Кузьмин Никита, ММП, 317.

Отчет по практическому заданию №2 "**Применение линейных моделей для определения токсичности комментария**".
Логистическая регрессия и градиентный спуск.

Содержание

1 Введение

В данном документе представлен отчет о проделанных экспериментах по практическому заданию №2, анализ результатов. Краткое описание задания: необходимо реализовать линейный классификатор с произвольной функцией потерь.

2 Теория

3 Эксперименты

В этом блоке приведены все обязательные эксперименты, которые изложены в формулировке задания. Все эксперименты проводились на упрощенном датасете (рассматривается задача бинарной классификации) из соревнования **Toxic Comment Classification Challenge**, в котором нужно определить токсичность комментария.

Стандартный дизайн эксперимента:

- Оценка качества и подбор параметров модели проводились на каждой эпохе с помощью отложенной тренировочной выборки (30%). Все графики ниже построены по значениям accuracy, посчитанным на отложенной выборке.
- В тренировочную выборку был добавлен признак, состоящий из всех единиц, который позволяет учитывать смещение (**bias**). Было решено не использовать смещение в $L2$-регуляризации, чтобы даже при плохом выборе коэффициента регуляризации решающая гиперплоскость не вырождалась в 0.
- В стохастическом градиентном спуске проверяется критерий останова на каждой эпохе (не итерации).

3.1 Исследование поведения градиентного спуска

Обновления весов модели при использовании градиетного спуска происходит по следующей формуле:

$$w_t = w_{t-1} - \frac{\alpha}{t^\beta} \times \frac{1}{N} \times \sum_{i=1}^{N} \nabla_w \mathcal{L}(x_i, y_i | w_{t-1}), \tag{1}$$

где t - номер итерации, β - **step_beta**, $\nabla_w \mathcal{L}(x_i, y_i | w_{t-1})$ - градиент функции потерь.

3.1.1 Параметр размера шага step_alpha

Параметр **step_alpha** (α) используется в градиентном спуске при обновлении весов в формуле 1. Рассмотрим следующие зависимости при разных значениях параметра **step_alpha**:

1. зависимость значения функции потерь от реального времени работы метода
2. зависимость точности (accuracy) от реального времени работы метода
3. зависимость значения функции потерь от итерации метода
4. зависимость точности (accuracy) от итерации метода

Соответствующие графики приведены на: рис. 1, 2, 3, 4.

Рис. 1: Зависимость значения функции потерь от реального времени работы градиентного спуска

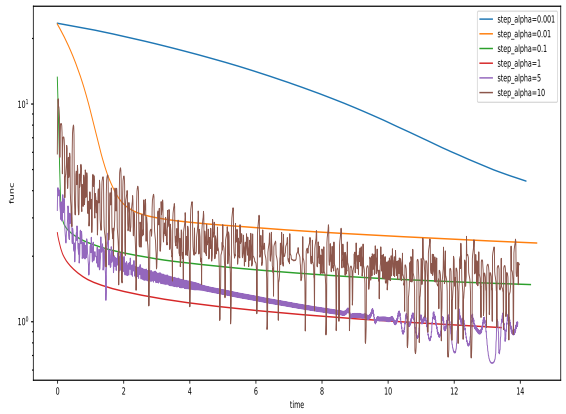


Рис. 2: Зависимость значения точности (accuracy) от реального времени работы градиентного спуска

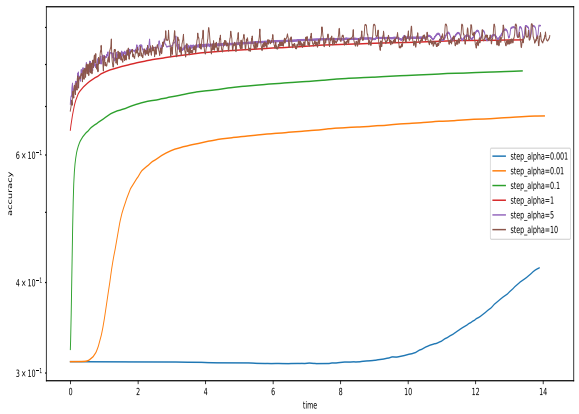


Рис. 3: Зависимость значения функции потерь от итерации метода градиентного спуска

