Домашнее задание №2 по курсу «Машинное обучение»: линейные модели

Колесов Алексей

12 сентября 2018 г.

В первой задаче вам необходимо прислать python-файл `features.py`, в комментариях которого должны быть указаны подобранные признаки и их мотивация.

Во второй задаче необходимо прислать `IPython-notebook` с решением.

1 Задания

1. В этой задаче вам необходимо будет произвести `feature-engineering` на примере датасета MNIST — `http://yann.lecun.com/exdb/mnist/`. Датасет представляет собой чёрно-белые изображения рукописных цифр, разделённых на десять классов: по признаку того, какая цифра была написана. На этом датасете можно решать 45 задач бинарной классификации: каждая цифра против каждой. Ваша задача для каждой из этих задач придумать по два признака, так чтоб логистическая регрессия решала бы задачу как минимум с точностью 75 процентов.

   Для решения этой задачи вам предоставлены два файла: `main.py` с кодом, который скачивает датасет, составляет 45 задач, обучает для каждой задачи логистическую регрессию и измеряет полученный результат. Этот файл вам менять не нужно, но стоит ознакомиться для понимания.

   Второй файл — `features.py` — пример которого вам тоже дан, должен быть написан вами. В нём необходимо реализовать набор функций, принимающих один аргумент — вектор размерности 784 (исходные картинки имеют размер 28×28) и выдаёт единственное вещественное число.

   Файл `features.py` должен экспортировать переменную `FEATURES` — `dict` из пары цифр в пару функций, вычисляющих признаки. Внутри этого файла нельзя использовать ничего кроме стандартной библиотеки и библиотеки `numpy`.

   Задача считается решённой, если скрипт отработал менее, чем за 20 минут и решил как минимум 40 случаев. Кроме того отдельные баллы будут засчитаны решениям, которые будут в топе по средней точности, минимальной точности (которую надо максимизировать) и количеству решённых задач.

   Запуск файла `main.py` выдаст вышеописанные числа на стандартный вывод.

2. Реализуйте два алгоритма построения линейной модели для решения задачи восстановления регрессии:

   - ridge-регрессия (рассматривали на лекции)
   - регрессия с функцией потерь $L(h) = \sum_{i=1}^{m} |h(x_i) - y_i|$ (рассматривали на семинаре)

   Для реализации нельзя пользоваться пакетами машинного обучения. Можно пользоваться пакетами линейной алгебры (в частности реализациями матричного умножения и SVD-разложения) и линейного программирования (непосредственно для решения задач в том виде, что она была поставлена на лекции).

   Рядом приложен csv-файл, на данных из которого вам необходимо протестировать работу ваших методов. Файл описывает игроком NBA и состоит из пяти колонок:

- высота в футах
- вес в фунтах
- вероятность удачного попадания с игры
- вероятность удачного попадания со штрафного
- среднее количество очков, набранных в игре

Вам необходимо построить линейную модель, приближающую последнюю величину из четырёх первых. Сравните полученные решения. Постройте график MSE в зависимости от τ в случае ridge-регрессии.

Предложите пример задачи, когда квадратичная функция потерь более естественна с точки зрения предметной области, чем модуль, и наоборот.