Отчет по заданию 1. Метрические алгоритмы классификации

Зотов Арсений, 317 группа

20 ноября 2020

Постановка задания

В данном задании мною была проделана следующая работа:

1. Написана собственная реализация линейного классификатора с произвольной функцией потерь.
2. Выведены формулы градиента функции потерь для бинарной и многоклассовой логистической регрессии.
3. Проведены соответствующие эксперименты с датасетом Toxic.

Теоретичесчкая часть

1. Формула градиента функции потерь для задачи бинарной логистической регрессии.

$y \in [1, -1]$
$L = \frac{1}{n} \times \sum_{i=1}^{n} L_i + \frac{\lambda}{2} \parallel w \parallel_2^2$, $L_i = log(1 + e^{-M_i})$,
$M_i = y_i(w_1 x_1 + ... + w_d x_d)$
$\nabla_w L = (\frac{\partial L}{\partial w_1}, ..., \frac{\partial L}{\partial w_d})$
$\frac{\partial L}{\partial w_j} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial L_i}{\partial M_i} \frac{\partial M_i}{\partial w_j} + \lambda w_j = \frac{1}{n} \sum_{i=1}^{n} \frac{-e^{-M_i}}{1+e^{-M_i}} y_i x_j + \lambda w_j =$
$= -\frac{1}{n} \sum_{i=1}^{n} \frac{1}{1+e^{M_i}} y_i x_j + \lambda w_j$

2. Формула градиента функции потерь для задачи многоклассовой логистической регрессии.

$y \in Y, \mid Y \mid = K$
$L = -\frac{1}{n} \times \sum_{i=1}^{n} L_i + \frac{\lambda}{2} \sum_{k=1}^{K} \parallel w_k \parallel_2^2$, $L_i = log(\frac{exp(\langle w_{y_i}, x_i \rangle)}{\sum_{k=1}^{K} exp(\langle w_k, x_i \rangle)})$

$$\nabla_{w_l} L = (\frac{\partial L}{\partial w_{l1}}, ..., \frac{\partial L}{\partial w_{ld}})$$
$$\frac{\partial L}{\partial w_{lj}} = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial L_i}{\partial w_{lj}} + \lambda w_{lj} =$$
$$= -\frac{1}{n}\sum_{i=1}^{n} \left([y_i = l] * x_{ij} - x_{ij} * \frac{exp(\langle w_l, x_i \rangle)}{\sum_{k=1}^{K} exp(\langle w_k, x_i \rangle)} \right) + \lambda w_j$$

3. Покажем, что при количестве классов = 2, задача мультиномиальной логистической регрессии сводится к бинарной логистической регрессии.

$$P(y_i = 1|x_i) = \frac{exp(\langle w_1, x_i \rangle)}{exp(\langle w_1, x_i \rangle) + exp(\langle w_{-1}, xi \rangle)}$$
$$P(y_i = -1|x_i) = \frac{exp(\langle w_{-1}, x_i \rangle)}{exp(\langle w_1, x_i \rangle) + exp(\langle w_{-1}, xi \rangle)}$$

Разделим числитель и знаменатель в обоих равенствах на $exp(\langle w_1, x_i \rangle)$. Получим:

$$P(y_i = 1|x_i) = \frac{1}{1 + exp(\langle w_{-1} - w_1, xi \rangle)}$$
$$P(y_i = -1|x_i) = \frac{exp(\langle w_{-1} - w_1, x_i \rangle)}{1 + exp(\langle w_{-1} - w_1, xi \rangle)}$$

Сделаем замену $w = w_1 - w_{-1}$, получим:

$$P(y_i = 1|x_i) = \frac{1}{1 + exp(-\langle w, xi \rangle)}$$
$$P(y_i = -1|x_i) = \frac{exp(-\langle w, x_i \rangle)}{1 + exp(-\langle w, xi \rangle)} = \frac{1}{1 + exp(\langle w, xi \rangle)}$$

Откуда,

$$P(y_i|x_i) = \frac{1}{1 + exp(-y_i \langle w, xi \rangle)}$$

Получили вероятность для бинарной логистической регрессии.