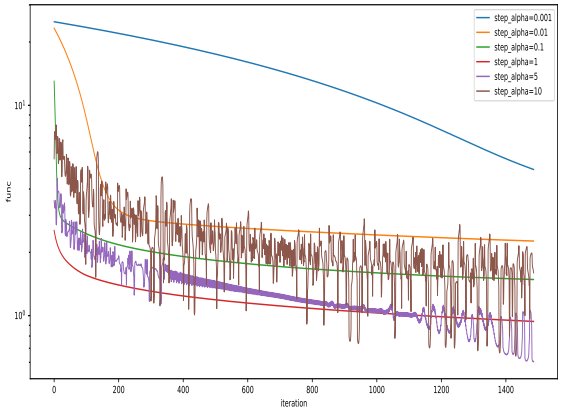
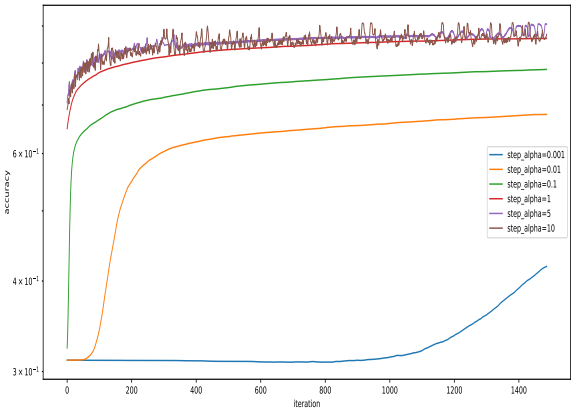


Рис. 4: Зависимость значения точности (accuracy) от итерации метода градиентного спуска



Из графиков видно, что при значениях α, близких к нулю алгоритму нужно больше времени для сходимости. С другой стороны, если значения слишком большие, то алгоритм становится крайне не стабильным.

3.1.2 Параметр размера шага step_beta

Параметр **step_beta** (β) используется в градиентном спуске при обновлении весов в формуле 1. Аналогично предыдущему пункту рассмотрим зависимости из 3.1.1 при разных значениях параметра **step_beta** и проанализруем соответсвующие графики, представленные на рис. 5, 6, 7, 8.

Рис. 5: Зависимость значения функции потерь от реального времени работы градиентного спуска

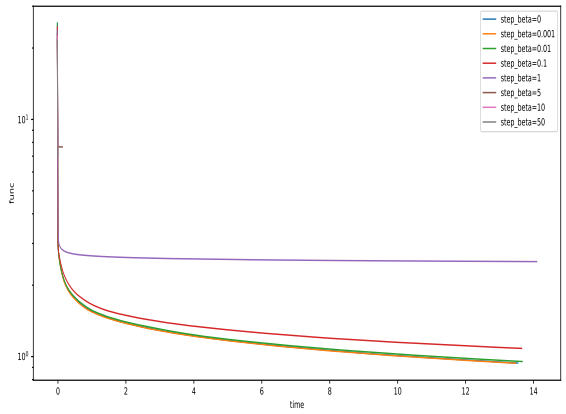


Рис. 6: Зависимость значения точности (accuracy) от реального времени работы градиентного спуска

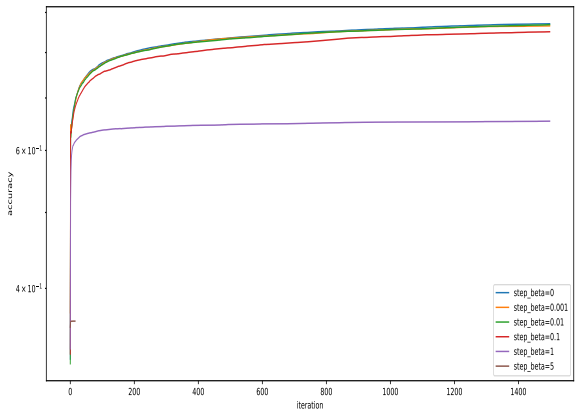


Рис. 7: Зависимость значения функции потерь от итерации метода градиентного спуска

