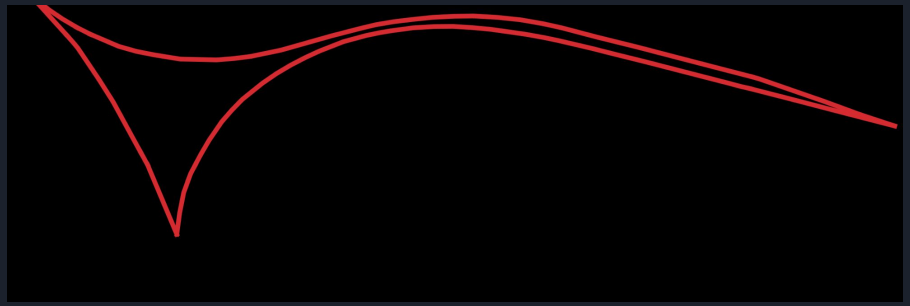
Поиск зон расширений и сужений в сложных фигурах

Курсовая работа на курсе ML Basic 04.2024

Приступин Константин

Задачи и цели работы

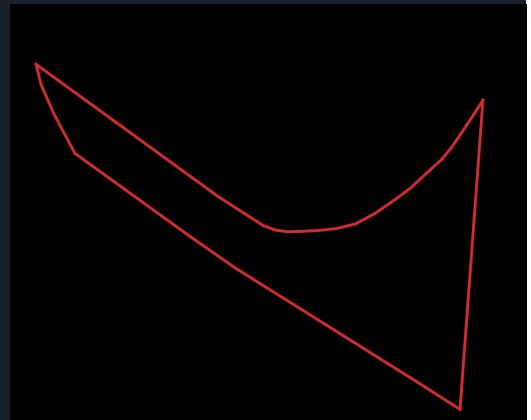
Задачей данного проекта является детекция точек на фигурах с нестабильной шириной, в которых можно разметить границы зон изменения состояния производной от ширины фигуры в данной точке. Шириной фигуры в точке будем считать расстояние между точкой на рассматриваемой линии и точкой пересечения перпендикуляра к линии в этой точке с другой линией (при наличии такого пересечения).

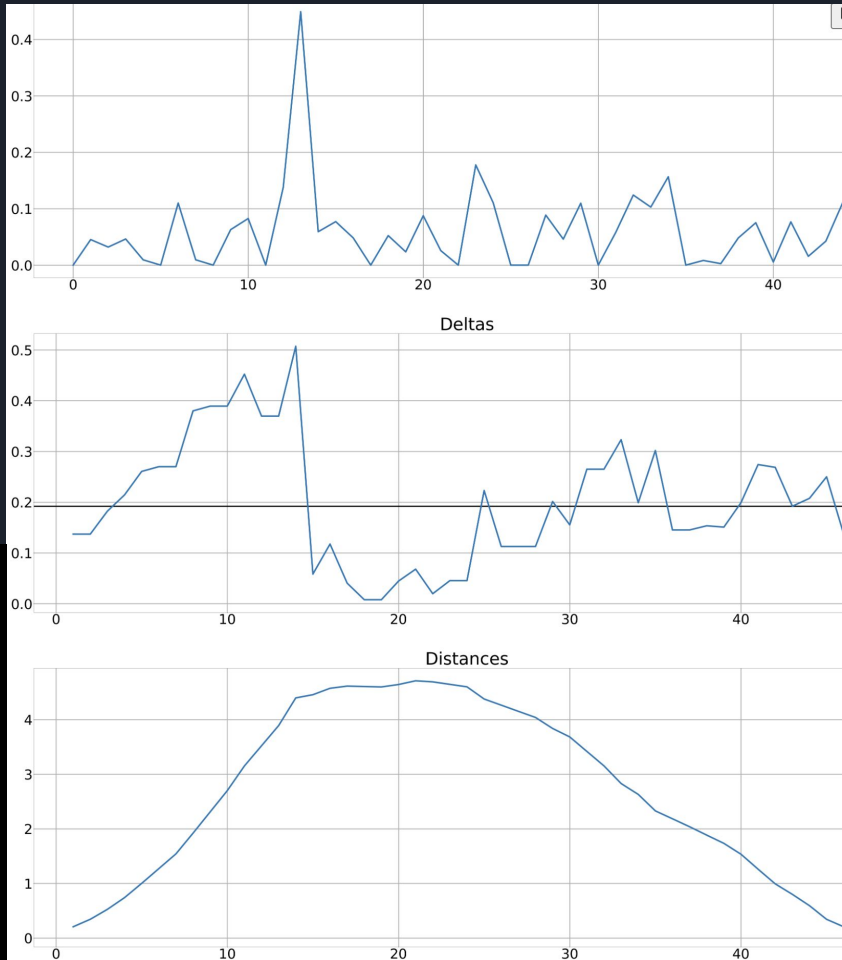


Пример рассматриваемой фигуры.

Задачи и цели работы

Проблемой алгоритмического решения, опирающегося на анализ изменения знака производной от ширины в данной точке, является то, что ширина в рассматриваемых фигурах изменяется хаотично. На графике справа, для примера, приведены графики, построенные для фигуры ниже, на графиках (снизу вверх: ширина в точке, производная 1-го порядка, производная 2-го порядка).