Результаты экспериментов

Перед началом экспериментов была произведена предобработка текста: все тексты приведены к нижнему регистру, все символы, не являющиеся буквами и цифрами, заменены на пробелы.

Выборка была преобразована в разреженную матрицу с помощью метода BagOfWords.

Эксперимент №1

В данном эксперименте было исследовано поведение градиентного спуска для задачи логистической регрессии в зависимости от следующих параметров:

- параметр размера шага step_alpha
- параметр размера шага step_beta
- начального приближения

Для этого был проведен анализ зависимости значения функции потерь от времени работы метода.

На рисунке 1 представленны результаты экспериментов для различных параметров шага. Для наглядности была использована логарифмическая шкала по оси

Оу.

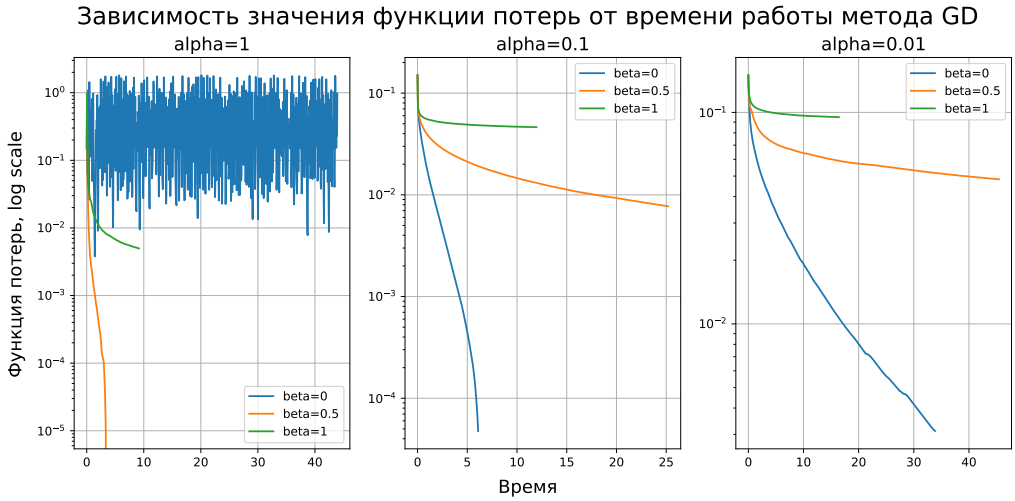

Рис. 1: Поведение GD при различных параметрах шага.

Видно, что при слишком большом beta метод не сходится к оптимальному решению. В дальнейшем для метода GD будем использовать параметры, показавшие лучший результат: alpha=0.1, beta=0.

На рисунке 2 представленны результаты экспериментов для различных начальных приближений. Для наглядности была использована логарифмическая шкала по оси Оу.