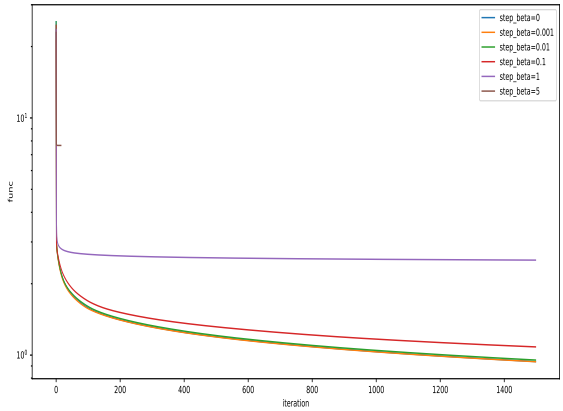
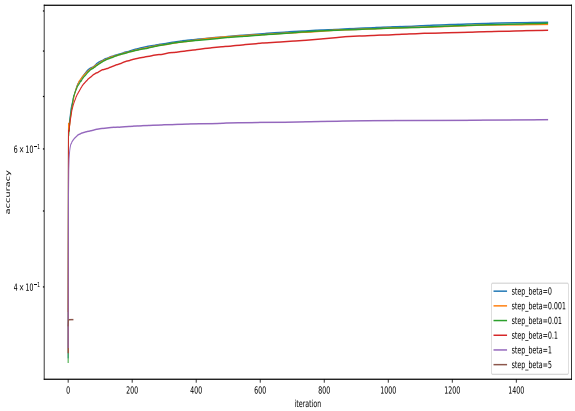


Рис. 8: Зависимость значения точности (accuracy) от итерации метода градиентного спуска



Из графиков можно заметить, что значения **step_beta**, близкие к 0, приводят к одинаковому качеству. С увеличением же параметра **step_beta** значение точности уменьшается. При **step_beta** $= 5$ изменение функции потерь так мало, что критерий остановки срабатывает до первых 200 итераций.

3.1.3 Начальное приближение w_0

Начальное приближение нужно для инициализации весов модели. В данной работе были рассмотрены следующие варианты задания w_0:

- нулевой вектор
- вектор с координатами из $U(0, 1)$
- вектор с координатами из $U(100, 500)$
- вектор с координатами из $U(1000, 5000)$
- вектор с координатами из $U(10000, 50000)$
- вектор с координатами из $N(0, 1)$
- вектор с координатами из $N(0.5, 0.5)$

Графики зависимостей 3.1.1 представлены на рис. 9, 10, 11, 12.

Рис. 9: Зависимость значения функции потерь от реального времени работы стохастического градиентного спуска

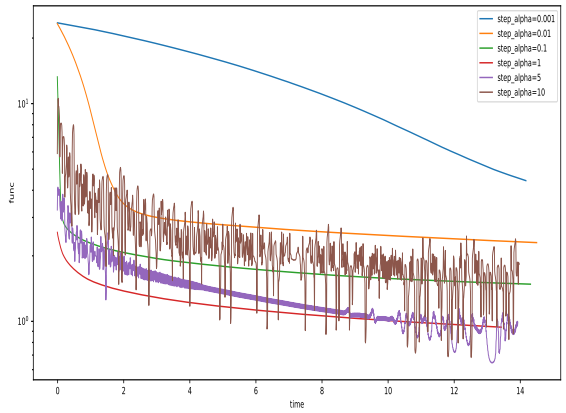


Рис. 10: Зависимость значения точности (accuracy) от реального времени работы стохастического градиентного спуска

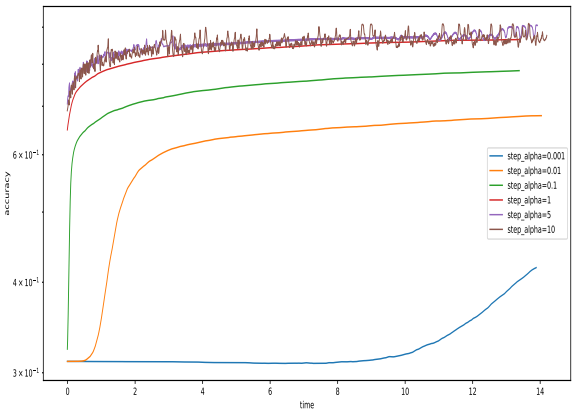


Рис. 11: Зависимость значения функции потерь от итерации метода стохастического градиентного спуска

