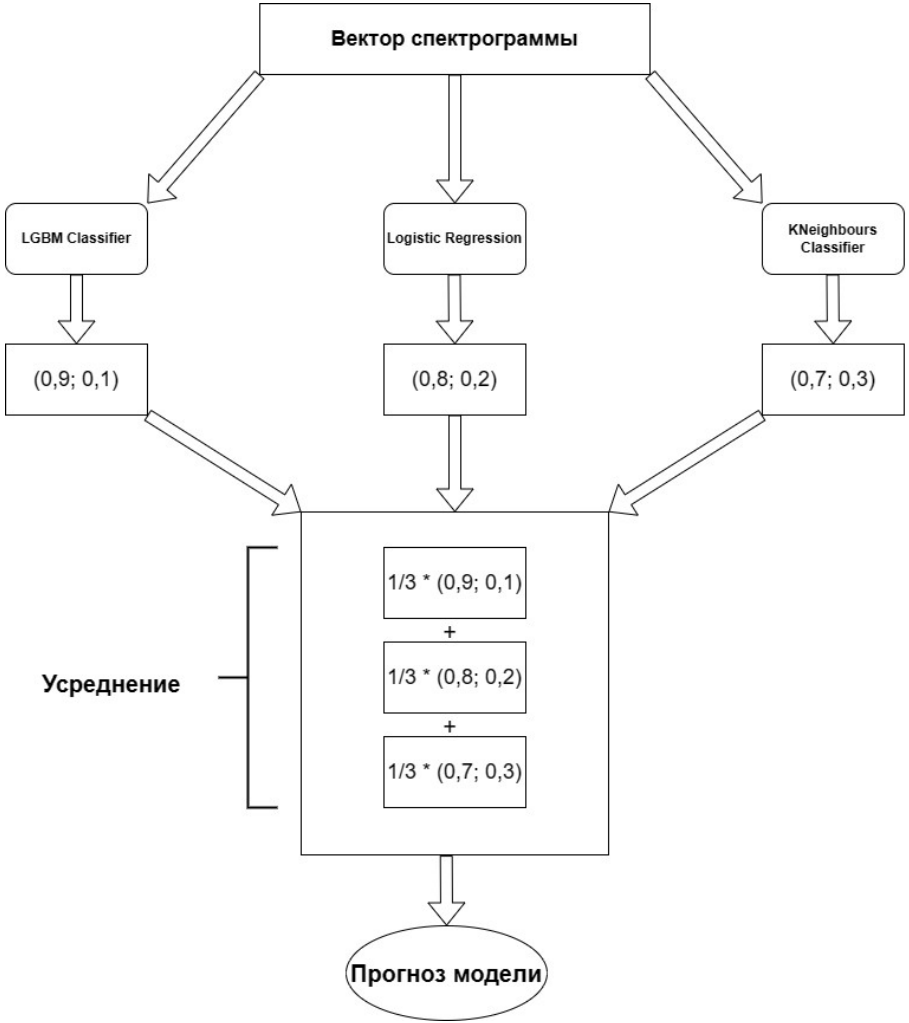
1) Логистическая функция $1/(1 + e^{-x})$применяется к обычной модели линейной регрессии.
2) Получившаяся модель обучается чтобы предсказывать вероятности классов.
3) Поиск оптимального набора весов осуществляется с помощью градиентного спуска.
4) Для предсказания класса нового объекта достаточно посчитать вероятности по имеющейся формуле с найденными весами.

LightGBM

Эталонный фреймворк для машинного обучения, реализует наиболее сложную модель из перечисленных - градиентный бустинг над решающими деревьями.

1) Строится первое решающее дерево, обученное на обучающей выборке.
2) Пусть A_n - композиция уже построенных деревьев. Обучим следующее так, чтобы оно частично компенсировало ошибку A_n.
3) С каждым новым деревом средняя ошибка модели уменьшается.