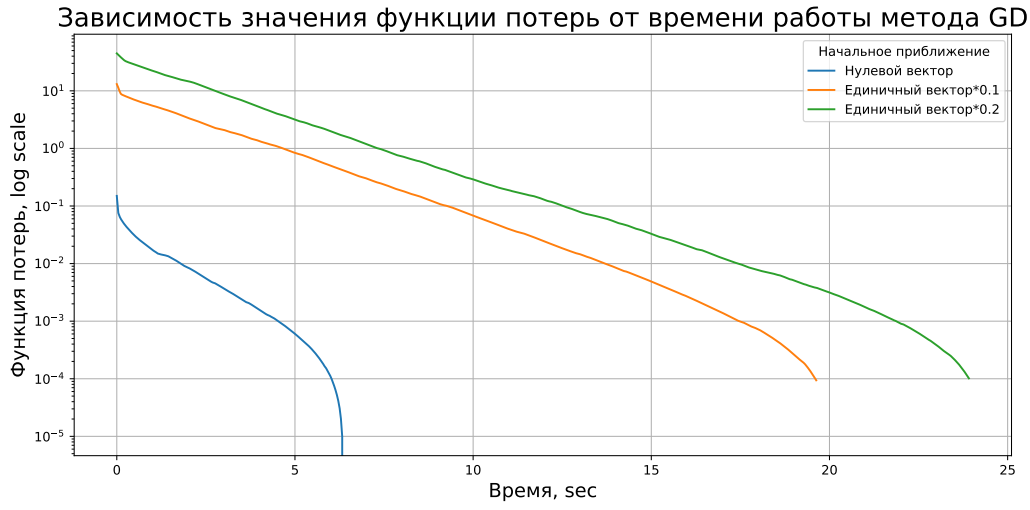

Рис. 2: Поведение GD при различных начальных приближениях.

Лучшим начальным приближение оказался нулевой вектор. В дальнейшем будем использовать его.

Эксперимент №2

В данном эксперименте было исследовано поведение стохастического градиентного спуска для задачи логистической регрессии в зависимости от тех же параметров, что и в Эксперименте 1.

Для этого был проведен анализ зависимости значения функции потерь от времени работы метода.

На рисунке 3 представленны результаты экспериментов для различных параметров шага. Для наглядности была использована логарифмическая шкала по оси Oy.

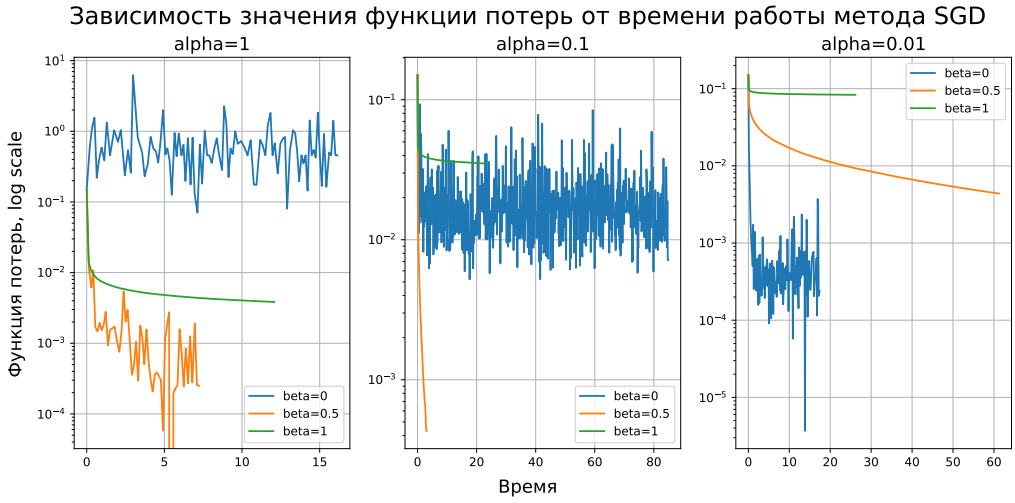

Рис. 3: Поведение SGD при различных параметрах шага.