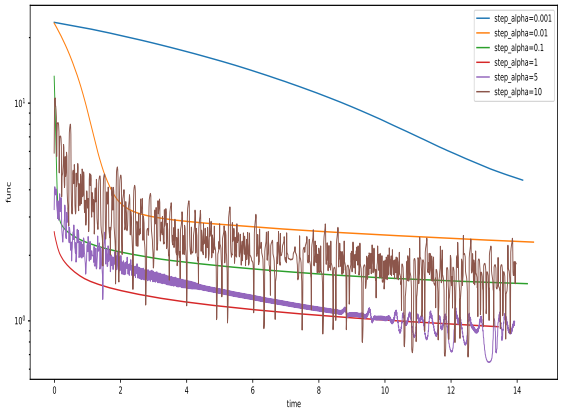
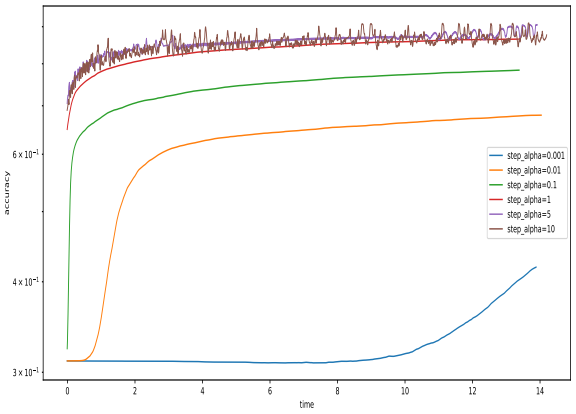


Рис. 12: Зависимость значения точности (accuracy) от итерации метода стохастического градиентного спуска



3.2 Исследование поведения стохастического градиентного спуска

Обновления весов модели при использовании стохастического градиетного спуска происходит по следующей формуле:

$$w_t = w_{t-1} - \frac{\alpha}{t^\beta} \times \frac{1}{|I|} \times \sum_{i \in I} \nabla_w \mathcal{L}(x_i, y_i | w_{t-1}), \tag{2}$$

где t - номер итерации, β - **step_beta**, I - некоторое подможнество индексов тренировочной выборки, $\nabla_w \mathcal{L}(x_i, y_i | w_{t-1})$ - градиент функции потерь.

3.2.1 Параметр размера шага step_alpha

Параметр **step_alpha** (α) используется в стохастическом градиентном спуске при обновлении весов в формуле 2. Рассмотрим следующие зависимости при разных значениях параметра **step_alpha**:

1. зависимость значения функции потерь от реального времени работы метода
2. зависимость точности (accuracy) от реального времени работы метода
3. зависимость значения функции потерь от эпохи метода
4. зависимость точности (accuracy) от эпохи метода

Соответствующие графики приведены на: рис. 17, 18, 19, 20.

Рис. 13: Зависимость значения функции потерь от реального времени работы градиентного спуска

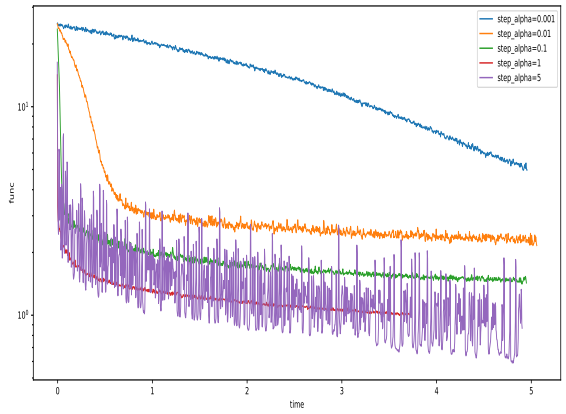


Рис. 14: Зависимость значения точности (accuracy) от реального времени работы градиентного спуска