Архитектура модели



Формат исходных данных, результат

	#X	#Y	#Wave	#Intensity
2	-38876.403657	14875.070833	3550.405273	29680.113281
3	-38876.403657	14875.070833	3549.173828	29676.039063
4	-38876.403657	14875.070833	3547.941406	29791.939453
5	-38876.403657	14875.070833	3546.708984	30056.468750
6	-38876.403657	14875.070833	3545.476563	29929.908203
7	-38876.403657	14875.070833	3544.243164	29793.218750
8	-38876.403657	14875.070833	3543.010742	29770.570313
9	-38876.403657	14875.070833	3541.777344	29874.099609
10	-38876.403657	14875.070833	3540.542969	29955.201172
11	-38876.403657	14875.070833	3539.309570	30103.390625
12	-38876.403657	14875.070833	3538.075195	29706.476563
13	-38876.403657	14875.070833	3536.840820	29990.878906
14	-38876.403657	14875.070833	3535.605469	29872.654297
15	-38876.403657	14875.070833	3534.371094	29929.261719
16	-38876.403657	14875.070833	3533.135742	29967.554688
17	-38876.403657	14875.070833	3531.900391	29936.755859
18	-38876.403657	14875.070833	3530.664063	29645.986328
19	-38876.403657	14875.070833	3529.428711	29907.677734
20	-38876.403657	14875.070833	3528.192383	29690.119141
21	-38876.403657	14875.070833	3526.955078	29769.001953
22	-38876.403657	14875.070833	3525.718750	29995.990234
23	-38876.403657	14875.070833	3524.481445	29731.876953
24	-38876.403657	14875.070833	3523.244141	29893.857422
25	-38876.403657	14875.070833	3522.006836	30011.148438
26	-38876.403657	14875.070833	3520.768555	29893.181641
27	-38876.403657	14875.070833	3519.530273	29927.300781
28	-38876.403657	14875.070833	3518.291992	29791.154297
29	-38876.403657	14875.070833	3517.053711	29997.509766
30	-38876.403657	14875.070833	3515.814453	29968.771484
31	-38876.403657	14875.070833	3514.575195	30085.873047

Out[64]: <catboost.core.CatBoostClassifier at 0x7fe910206b50>

Предсказания для переданных данных:

0 - здоровый, 1 - больной

Ввод [65]:
```
predictions = model.predict(X)
print(predictions)
```

```
[0 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 1 0 1 0 1 0 1 1 0 1 1 1 1 1
 1 1 1 1 1 0 1 1 1 1 1 1 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 0 1 1 0
 1 1 1 1 1 1 1 1 0 0 1 1 1 0 1 1 1 1 1 1 1 1 0 1 1 0 0 0 1 1 1 1 1 1 1 0 1
 1 0 0 0 1 0 1 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 0 1 1 1 1 1 1 1 0 0 1 1 0 1 1 1 1
 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 1 0 1 0 0 1 1 1 1 1 1
 1 1 1 0 1 1 0 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 1 0 1 0 0 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1
 1 1 1 0 1 0 0 0 1 0 0 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 0 1 1 1 0 1 1 1 1 1 1 0 0 1
 1 1 1 1 1 1 1 1 0 0 0 1 1 1 1 1 1 1 1 1 1 1 0 1 1 0 0 0 1 1 1 1 1 1 1 1 1 0
 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 0 0 1 1 1 1 1 0 0 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 0
 0 0 1 1 1 1 0 1 1 1 1 1 1 0 1 1 1 1 1 1 0 1 1 1]
```

Вероятности для каждого предсказания

Ввод [66]:
```
predictions_probs = model.predict_proba(X)
print(predictions_probs[:20])
```

```
[[0.74706575 0.25293425]
 [0.13466442 0.86533558]
 [0.01250987 0.98749013]
 [0.03758775 0.96241225]
 [0.02934774 0.97065226]
 [0.70568201 0.29431799]
 [0.31611231 0.68388769]
```

Результаты

- На кросс-валидации точность модели составила 97.6%
- За счёт использования универсальных фреймворков и языка Python, модель может быть использована на операционной системе Alt Linux p10.
- http://astera.laser.ru/umkiurban/