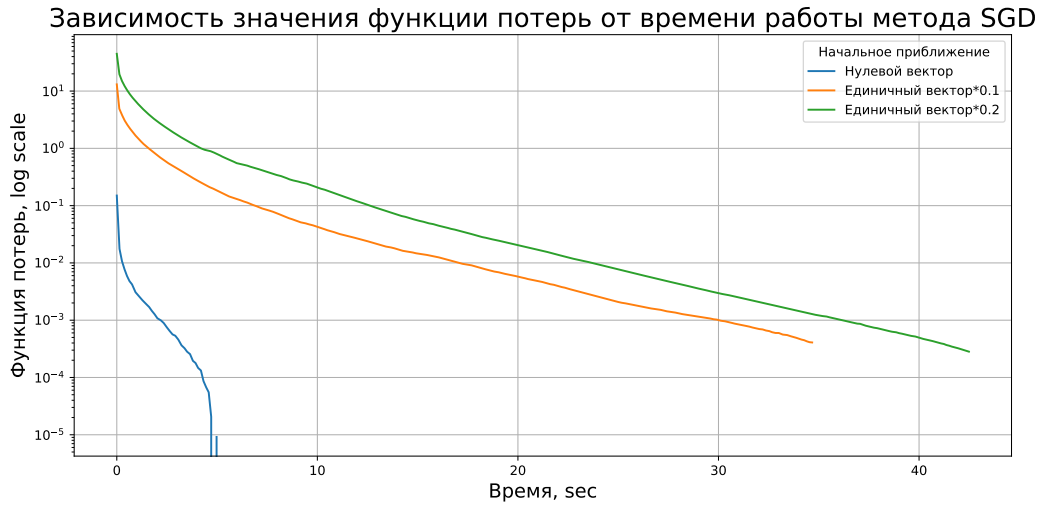

Рис. 4: Поведение SGD при различных начальных приближениях.

Видно, что при слишком больших и маленьких alpha метод сходится слишком

быстро, не достигнув оптимального решения. В дальнейшем для метода GD будем использовать параметры, показавшие лучший результат: alpha=0.1, beta=0.5.

На рисунке 4 представленны результаты экспериментов для различных начальных приближений. Для наглядности была использована логарифмическая шкала по оси Oy.

Лучшим начальным приближение оказался нулевой вектор. В дальнейшем будем использовать его.

Эксперимент №3

Сравним поведение двух методов между собой.

По предыдущим графикам видно, что по достижении точки сходимости SGD начинает "колебаться"около этой точки, при этом график получается ломанным, графики GD же более гладкие. Однако SGD сходится быстрее, что можно видеть на рисунке 5, где был проведен анализ зависимости точности на валидационной выборке от времени работы метода и эпохи.

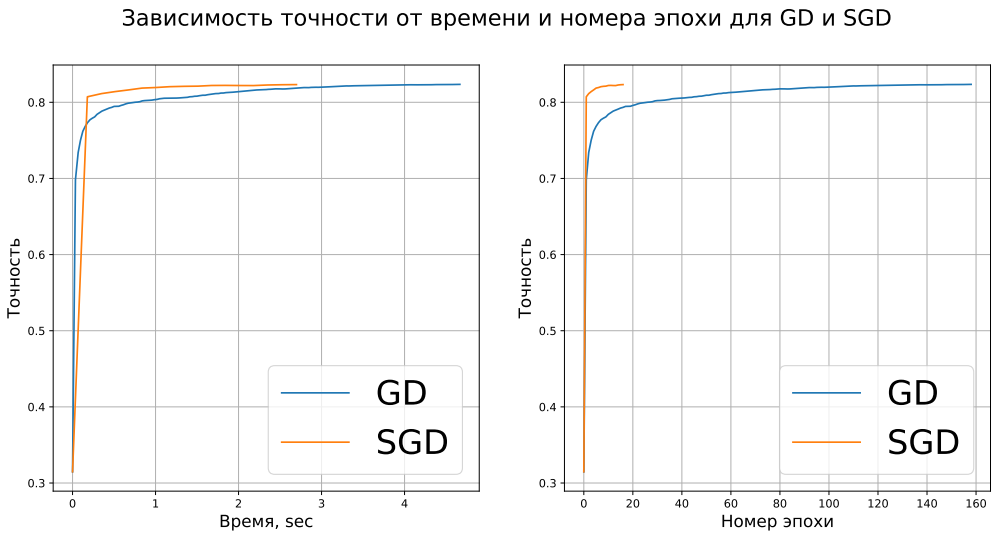

Рис. 5: Сравнение сходимости GD и SGD.

Можно заметить, что одна эпоха SGD занимает больше времени, чем GD, что связано с ралицацией метода на Python, в котором итерация с помощью for занимает значительное время. Для дальнейших экспериментов будем испольовать SGD.

