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

Колесов Алексей

15 сентября 2017 г.

Вам необходимо прислать pdf-файл с доказательствами, графиками и прочим. Отдельно нужно прислать код программы и способ её запустить (в идеале, ссылку на IPython-notebook).

1 Задания

1. Предложите алгоритм для генерации случайной линейно разделимой выборки в двумерном пространстве. Формально, пусть ваш алгоритм принимает на вход следующие величины:

   - $\min_x$, $\max_x$ ($\min_x < \max_x$) — допустимые границы для 1-й координаты
   - $\min_y$, $\max_y$ ($\min_y < \max_y$) — допустимые границы для 2-й координаты
   - N — количество объектов отрицательного класса
   - P — количество объектов положительного класса

   Алгоритм должен за время $\mathcal{O}(N + P)$ породить выборку длиной $N + P$, в которой ровно N объектов одного класса, ровно P другого, каждый объект лежит в прямоугольнике $[\min_x; \max_x] \times [\min_y, \max_y]$. Кроме того, для обеспечения разнообразности ваш алгоритм должен удовлетворять следующим ограничениям:

   - вероятность того, что два объекта выборки имеют равные координаты должна быть равна нулю
   - вероятность того, что прямая с угловым наклоном k окажется разделяющей для полученной выборки должна не зависеть от k

   Опишите алгоритм, докажите его свойства. Реализуйте предложенный алгоритм

2. Реализуйте алгоритм Batch Perceptron в случае двумерного пространства и класса **неоднородных** линейных моделей. Запустите алгоритм на выборках, полученных алгоритмом из предыдущего задания. На выходе алгоритм должен выдавать параметры модели и минимальный отступ этой модели по объектам выборки. Пусть $N = P = k$, $\min_x = \min_y = -10$, $\max_x = \max_y = 1$. Постройте график среднего количества шагов, необходимых алгоритму в зависимости от k.

3. Бонусное задание (+1 балл) Реализуйте визуализацию алгоритма Batch Perceptron. На вход программа должна принимать параметры для алгоритма генерации выборки (см. 1-е задание). На выходе — пошаговую визуализацию алгоритма: выбор неправильно классифицируемого объекта, сдвиг w.

4. 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.