МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики»

Факультет информационных технологий и программирования

Кафедра информационных систем

Интеллектуальные системы и технологии
Лабораторная работа 1

Выполнила студентка группы № М3403

Слобода Полина

Задача классификации – необходимо определить, к какому классу относится объект.

Задача регрессии – необходимо выдать не дискретную метку класса, а одно из значений аналогового множества значений.

Обучающий набор скачала с предложенного сайта, это файл данных csv (/data/glass). В нем 10 атрибутов, 9 признаков и 1 целевой атрибут ("Туре"). Этот тип может принимать 6 различных значений, то есть принадлежать к одному из 6 классов (задача классификации). 9 признаков — дробные числа. Так как они все измеряются в разных величинах (единицы, десятки), а датасет небольшой, необходимо нормализовать данные (привести их к среднему значению 0 и дисперсии 1). Дисперсия — мера разброса величины относительно мат ожидания. Это можно получить, вычитая из данных признака среднее значение признака и деля полученные значения на их стандартное отклонение.

От задачи классификации переходим к задаче регрессии.

Наивный способ

Каждое значение класса представляется одним числом. Во время предсказания полученный ответ округляется до ближайшего целого числа.

Для подбора лучшей комбинации функции расстояния, окна, ядра и метода используется алгоритм, в котором перебираются все возможные комбинации параметров и оцениваются по f-мере запуска алгоритма с такой комбинацией на всех данных (на предсказании класса каждого имеющегося вектора по прочим). Лучшей комбинацией считается та, у которой выходит максимальная f-мера.

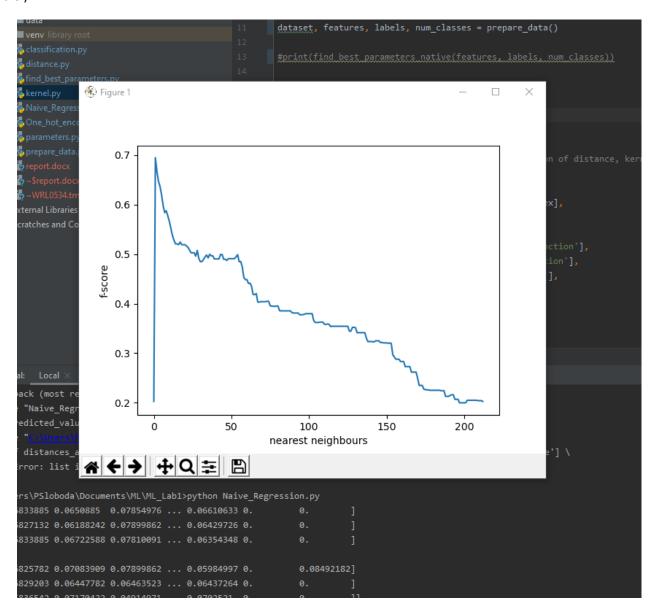
Скриншот части вывода данных долгой обработки данных для поиска лучших параметров:



На скриншоте можно заметить, что для нескольких значений параметров f-мера одинакова и максимальна. При этом среди всех лидеров функция расстояния манхеттеновская, а метод окна — взятие радиуса по ближайшим соседям (параметр 1, что логично). Функция ядра может быть различной, максимальные значения получились при вариантах 'uniform', 'triangular', 'epanechnikov', 'quartic', 'triweight', 'tricube', 'cosine'. Ядра 'gaussian', 'logistic', 'sigmoid' показали результат хуже.

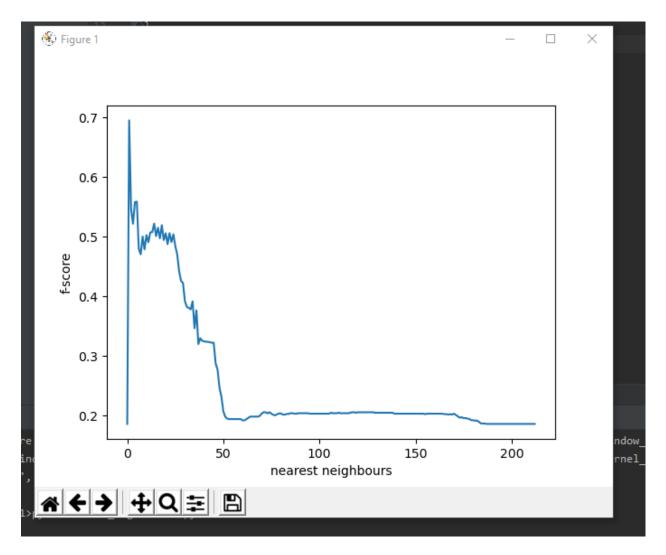
Зафиксируем лучшими параметрами комбинацию 'manhattan', 'triweight', 'variable'.