Эксперимент №4

Применим алгоритм лемматизации к коллекции. Удалим из текста стоп-слова.

Исследуем, как предобработка корпуса повлияла на точность классификации, время работы алгоритма и размерность признакового пространства.

На рисунке 6 представленны результаты экспериментов для различных параметров шага.

Рис. 6: Сравнение качества классификации до и после предобработки.

Точность до предобработки: 0.8256

Точность после предобработки: 0.8271

Размерность признакового пространства до предобработки(BagOfWords): 18210

Размерность признакового пространства после предобработки(BagOfWords): 16189

Предобработка помогла уменьшить размерность признакового пространства и улучшить качество классификации на валидационной выборке. При этом время работы алгоритма незначительно выросло.

Эксперимент №5

Исследуем качество, время работы алгоритма и размер признакового пространства в зависимости от следуюших факторов:

• использовалось представление BagOfWords или Tfidf

• параметров min_df и max_df конструкторов.

Результаты экспериментов представлены на рисунке 7.

Видим, что увеличение min_df и уменьшение max_df приводит к значительному уменьшению размеров признакового пространства. Но чрезмерные ограничения приводят к потере качества на валидационной выборке.

Оптимальные параметры:

•Представление BagOfWords

•min_df=10, без ограничения сверху

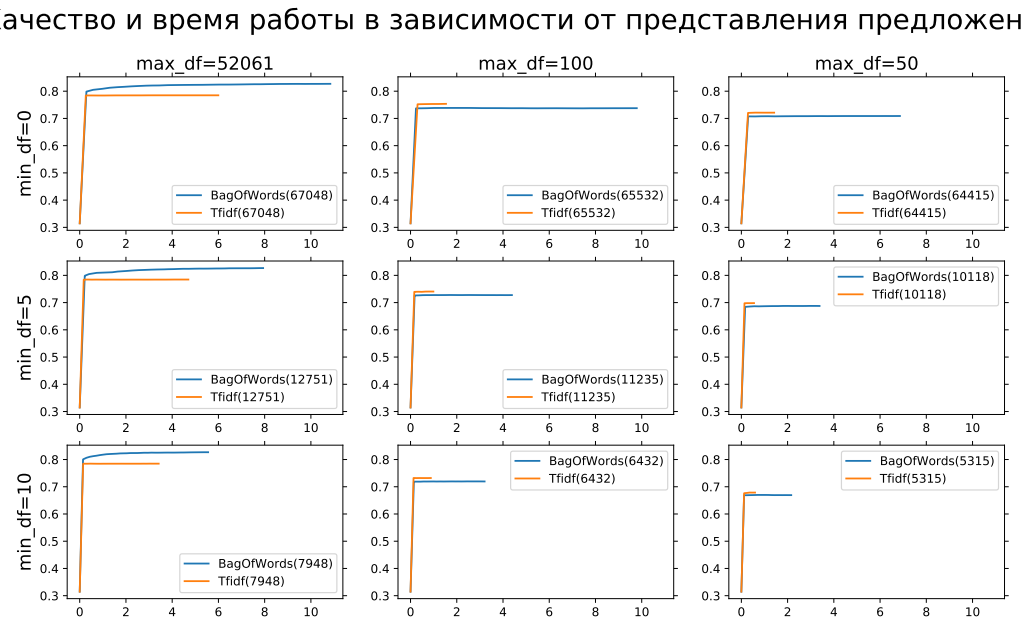

Рис. 7: Сравнение качества алгоритма при разных способах представления данных. В скобках в легенде записаны соответсвующие размеры признакового пространства.

Эксперимент №6

Применим полученный лучший алгоритм для тестовой выборки.

Полученная точность: 0.82564

Полученная точность сравнима с точностью на валидационной выборке.

Проанализируем ошибки алгоритма.

