

Решение задачи классификации.

Лукьяненко Иван Андреевич, Б05-906

Студент МФТИ ФПМИ 2 курс

Собеседование на кафедре ИАД.

Постановка задачи.

- **Задача:** классификация;
- **Данные:** синтетическая выборка и <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer>;
- **Используемые модели:** логистическая регрессия, нейронная сеть, градиентный бустинг;
- **Структурные параметры:** состав признаков, структура модели, количество параметров модели;
- **Критерии качества:** ROC AUC, PR кривая, сложность модели (ввести определение).

- Выборка: <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer>
- Количество объектов: 286 штук.
- Распределение по меткам класса: 201 объект принадлежит к одному классу, 85 к другому.
- Количество признаков: 9 признаков.

Стоит отметить, что все признаки категориальные либо порядковые.

Кодирование категориальных признаков с помощью:

- One-Hot-Encoding

Вместо одного признака создаются n признаков, когда исходный признак для объекта был равен i -ому уникальному значению, i -ый признак из новосозданных равен 1, остальные $n-1$ новосозданных признаков заполняются нулями.

- Ordinal-Encoding

Нумеруем уникальные значения признака, и отображаем значение признаков в отрезок $[1; n]$

Данные: предобработка категориальных признаков

Всего имеем 5 категориальных признаков:

menopause, node-caps, breast, breast-quad, irradiat.

Данные категориальные признаки мы будем кодировать с помощью One-Hot-Encoding.

Корреляция признаков: breast и breast-quad.

Данные: предобработка порядковые признаки

Всего имеем 4 порядковых признака:

age, tumor-size, inv-nodes, deg-malig.

Данные категориальные признаки мы будем кодировать с помощью Ordinal-encoding

Сохранение порядка.

Сравнение моделей:

Будем сравнивать модели:

- Логистическая регрессия
- Градиентный бустинг
- Нейронная сеть

Сравнение моделей: логистическая регрессия

- Логистическая регрессия

Случай бинарной классификации. Метки класса $Y = \{-1, 1\}$
Поиск вектора параметров w , такого что функция $a(x, w) : X \rightarrow Y$,
где X - пространство признаков.

Функция потерь: $Q(w) = \sum_{i=1}^m \ln(1 + \exp(-y_i \langle x_i; w \rangle)) \rightarrow \min$

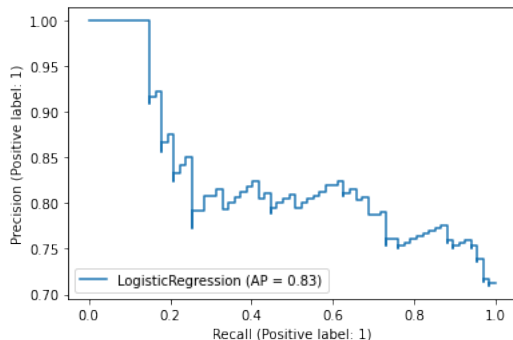
Вычисление вероятности принадлежности классу: $\mathbb{P}(y|x) = \sigma(y_i \langle x_i, w \rangle)$

Сравнение моделей: логистическая регрессия

Метрики качества при l_2 - регуляризации:

ROC-AUC score: 0.636727078891258

PR-Curve:

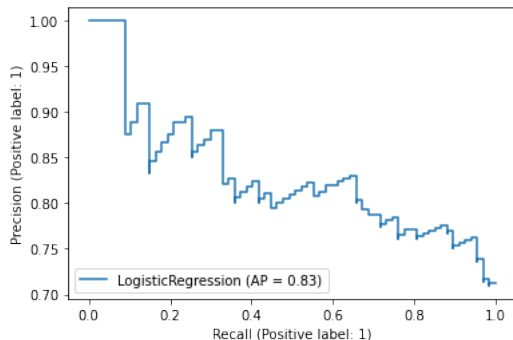


Сравнение моделей: логистическая регрессия

Метрики качества при l1 - регуляризации:

ROC-AUC score: 0.6292643923240938

PR-Curve:



Сравнение моделей: градиентный бустинг

- Градиентный бустинг.

