Методы оптимизации. Лабораторная работа 1

Daniil Bakushkin

24 апреля 2023 г.

Матрично-векторная форма для бинарной логистической регрессии, ее производной и гессиана

Функция логистической регрессии

n – feautres, m – samples
$b = \{-1, 1\}^m$
X – samples / features matrix
w – weights of the model
$lambda$ – regularization coefficient

$$f(w) = \frac{1}{m}\langle(1_m + exp(-b \odot \langle X; w\rangle)); I_m\rangle + \frac{1}{2}\|x\|_2^2$$

Градиент логистической регрессии

$$\nabla f(w) = -\frac{1}{m}X^T\left(b \odot \left(1 - \frac{1}{1 + exp(-b \odot Xw)}\right)\right) + \|x\|_2$$

Гессиан логистической регрессии

Q – diagonal matrix with elements b_i $q_i = \sigma(-y_i w^T x_i) \cdot (1 - \sigma(-y_i w^T x_i))$

$$\nabla^2 f(w) = \frac{1}{N}X^T Q X + \lambda I_n$$

Эксперимент: Траектория градиентного спуска на квадратичной функции

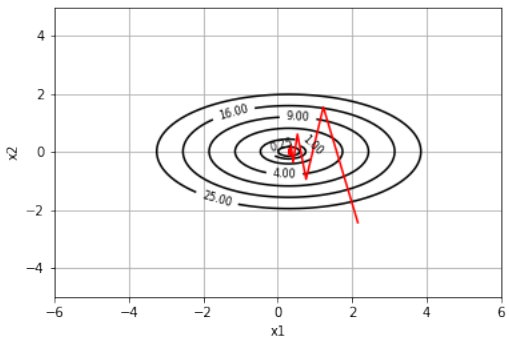
Используемые квадратичные функции:

1.
$$\frac{1}{2}x^T Ax - bx$$

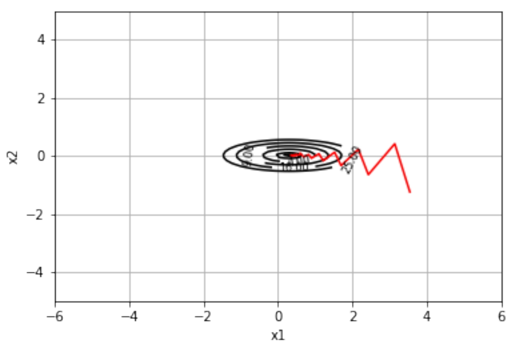
2.
$$\frac{1}{2}\|Ax - b\|^2$$

Траектории градиентного спуска для функций. По осям – параметры модели $x_1,\ x_2$.

Траектория первой функции: значения параметров, которые выбирались на шагах оптимизации.



Траектория второй функции: значения параметров, которые выбирались на шагах оптимизации.



В задании мы оптимизируем функцию лосса для квадратичной функции с помощью градиентного спуска.

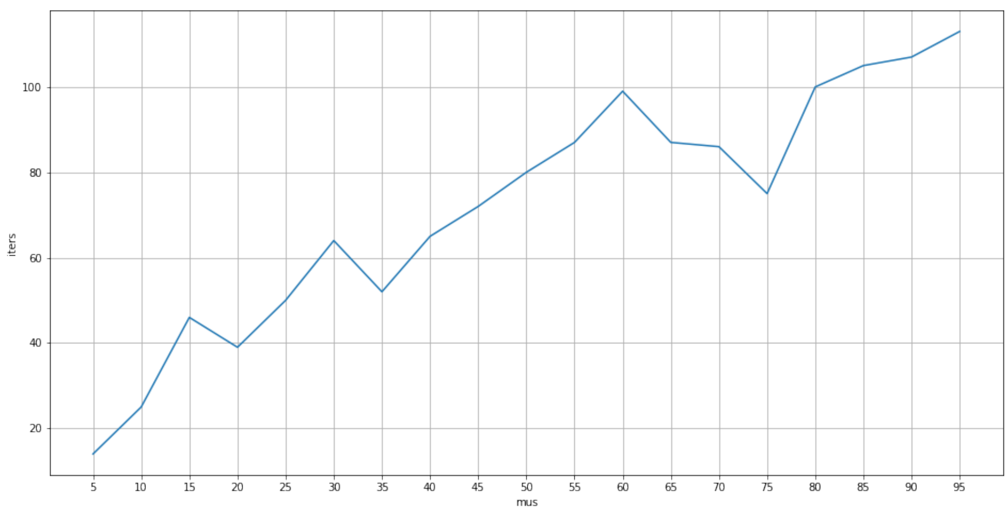
1. для обоих методов используется Armijo с дефолтными параметрами
2. коэффициент обусловленности – 20
3. Начальные точки генерируются случайно из нормального распределния.