Из объектов, на которых были допущены ошибки, можно выделить несколько общих черт: 1) Комментарий написан не на английском языке.

2) Комментарий состоит из набора символов, не несущей смысловой нагрузки.

Пример:-= hrthrtdghrsdtghtrsdhtrhdgthjrtgh == ffffffffffffffffffffffffffffff..."

Верный ответ: False

Ответ классификатора: True

3) Длинный токсичный комментарий без явной агрессии и токсичных слов.

Пример:-= stop being an ass ==

It seems you are a German in love with your German magazine, thats just great. One magazine has times, the rest of the world has times, dont try to suggest that your magazines times are more important. The rest of the world doesnt really care about your magazine, they just care about times."

Верный ответ: True

Ответ классификатора: False

4) Не токсичный комментарий, в котором есть токсичные слова.

Пример:-= black mamba ==

It.is ponious snake of the word and but it not kills many people but king cobra kills

many people in India"

Верный ответ: False

Ответ классификатора: True

Бонусные задания

1) В признаковое пространство были добавлены n-граммы и исследовано их влияние на качество и скорость алгоритма.

Результаты исследований представлены на рисунке 8.

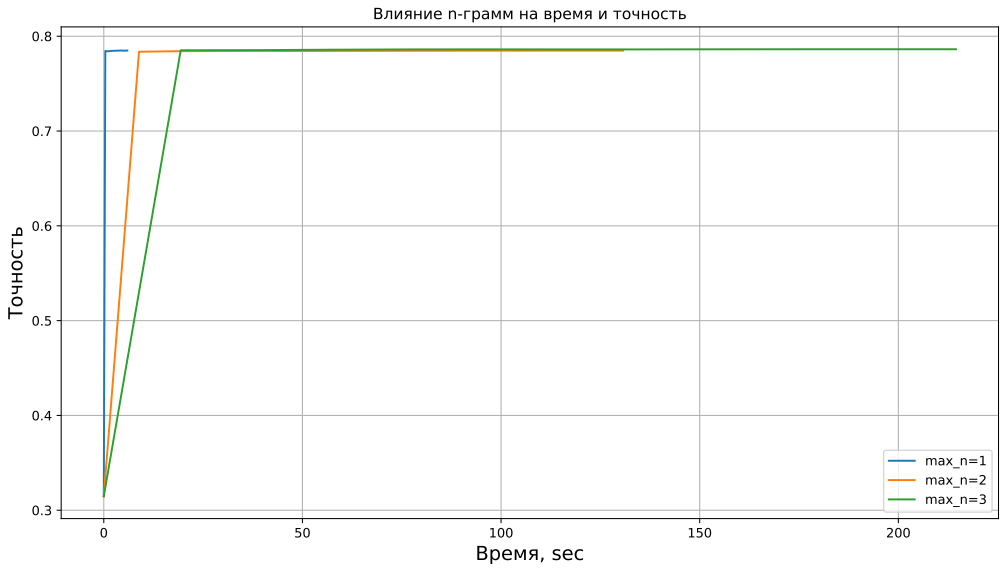

Рис. 8: Использование n-грамм.

max_n=1: время=5.9 sec, точность=0.78487

max_n=2: время=130.7 sec, точность=0.78482

max_n=3: время=214.4 sec, точность=0.78627

Заметим, что использование 2-грамм не улучшило качество, однако использование 3-грамм заметно увеличило качество алгоритма. При этом время работы алгоритма возросло в десятки раз.

2) Реализован режим работы алгоритма SGDClassifier, при котором вся обучающая выборка не хранится в оперативной памяти. Реализован требуемый итератор и специальный режим обучения алгоритма стохастического градиентного спуска.

Код представлен в optimization.py

Выводы

По итогам задания были изучены линейные модели. Были реализованы и исследованы GD и SGD.

Результаты экспериментов продемонстрировали, что с помощью кросс-валидации и аугментации выборки можно добиться лучших показателей на тестовой выборке.