Ансамблевый алгоритм из нескольких базовых алгоритмов.

$$a_N = \sum_{n=1}^N b_n(x)$$

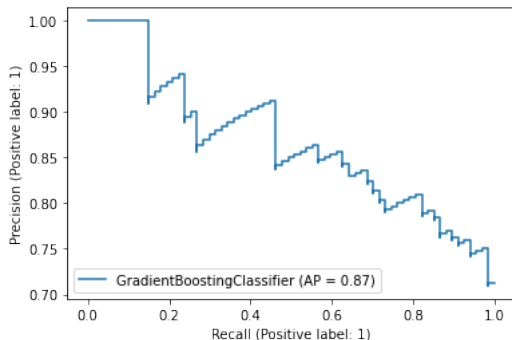
Каждый i -ый алгоритм корректирует ошибки предыдущих $i-1$ алгоритмов.

Сравнение моделей: градиентный бустинг

Метрики качества при 900 базовых алгоритмов и `learning_rate = 0.01`:

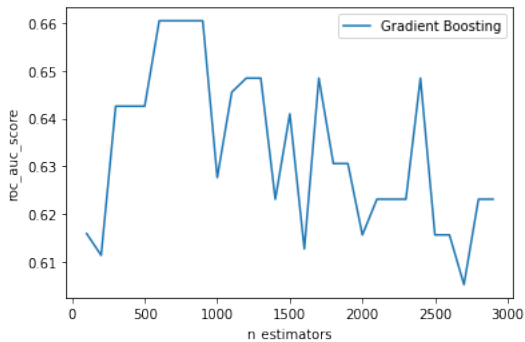
ROC-AUC score: 0.6708422174840085

PR-curve:



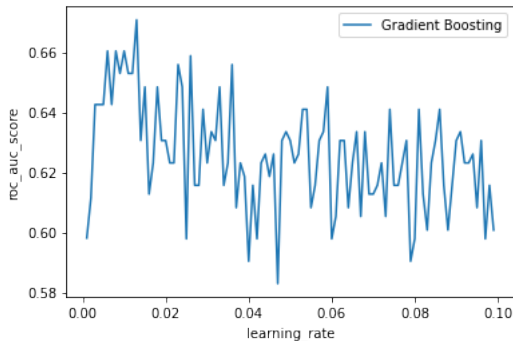
Сравнение моделей: градиентный бустинг

Зависимость `roc_auc_score` от количества базовых алгоритмов.



Сравнение моделей: градиентный бустинг

Зависимость roc_auc_score от показателя learning_rate.



- Нейронная сеть

В данной задаче мы будем рассматривать **нейронную сеть прямого распространения** или **многослойный перцептрон**.

Цель сети прямого распространения - аппроксимировать некоторую функцию f^*

В нашей задаче $y = f^*(x)$ целевая функция отображает вход x в метку класса y .

Сравнение моделей: нейронная сеть

Модель с параметрами:

Количество слоев: 8, количество параметров в каждом по 18, в выходном слое 1

2-ой и выходной слой с сигмоидальной функцией активации, остальные с линейными.

optimizer: sgd;

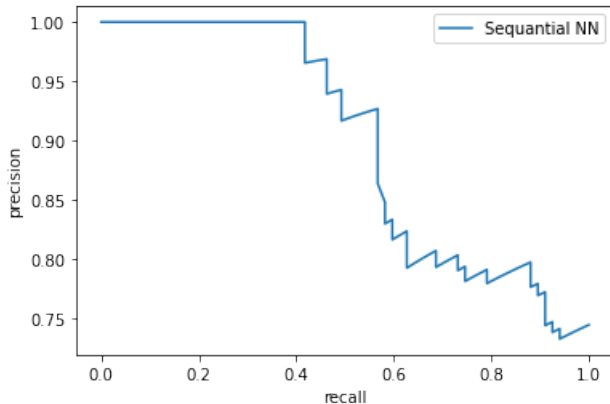
loss: mse;

Лучший показатель AUC-ROC: 0.7513326226012793

Accuracy score: 0.7684210526315789

Сравнение моделей: нейронная сеть

PR-curve:



Определение

Сложность модели - это интерпретируемость модели; показатель, того насколько легко модель подается понимаю.

