Московский авиационный институт

(Национальный исследовательский университет)

Факультет: «Информационные технологии и прикладная математика»
Кафедра: 806 «Вычислительная математика и программирование»
Дисциплина: «Искусственный интеллект»

Лабораторная работа № 2

Тема: Алгоритмы машинного обучения.

Студент: Шевчук П.В.

Группа: М80-304Б

Преподаватель: Ахмед Самир Халид

Москва, 2019

1. Постановка задачи

Требуется реализовать класс на выбранном языке программирования, который реализует один из алгоритмов машинного обучения. Обязательным является наличия в классе двух методов fit, predict. Необходимо проверить работу вашего алгоритма на ваших данных (на таблице и на текстовых данных), произведя необходимую подготовку данных. Также необходимо реализовать алгоритм полиномиальной регрессии, для предсказания значений в таблице. Сравнить результаты со стандартной реализацией sklearn, определить, в чем сходство и различие ваших алгоритмов. Замерить время работы алгоритмов.

2. Выбор алгоритма

Вариант: 14% 6 + 1 = 3

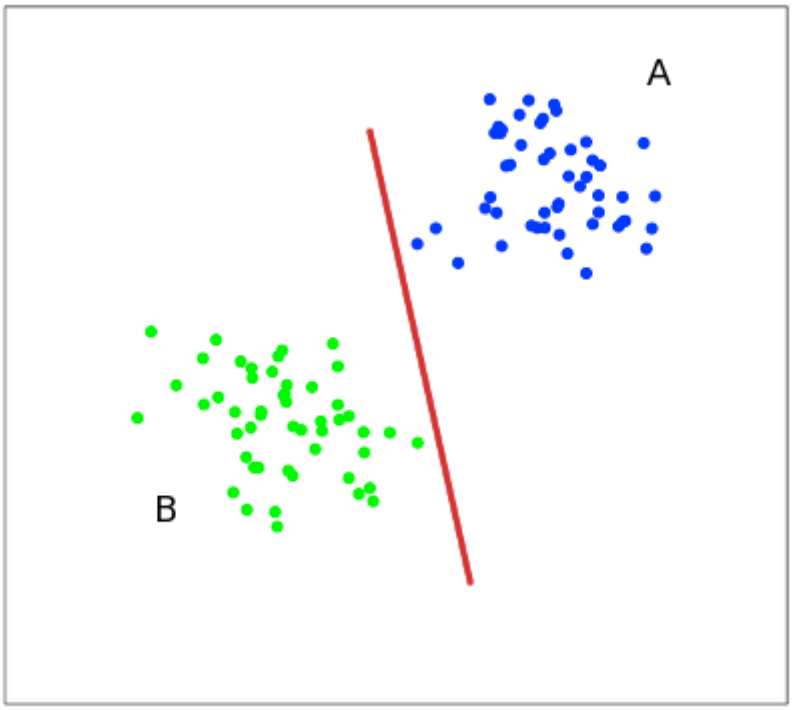
3) SVM (Метод опорных векторов)

3. Решение задачи

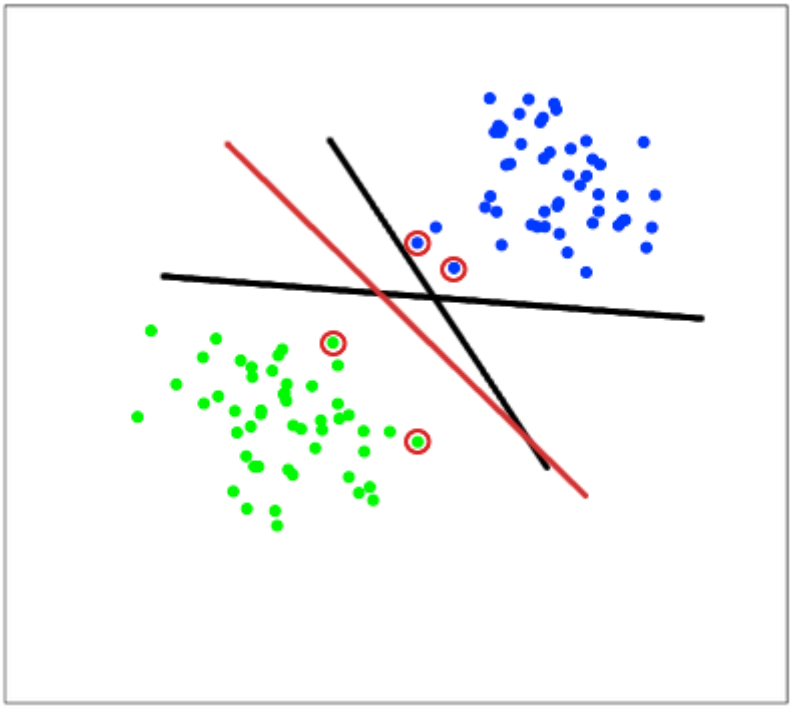
Задача 1:

Метод опорных векторов (support vector machine, SVM) — один из наиболее популярных методов обучения, который применяется для решения задач классификации и регрессии. Основная идея метода заключается в построении гиперплоскости, разделяющей объекты выборки наиболее оптимальным способом. Алгоритм работает в предположении, что чем больше расстояние (зазор) между разделяющей гиперплоскостью и объектами разделяемых классов, тем меньше будет средняя ошибка классификатора.

Идею метода удобно проиллюстрировать на следующем простом примере: даны точки на плоскости, разбитые на два класса. Проведем линию, разделяющую эти два класса. Далее, все новые точки (не из обучающей выборки) автоматически классифицируются следующим образом: точка выше прямой попадает в класс A, точка ниже прямой — в класс B.



Такую прямую назовем разделяющей прямой. Однако в пространствах высоких размерностей прямая уже не будет разделять наши классы, так как понятие «ниже прямой» или «выше прямой» теряет всякий смысл. Поэтому вместо прямых необходимо рассматривать гиперплоскости — пространства, размерность которых на единицу меньше, чем размерность исходного пространства. В $\mathbb{R}^3$, например, гиперплоскость — это обычная двумерная плоскость. В нашем примере существует несколько прямых, разделяющих два класса (рис. 2):



С точки зрения точности классификации лучше всего выбрать прямую, расстояние от которой до каждого класса максимально. Другими словами, выберем ту прямую, которая разделяет классы наилучшим образом (красная прямая на рис.2). Такая прямая, а в общем случае — гиперплоскость, называется оптимальной разделяющей гиперплоскостью.

Вектора, лежащие ближе всех к разделяющей гиперплоскости, называются опорными векторами(support vectors). На рисунке они помечены красным.

Задача выбрать такие w — нормальный вектор к разделяющей гиперплоскости, b — вспомогательный параметр, чтобы они максимизировали расстояние каждого класса.

Результаты:

При реализации были получены следующие результаты, для нормализованного вектора столбца GRE SCORE.

```
PS C:\Users\Паша\Downloads\AI\лаба 2> python .\test.py
Time: 3.6287723052
Accuracy: 0.736
Time: 0.0118376594
Accuracy: 0.81
PS C:\Users\Паша\Downloads\AI\лаба 2>
```

Как видим, наша реализация несильно уступает в точности, но гораздо больше идёт по времени. Возможно, это связано с пересчётом весов, т.к. они часто пересчитываются.

Задача 2:
Построим задачу линейной регрессии с помощью библиотеки sklearn и и вывод предсказанных данных, также вывод коэффициента среднеквадратической ошибки (mean squared error = MSE)

Выражаясь простым языком, модель регрессии в математической статистике строится на основе известных данных, в роли которых выступают пары чисел. Количество таких пар заранее определено. Если представить себе, что первое число в паре – это значение координаты x, а второе – y, то множество таких пар чисел можно представить на плоскости в декартовой системе координат в виде множества точек. Данные пары чисел берутся не случайно. На практике, как правило, второе число зависит от первого. Построить регрессию – это значит подобрать такую линию (точнее, функцию), которая как можно точнее приближает к себе (аппроксимирует) множество вышесказанных точек.
Результаты:

Предскажем значения TOEFL Score от GRE Score.

```
PS C:\Users\Паша\Downloads\AI\лаба 2> python .\poly.py
MSE: 1564.4504935381
[ 1236  899  4356  832  2476  5498  3492  1231  891  6743]
[ 1112  925  5942  890  2287  5214  3674  1334  716  5210]
PS C:\Users\Паша\Downloads\AI\лаба 2>
```

Наблюдаем большой коэффициент среднеквадратичной ошибки. Это связано с тем, что среднеквадратичная ошибка является отрицательным значением ожидаемого значения одной конкретной функции полезности , квадратичной функции полезности, которая может не подходить для соответствующей функции полезности при данном наборе обстоятельств.

4. Вывод

В ходе данной лабораторной работы освоен алгоритм метод опорных векторов. А также были получены навыки в использовании библиотек cxopt и sklearn.

Ссылка на репозиторий github: https://github.com/pavels-k/AI/tree/master/lab2