4. Матрица генерируется с заданным числом обусловленности согласно схеме из описания эксперимента 3.2

Первому методу понадобилось 15 итераций, чтобы сойтись, второму – 22.

Как отличается поведение метода в зависимости от числа обусловленности функции?

Для проверки были сгенерированы матрицы с коэффициентами обусловленности от 5 до 100 и отрисован график с количеством итераций, необходимым для сходимости метода. Использовалась квадратичная функция 1.

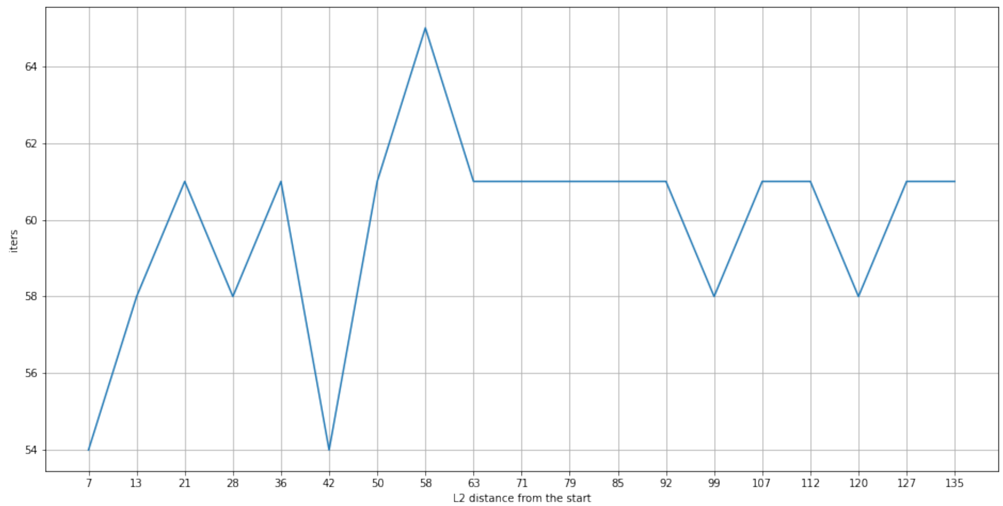


По оси абсцисс – число обусловленности, по оси ординат – количество итераций до схождения.

Вывод: количество итераций градиентного спуска растет линейно в зависимости от числа обусловленности функции.

Как отличается поведение метода в зависимости от выбора начальной точки?

Для проверки были сгенерированы начальные точки с различным L2 расстоянием до оптимума и отрисован график с количеством итераций, необходимым для сходимости метода. Использовалась квадратичная функция 1.

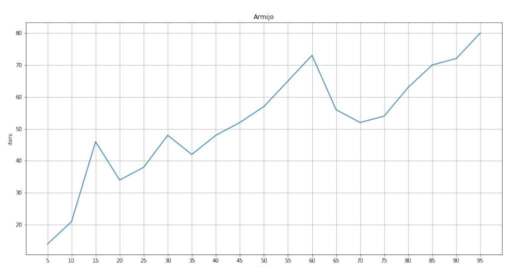


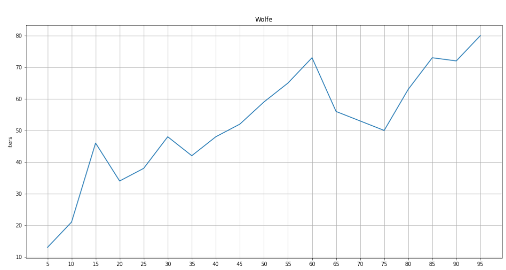
По оси абсцисс – L2-расстояние, по оси ординат – количество итераций до схождения.