Ранжировать рассматриваемые в задаче модели можно следующим образом(чем ниже, тем сложнее):

- Логистическая регрессия
- Градиентный бустинг
- Нейронная сеть

Сравнение моделей: синтетическая выборка

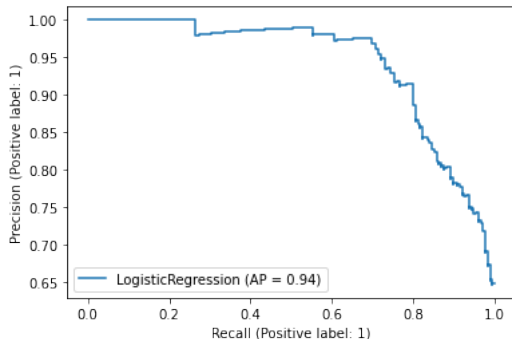
В качестве синтетической выборки был выбран датасет сгенерированный с помощью `sklearn.datasets.make_classification`. В данной выборке все признаки числовые, принимают значения чисел с плавающей точкой. Согласно документации признаки уже нормализованны и не нужно проводить нормализацию на этапе предобработки

Сравнение моделей: синтетическая выборка

Метрики качества логистической регрессии при l_2 - регуляризации:

ROC-AUC score: 0.8340092165898618

PR-Curve:

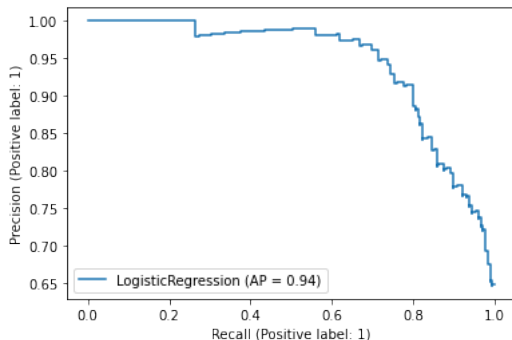


Сравнение моделей: синтетическая выборка

Метрики качества логистической регрессии при l_1 - регуляризации:

ROC-AUC score: 0.8340092165898618

PR-Curve:

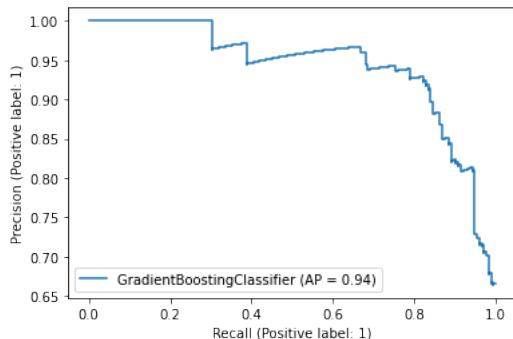


Сравнение моделей: синтетическая выборка

Метрики качества градиентного бустинга при 1000 базовых алгоритмах и `learning_rate = 0.01`:

ROC-AUC score: 0.8504147465437788

PR-curve:



Сравнение моделей: синтетическая выборка

Модель с параметрами:

Количество слоев: 8, количество параметров в каждом по 20, в выходном слое 1

2-ой и выходной слой с сигмоидальной функцией активации, остальные с линейными.

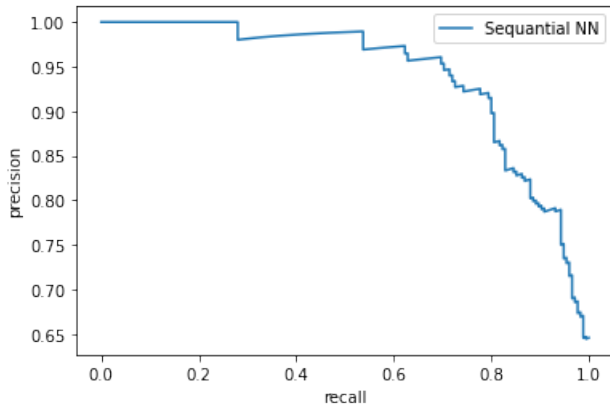
optimizer: sgd;

loss: mse;

Лучший показатель AUC-ROC: 0.928589861751152

Сравнение моделей: синтетическая выборка

PR-curve:



В ходе экспериментов с данными и параметрами моделей можно сделать следующие выводы:

- Нейронная сеть показала лучшее качество на реальных данных и синтетических.
- На синтетических данных оправдана сложность модели, в том смысле, что более сложные модели показывают лучше качество, чем более простые.