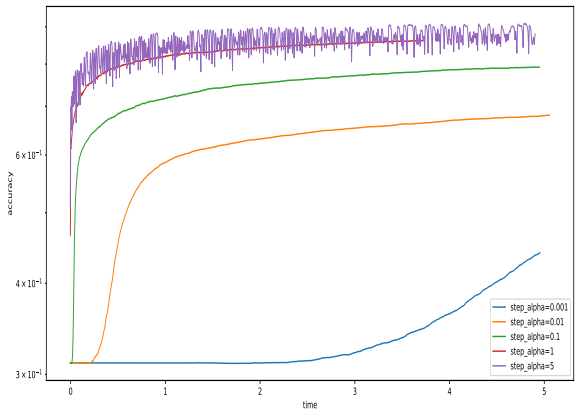


Рис. 15: Зависимость значения функции потерь от эпохи метода градиентного спуска

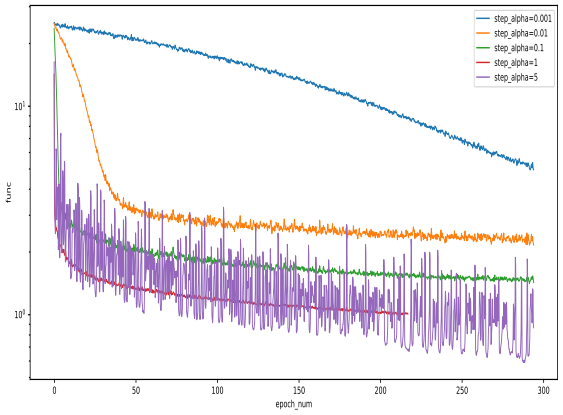
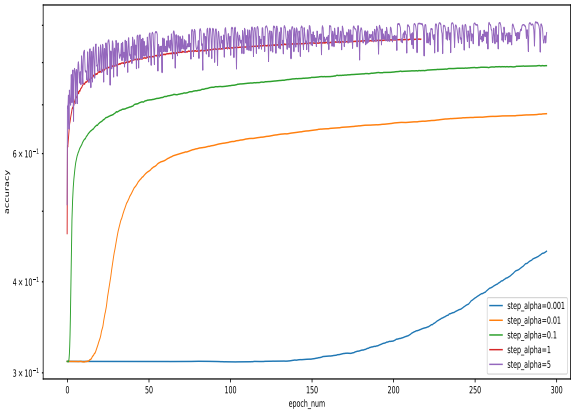


Рис. 16: Зависимость значения точности (accuracy) от эпохи метода градиентного спуска



На графиках просматривается аналогичная ситуация, что и с градиентным спуском: при значениях **step_alpha**, близких к 0 возникает эффект недообучения, а при больших - появляется нестабильность кривой обучения, но есть точки, в которых достигается наивысшая точность. В таком случае можно производить сохранение весов модели на итерации со значением наилучшей точности. Возможно, такой метод поможет справиться с нестабильностью.

3.2.2 Параметр размера шага step_beta

Параметр **step_beta** (β) используется в градиентном спуске при обновлении весов в формуле 1. Аналогично предыдущему пункту рассмотрим зависимости из 3.2.1 при разных значениях параметра **step_beta** и проанализируем соответсвующие графики, представленные на рис. **??**, **??**, **??**, **??**.

Рис. 17: Зависимость значения функции потерь от реального времени работы градиентного спуска

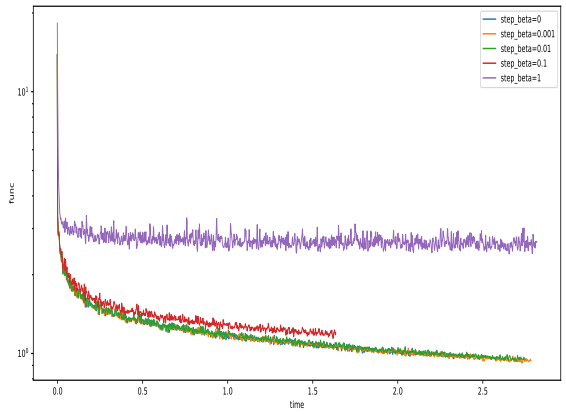


Рис. 18: Зависимость значения точности (accuracy) от реального времени работы градиентного спуска

