Лабораторная работа №4

Машина опорных векторов

Машины опорных векторов (support vector machine, SVM) один из крайне популярных алгоритмов машинного обучения. Данное семейство алгоритмов может применяться как для решения задач классификации, так и для задач регрессии. С одной стороны, он относится к классу линейных моделей. И не смотря на свою простоту может давать уверенные результаты. С другой стороны, алгоритм допускает решение задач классификации в случае, если выборка не является линейно разделимой. Данный подход (kernel trick) существенно расширяет возможности алгоритма, позволяя ему быть (даже буквально, геометрически) более гибким, чем другие линейные модели классификации.

Ход выполнения работы

1. Реализовать генератор входных данных, которые будут использоваться для обучения алгоритма и анализа качества обучения с помощью метрик после его обучения. Требования:

   (a) Признаки: $(x, y) \in [-1, 1] \times [-1, 1]$. Иными словами, пространство признаков - квадрат в плоскости $\mathbb{R}^2$.

   (b) Граница разделения классов: $x^2 + y^2 = \frac{1}{4}$. Объекты одного класса лежат внутри окружности $R = \frac{1}{2}$, объекты другого класса лежат вне окружности.

   (c) Входной параметр генератора: размер выборки.

2. Реализовать функции метрик качества: accuracе, precision, recall, F-мера. Входные данные: истинные метки классов, предсказанные метки классов. Выходные данные: значение метрики

3. Обучить ансамбль моделей NuSVC с различными условиями:

   (a) Выбор ядра SVM (линейное, полиномиальное, гауссово (rbf), сигмоид). Построить графически классы с разными метками, а также разделяющую гиперповерхность для каждого из ядер. Объем обучающей выборки произвольный, но одинаковый для сравнения построенной поверхности для различных ядер. Сравнить метрики качества в зависимости от выбора ядра.

   (b) Объем обучающей выборки. Исследовать зависимость метрик качества от объема обучающей выборки.

   (c) Параметр ν - нижняя граница доли опорных векторов. Сравнить графически и на основе метрик качества.

   В пунктах (b) и (c) ядро допускается выбрать фиксированным, например rbf.

Рекомендации к выполнению

Для представления данных рекомендуется минимально использовать встроенные структуры python, такие как списки (list), кортежи (tuple), словари (dict). Более оптимально использовать пакеты *numpy* и *pandas* и структуры *numpy.array* и *pandas.DataFrame*.

Для настройки и обучения алгоритма $NuSVC$ рекомендуется использовать пакет *sklearn*.