Домашнее задание №1.

Линейные модели.

Теоретические задачи.

1. Вывести формулу линейной регрессии для одномерного случая. Дана выборка $\{x_i, y_i\}_{i=1}^N$, методом наименьших квадратов определить коэффициенты линейной регрессии: $a(x) = w_0 + w_1 x$.

2. Рассмотрим задачу обучения линейной регрессии:

$$Q(\vec{w}) = \sum_{i=1}^{N} q(\vec{w}, \vec{x_i}) = \sum_{i=1}^{N} ((\vec{w}, \vec{x_i}) - y_i)^2,$$

   и будем решать ее с помощью стохастического градиентного спуска, т.е. на каждом шаге оптимизировать значение функционала на i-ом объекта $q(\vec{w}, \vec{x_i})$.

   Шаг итерационного процесса имеет вид:

$$\vec{w}^{(t+1)} = \vec{w}^t - \eta_t \nabla_w q(\vec{w}^t, \vec{x_i}).$$

   Определите длину шага, соответствующую наискорейшему спуску, т.е :

$$Q(\vec{w}^{(t+1)}) \to \min_{\eta}.$$

3. Пусть дана некоторая выборка X и классификатор $b(x)$, возвращающий в качестве оценки принадлежности объекта x к положительному классу 0 или 1 (а не вероятности).

   1) Постройте ROC-кривую для классификатора $b(x)$ на выборке X.

   2) Покажите, что AUC-ROC классификатора $b(x)$ может быть выражена через долю правильных ответов и полноту классификатора $a(x;t)$, получающегося при выборе некоторого порога $t \in (0;1)$ ($a(x) = [b(x) > 1]$). Помимо указанных величин в формулу могут входить N, N_+, N_-, число объектов, число положительных и отрицательных объектов в выборке X соответсвенно.

Практические задачи.

«Федералист» — сборник из 85 статей в поддержку ратификации Конституции США. Статьи выходили с октября 1787 года по август 1788 года в нью-йоркских газетах «The Independent Journal» и «The New York Packet». Сборник всех статей под заглавием «Федералист» увидел свет в 1788 году. «Федералист» считается не только ценнейшим источником толкования Конституции США (в сборнике значение положений Конституции разъясняется самими её авторами), но и выдающимся философским и политическим произведением. «Федералист» был написан группой авторов: А. Гамильтон, Д. Мэдисон, Д. Джей, но при издании отдельных статей в газетах каждая была подписана псевдонимом Публий. Известно, что при установлении авторства 12 спорных работ эксперты не пришли к единому мнению, написаны ли они Гамильтоном или Мэдисоном.

В данной задаче рассматриваются статьи авторства Гамильтона и Мэдисона. Выборка разделена на три части: “train 86 by 71.txt” (содержит информацию о 86 статьях), “valid 20 by 71.txt” (20 статей) , и “test 12 by 70.txt” (12 спорных статей). Каждый файл, во-первых, содержит авторство статьи (первый столбец в “train 86 by 71.txt”, “valid 20 by 71.txt”), 1 соответствует Гамильтону, 2 – Мэдисону, в тестовой выборке такой столбец отсуствует. Далее, следуют столбцы 70 признаков, которые представляют собой частоты встречаемости отдельных слов (на 1000 слов) в соответсвующих статьях.

1. Постройте модель логистической регрессии для классификации статей по двум авторам (Гамильтон и Мэдисон), используйте модели из модуля sklearn с параметрами по умолчанию. Вычислите метрики качества алгоритма: точность, полнота, выведите матрицу ошибок. Вычислите интегральные матрики модели: AUC-ROC, AUC-precision/recall. В качестве обучающей выборки используйте “train 86 by 71.txt”, в качестве тестовой “valid 20 by 71.txt”.

2. Проведите нормировку признаков, как поменяются метрики качества алгоритма?

3. Попробуйте поменять дефолтные параметры модели логистической регрессии (коэффициент C, тип регуляризации), как при этом меняются метрики качества?

4. На обучающей выборке “train 86 by 71.txt”, постройте метод опорных векторов. Используя выборку “valid 20 by 71.txt”, определите оптимальное значение коэффициента C , соответсвующее наибольшой точности на тестовой выборке.

5. Представьте таблицу предсказаний (авторство статей) для выборки “test 12 by 70.txt” для построенных моделей.