Домашнее задание №1. Идентификация видов стекла

Михаил Лепехин, группа 694

21 февраля 2019 г.

Постановка задачи

Выборка состоит из 9 признаков - химических параметров образцов и 214 объектов. Необходимо каждому образцу сопоставить один из 6 классов (например: стекло автомобиля, осколок посуды, окно здания).

Метрика качества

Доля правильных ответов классификатора (*accuracy*).

Пусть n - количество элементов в тестовой выборке, $y \in \{1, ..., 6\}^n$ - истинные значения классов стекла, $\widehat{y} \in \{1, ..., 6\}^n$ - предсказания алгоритма.

Тогда *accuracy* можно записать следующим образом:

$$accuracy = \frac{1}{n} \sum_{i=1}^{n} I(y_i = y)$$

Цель

1. Решить задачу с использованием следующих методов:

1) Алгоритм k ближайших соседей,

2) Алгоритм решающего дерева

и сравнить качество их работы.

2. Отмасштабировать признаки и проверить, даст ли такая процедура увеличение значения *accuracy*.

3. Выбрать наиболее точный из алгоритмов и исследовать зависимость *accuracy*, достигаемое этим алгоритмом на разных количествах признаков.

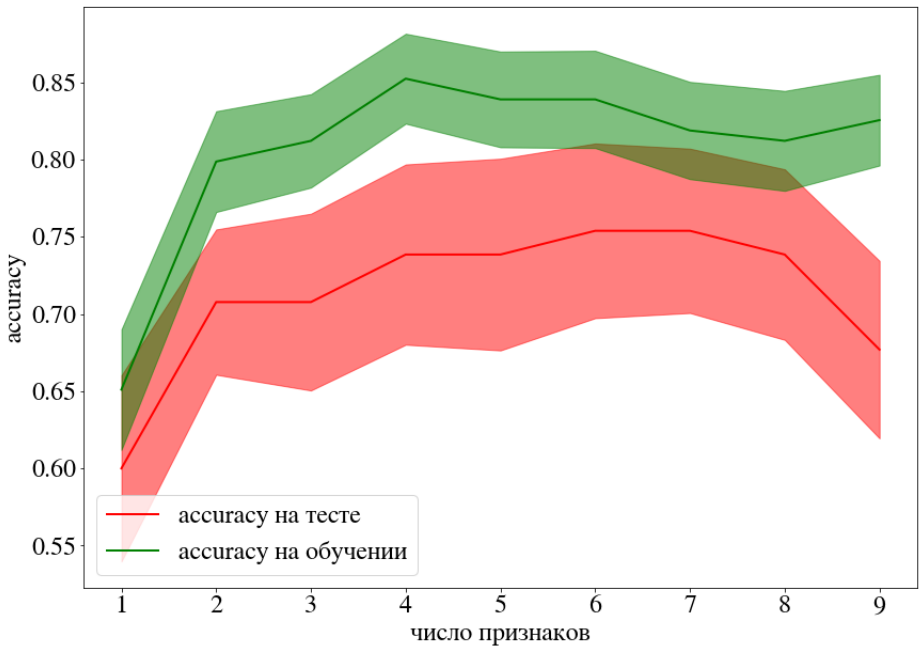
Метод отбора признаков

Отбор признаков быд производён жадным образом. А именно, будем на каждом шаге из недобавленных признаков выбирать такой, что его добавление максимизирует метрику качества на тесте.

Результаты эксперимента

После масштабирования признаков точность метода k ближайших соседей повысилась, а метода решающих деревьев - осталась такой же, как и до масштабирования.

Поскольку точность метода k ближайших соседей оказалась выше, будем в дальнейших экспериментах использовать именно его.



Вывод

Как можно увидеть по построенному графику, начиная с 6 признаков, точность работы алгоритма на тестовой выборке начинает ухудшаться. Кроме того, наилучшее значение ‘accuracy‘ на обучающей выборке достигается, когда число признаков равно 4.

Это значит, что в данных есть как минимум 3 незначимых признака, от которых необходимо избавиться, чтобы избежать переобучения.