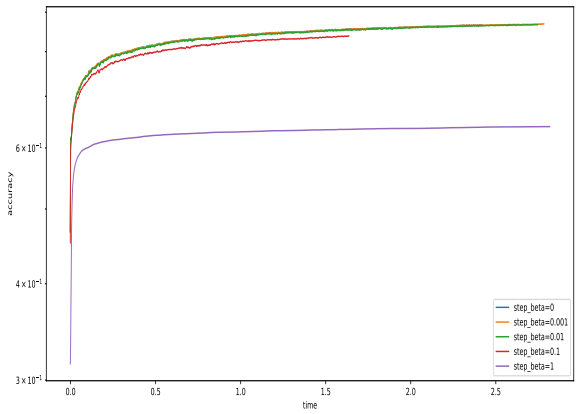


Рис. 19: Зависимость значения функции потерь от эпохи метода градиентного спуска

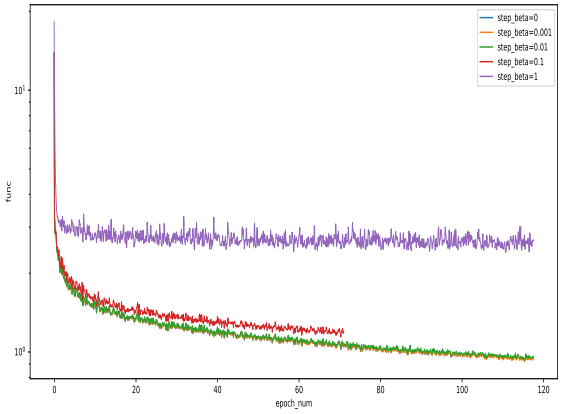
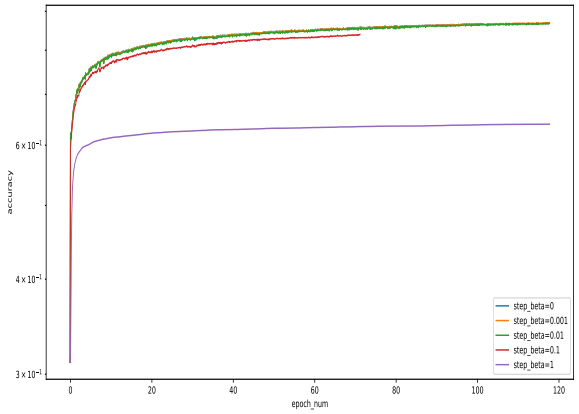


Рис. 20: Зависимость значения точности (accuracy) от эпохи метода градиентного спуска



Аналогично обычному градиентному спуску - с увелничением значения **step_beta** увеличивается значение функции потерь и уменьшается точность (accuracy).

3.2.3 Размер подвыборки batch_size

Размер подвыборки определяет количество элементов тренировочной выборки, которые будут использованы для подсчета градиента.

Соответствующие графики приведены на рис. 21, 22, 23, 24.

Рис. 21: Зависимость значения функции потерь от реального времени работы градиентного спуска

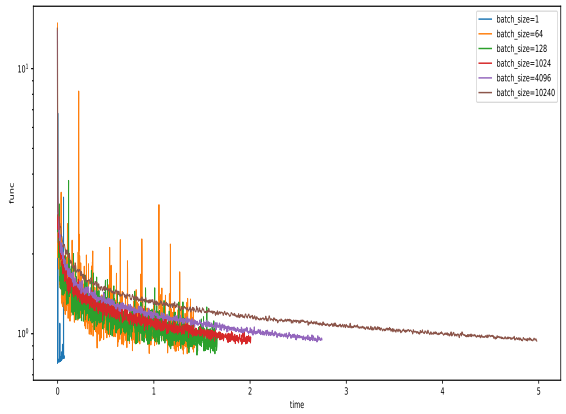


Рис. 22: Зависимость значения точности (accuracy) от реального времени работы градиентного спуска