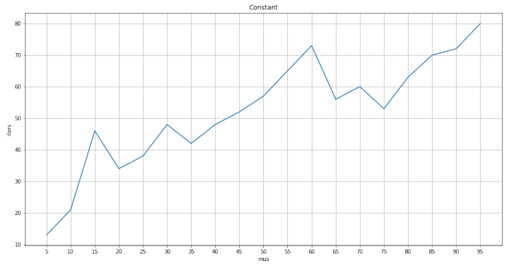
Вывод: количество итераций градиентного спуска не зависит от расстояния до точки оптимума.

Как отличается поведение метода в зависимости от метода подбора шага?

Для проверки были сгенерированы итерации для разных чисел обусловленности с разными методами (все методы брались с дефолтными парамтерами).







По оси абсцисс – число обусловленности, по оси ординат – количество итераций до схождения.

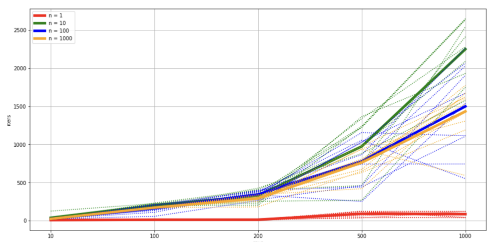
Вывод: как можно заметить, методы не сильно отличаются в количестве итераций до схождения, но во время экспериментов на других типах функций было выявлено, что метод подбора шага с константой очень часто не сходится вообще. Армихо и Вульф всегда работали +- одинаково.

Эксперимент: Зависимость числа итераций градиентного спуска от числа обусловленности и размерности пространства

В данной задаче было необходимо проверить, как зависит количество итераций необходимое для сходимости градиентного спуска от числа обусловленности функции, а также посмотреть, какая существует зависимость от размерности пространства.

1. для всех экспериментов используется Armijo с дефолтными параметрами
2. коэффициент обусловленности – от 5 до 1000
3. размерность пространства – от 10 до 10000
4. Начальные точки генерируются случайно из $U \sim [1, mu]$, где mu – число обусловленности
5. Матрица генерируется с заданным числом обусловленности согласно схеме из описания эксперимента 3.2

В задании мы оптимизируем функцию лосса для квадратичной функции с помощью градиентного спуска.



По оси абсцисс – число обусловленности, по оси ординат – количество итераций до схождения. В легенде подписаны размерности пространства.

Вывод: градиентный спуск растет линейно в зависимости от числа обусловленности. Размерность пространства влияет на результаты незначительно.

Эксперимент: Сравнение методов градиентного спуска и Ньютона на реальной задаче логистической регрессии

Теоретическая выкладка:

n – размерность пространства q – количество примеров

method	GD	Netwon
memory	$O(n)$	$O(n^2)$
time	$O(n)$	$O(n^3)$

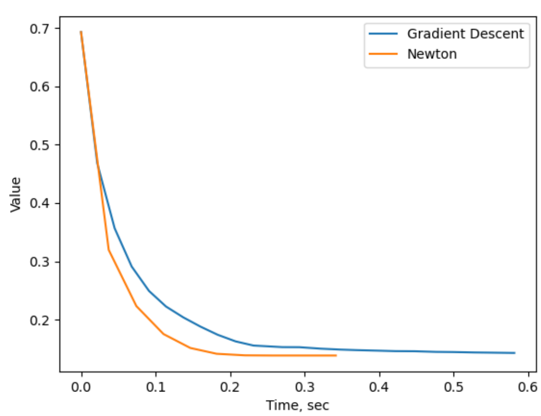
В таблице указано время для метода Ньютона при обращении Гессиана, но существуют более продвинутые методы решения с помощью матричных разложений. Либо другие подходы: сопряженные градиенты – time $O(n^3)$, HFN – time $O(Nd^2) + O(d^3)$.

