Лабораторная работа №3
Метрические алгоритмы, выступ, типичность объектов, проклятие размерности

1. KNeighborsClassifier

Примените метод k ближайших соседей KNeighborsClassifier к вашей задаче бинарной классификации. Метрические алгоритмы чувствительны к масштабу признаков, поэтому предварительно нормализуйте ваш датасет с помощью StandardScaler.

Нарисуйте на одном графике зависимость ошибок предсказания на Train и Test датасетах от количества соседей n_neighbors. Какое значение n_neighbors является оптимальным? Для каких значений n_neighbors метод недообучен, переобучен?

2. KNeighborsRegressor

Подготовьте нормализованный датасет для задачи регрессии. Сгенерируйте 1000 точек, равномерно заполняющих отрезок AB в пространстве X вашего датасета, где A - 10% квантиль всех признаков, B - 90% квантиль всех признаков. Нарисуйте на одном графике предсказания для точек на отрезке AB методом KNeighborsRegressor с разными значениями параметра n_neighbors. Объясните отличия в графиках.

Нарисуйте на одном графике предсказания для точек на отрезке AB методом KNeighborsRegressor с разными значениями параметра weights. Подберите остальные параметры так, чтобы графики отличались. Объясните отличия в графиках.

Нарисуйте на одном графике предсказания для точек на отрезке AB методами KNeighborsRegressor и RadiusNeighborsRegressor с одинаковыми значениями общих параметров. Подберите остальные параметры так, чтобы графики отличались. Объясните отличия в графиках.

3. Проклятие размерности

В пространствах большой размерности наблюдается удивительное явление: расстояния до всех точек датасета совпадают и ближайшие соседи выбираются случайным образом. Продемонстрируйте этот эффект, нарисовав и сравнив гистограммы распределения расстояний между точками в вашем датасете разным цветом для малого и большого количества используемых в метрике признаков.

4. Типичность объектов

Для найденного в п.1 наилучшего значения параметра n_neighbors вычислите по формуле из лекций значения выступа всех объектов обучающей выборки Train (вместо Г используйте вероятности, возвращаемые KNeighborsClassifier.predict_proba). Нарисуйте график отсортированных значений выступов (как в лекции). В комментариях напишите, где на графике случайные выбросы, периферийные объекты и эталоны. Работает ли для вашего датасета критерий крутого склона?