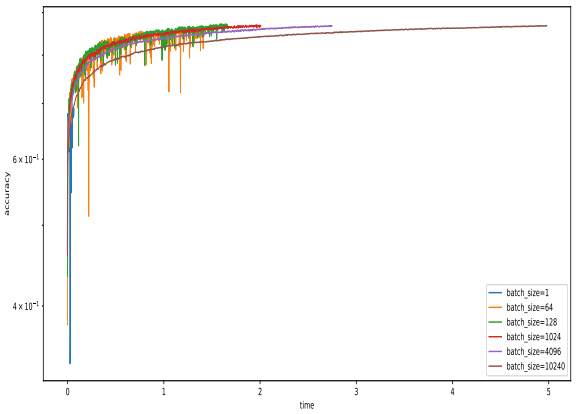


Рис. 23: Зависимость значения функции потерь от эпохи метода градиентного спуска

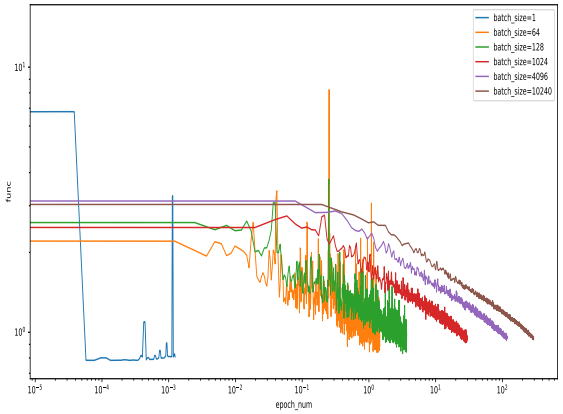
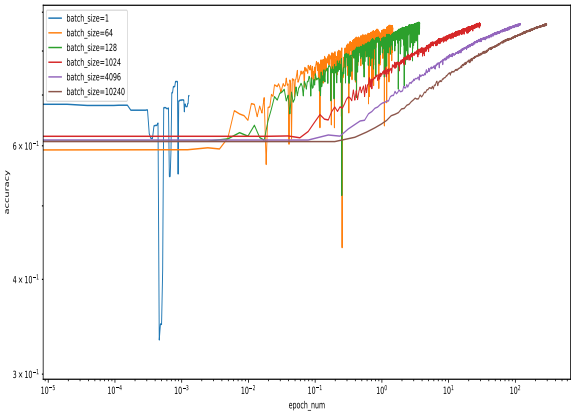


Рис. 24: Зависимость значения точности (accuracy) от эпохи метода градиентного спуска



Можно заметить сильное увеличение дисперсии значений функции потерь с уменьшением размера подвыборки. Это объясняется тем, что при изменении весов на каждой итерации используется подвыборка для вычисления градиента, то есть приближенное значение, которое может сильно меняться от итерации к итерации, если размер подвыборки достаточно маленький.

3.2.4 Начальное приближение w_0

Начальное приближение нужно для инициализации весов модели. В данной работе были рассмотрены следующие варианты задания w_0:

- нулевой вектор
- вектор с координатами из $U(0, 1)$
- вектор с координатами из $U(100, 500)$
- вектор с координатами из $U(1000, 5000)$
- вектор с координатами из $U(10000, 50000)$
- вектор с координатами из $N(0, 1)$

- вектор с координатами из $N(0.5, 0.5)$

Графики зависимостей 3.2.1 представлены на рис. 25, 26, 27, 28.

Рис. 25: Зависимость значения функции потерь от реального времени работы градиентного спуска

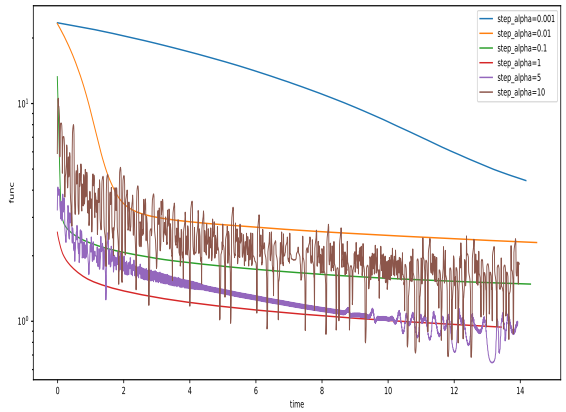


Рис. 26: Зависимость значения точности (accuracy) от реального времени работы градиентного спуска