После получения лучшей комбинации для данного датасета производится построение графика зависимости f-меры от числа ближайших соседей (так как в данном случае используется метод 'variable'):



На графике видно, что при нуле соседей f-мера нулевая, при одном — максимальна, далее убывает и стремится к нулю, что вполне соответствует здравому смыслу. Параметр окна 1, как и в поиске лучших параметров, так и в проверке на графике, выдает максимальную f-меру.

Ради интереса построила еще график с использованием другой функции ядра ('manhattan', 'uniform', 'variable'):



По сравнению с предыдущим методом зависимость обладает большими скачками, не очень понятно куда стремится, поэтому эта функция ядра менее подходящая как минимум для данного датасета.

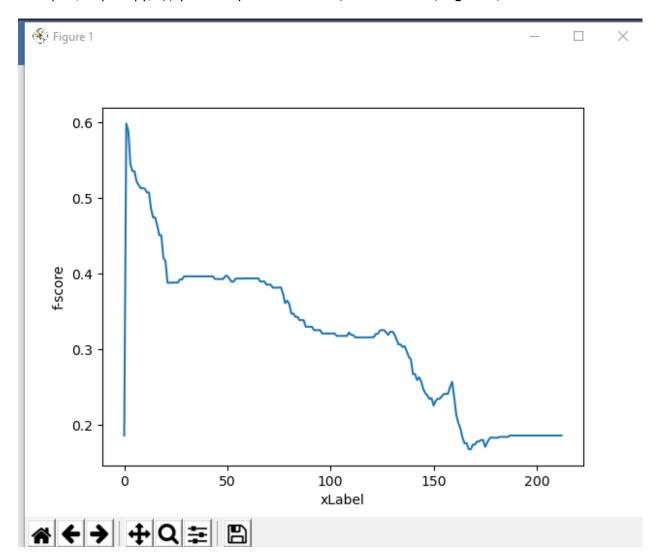
OneHot

Вместо одного целевого признака в набор данных добавляется столько новых числовых переменных, сколько в нём содержится классов. В соответствие каждому вектору ставится многомерный ответ (к примеру, [0 0 1 0 0 0] — принадлежность объекта к третьему классу). Каждую часть ответа надо будет вычислить отдельно, после по массиву определить, к какому классу в итоге относился объект. В остальном все похоже на наивный способ — лучшая комбинация определяется из всех возможных максимальной f-мерой.

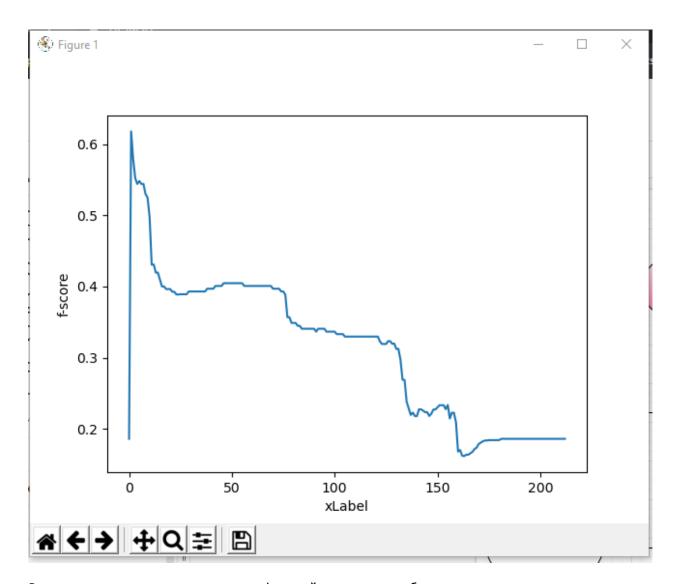


В данном случае одинаковых мер получилось гораздо больше. Функции расстояния в лучших комбинациях встречаются все три, функции ядра тоже все (все 10 вариантов). Неизменным остался метод подсчета окна по количеству соседей с параметром 1.

Выберем, к примеру, одну из наилучших комбинаций 'manhattan', 'sigmoid', 'variable'



И тоже сравнить график с получаемым от другой функции расстояний ('euclidean', 'sigmoid', 'variable'):



Разница в использовании различных функций расстояния небольшая.