Задачи и цели работы

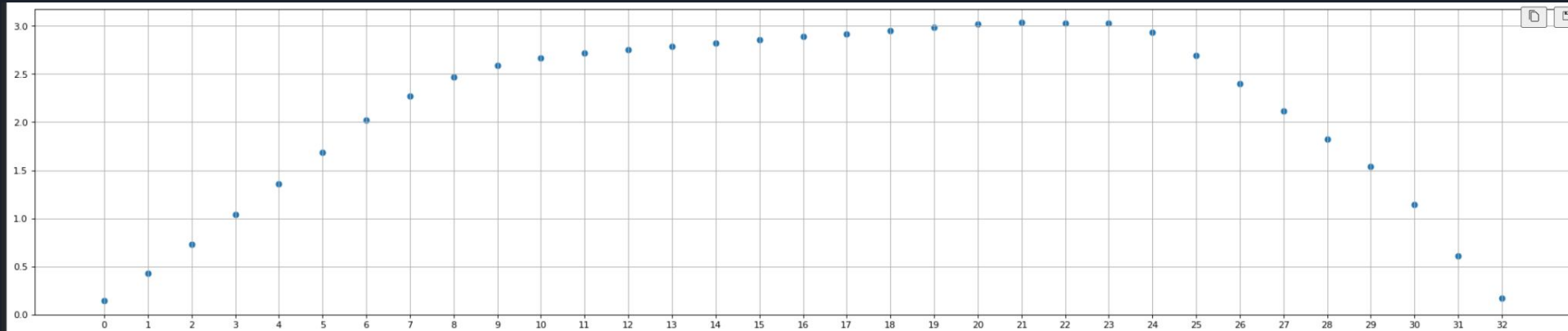
Цель проекта - выделение признаков в точке достаточных для создания датасета для эффективного использования градиентного бустинга для обучения модели, которая сможет дать оценку вероятности того, что данная точка может быть границей условной зоны изменения состояния производной ширины данной фигуры.

Датасет

Итоговый датасет содержит следующие поля:

- ширина фигуры в точке
- производная 1-го порядка от ширины в точке
- производная 2-го порядка от ширины в точке
- разница ширины в точке и медианного значения
- количество точек до ближайшего изменения знака производной 1-го порядка
- целевое значение (может ли точка быть граничной)

Все эти данные созданы на основе замеров ширины фигуры в точке по всей длине опорной линии (пример ниже)



Пример датасета

	dist	delta	delta_delta	dif_from_median	min_distance_from_signchanges	should_cut
0	0.821307	0.211402	0.000000e+00	0.181989	13	False
1	1.097101	0.275794	6.439259e-02	0.246381	12	False
2	1.386558	0.289456	1.366181e-02	0.260043	11	False
3	1.676014	0.289456	1.000000e-09	0.260043	10	False
4	1.944390	0.268376	2.108001e-02	0.238963	9	False
...	...	...	...	...	...	...
17	1.751540	0.047543	2.958920e-02	0.000869	0	True
18	2.635020	0.294496	2.469531e-01	0.246084	1	False
19	4.090320	0.485099	1.906030e-01	0.436687	2	False
20	7.193710	1.034460	5.493610e-01	0.986048	3	False
21	14.016300	2.274180	1.239720e+00	2.225768	4	False

Выбор модели обучения

Для обучения модели был выбран градиентный бустинг, как наиболее эффективный алгоритм среди ансамблевых моделей.

В полученных данных фактически отсутствуют линейные зависимости, а корреляция между данными из разных колонок будет высокой (и фактической пользы не принесет) по причине того, что все данные основываются на значениях первой колонки, поэтому выбор модели обучения в пользу использования ансамбля модели является обоснованным.

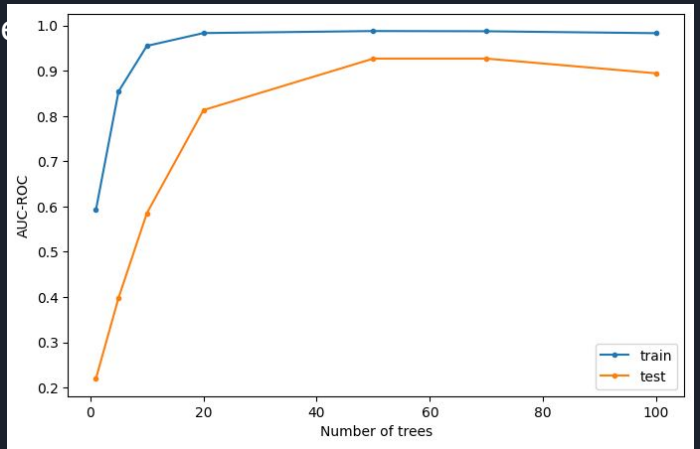
Для реализации модели была выбрана библиотека с открытым исходным кодом CatBoost.

Для разбиения датасета, построения метрик и нормализации данных был использован sclearn.

Результат и вывод

Оптимальное количество моделей в ансамблей было подобрано вручную “по сетке” на основании анализа AUC-ROC кривых для тренировочного и проверочного датасета. Оптимальное количество моделей - 50.

Точность в тренировочном сете, стремящаяся к 1 объясняется недостаточным размером итогового датасета (< 1000 строк, что связано с недостатком данных для обучения.) и подавляющим большинством одного ожидаемого результата (точка не является границей) в датасете из-за специфики решаемой задачи, который предсказывается всегда верно. В ходе дальнейшей работы данных для обучения будет становиться больше. Однако, даже при таком скромном объеме датасета около трети точек, которые могут быть границами зон изменения широты , размечены ве



Спасибо за внимание