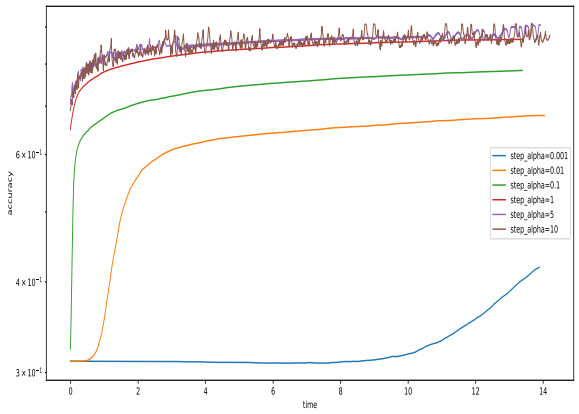


Рис. 27: Зависимость значения функции потерь от эпохи метода градиентного спуска

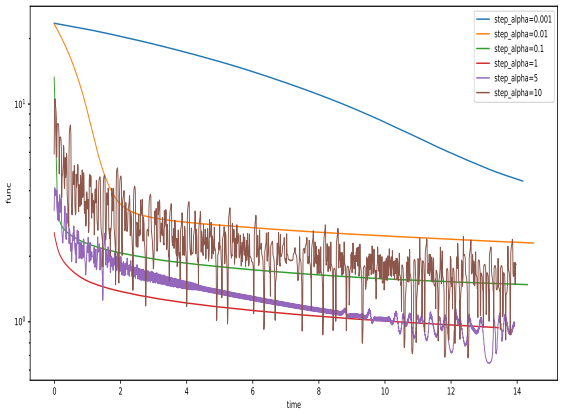
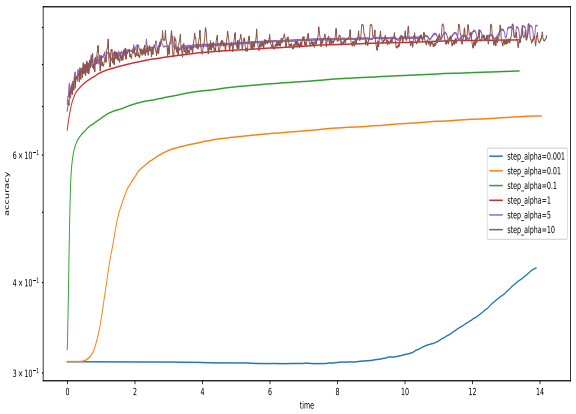


Рис. 28: Зависимость значения точности (accuracy) от эпохи метода градиентного спуска



Из графиков видно, что лучшим вариантом инициализации весов оказался нулевой вектор.

3.3 Сравнение градиентного спуска и стохастического градиентного спуска

В данном разделе проведено сравнение методов по трем характеристикам:

- значения функции потерь
- точность (accuracy)
- время работы метода

Результаты экспериментов приведены на рис. 29, 30 и в таблице 1

Таблица 1: Среднее время работы методов оптимизации

Метод	Среднее время работы, с
Градиентный спуск	8.852
Стохастический градиентный спуск	1.964

Рис. 29: Зависимость значения функции потерь от реального времени работы обычного и стохастического градиентного спуска

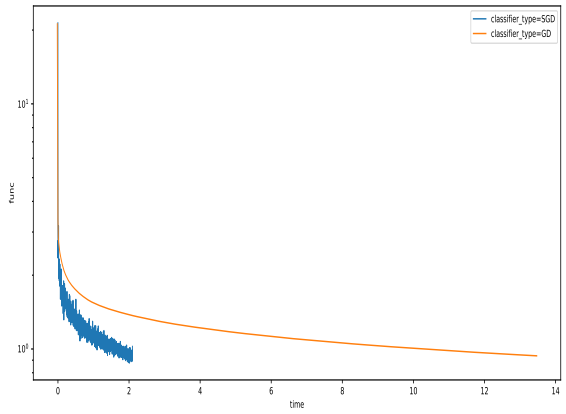
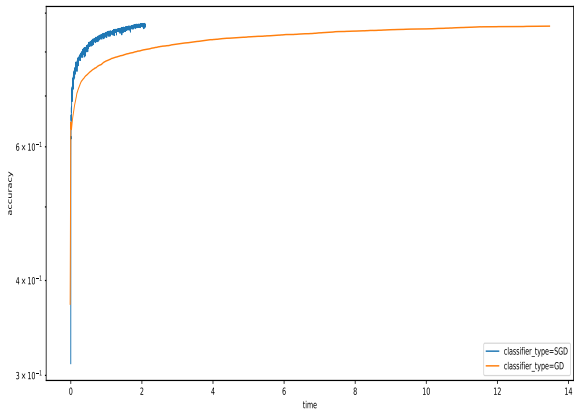


Рис. 30: Зависимость значения точности (accuracy) от реального времени работы обычного и стохастического градиентного спуска



Из приведенных выше графиков видно, что стохастический градиентный спуск сходится быстрее, причем к более хорошему оптимуму, чем обычный градиентный спуск, но сходимость с бóльшей дисперсией значений функции потерь. Таблица 1 показывает, что стохастический градиентный спуск работает в несколько раз быстрее.

3.4 Лемматизация и удаление стоп-слов

3.4.1 Простая лемматизация с удалением стоп-слов

В этом эксперименте проводится лемматизация, в которой считается, что все слова в тексте - глаголы. Удаляются все стоп-слова. В таблице **??** приведено сравнение метода градиентого спуска до преобразования текста и после.

3.4.2 Лемматизация, учитывающая части речи, с удалением стоп-слов

В этом эксперименте проводится лемматизация, в которой учитываются части речи слов в тексте. Удаляются все стоп-слова. В таблице **??** приведено сравнение метода градиентого спуска до преобразования текста и после.

3.5 Сравнение представлений BagOfWords и TF-IDF с различными параметрами

4 Применение лучших алгоритмов с каждого эксперимента к тестовой выборке