В задании требуется построить зависимости значения лосса от времени работы для Ньютона и GD.

В задании нужно сравнить методы Ньютона и Градиентного спуска на 3-х различных датасетах.

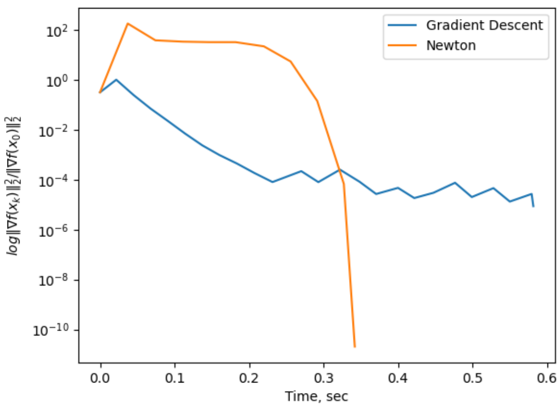
1. для всех экспериментов используется Armijo с дефолтными параметрами
2. начальная точка $x0 = 0$
3. коэффициент регуляризации – $\lambda = \frac{1}{m}$
4. для прогонки использовал 60-80% датасета, потому что считалось очень долго

Датасет **w8a**



По оси абсцисс – время, по оси ординат – значения лосс функции.

На датасете $w8a$ метод Ньютона быстрее уменьшает значение лосс-функции.

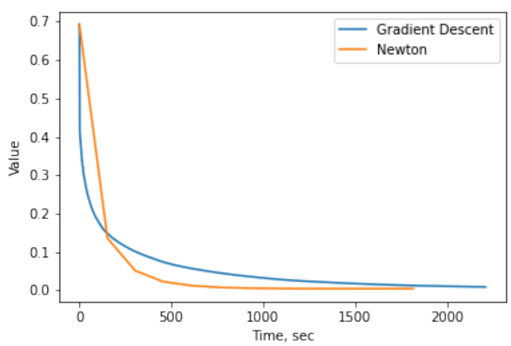


По оси абсцисс – время, по оси ординат – норма невязки.

На датасете $w8a$ у метода GD линейная скорость убывания нормы невязки. У Newton – квадратичная скорость убывания.

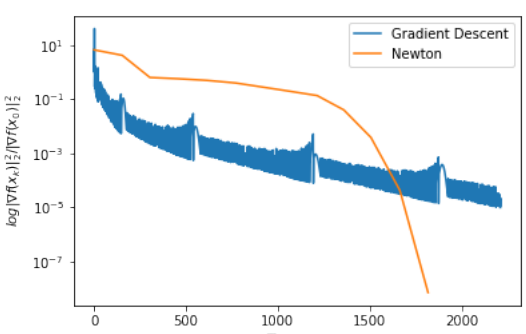
Норма невязки для GD после 10^{-5} начинается уменьшаться очень медленно.

Датасет **gisette_scale**



По оси абсцисс – время, по оси ординат – значения лосс функции.

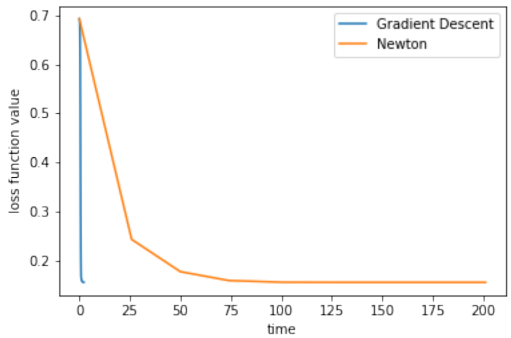
На датасете *gisette_scale* метод Ньютона быстрее уменьшает значение лосс-функции.



По оси абсцисс – время, по оси ординат – норма невязки.

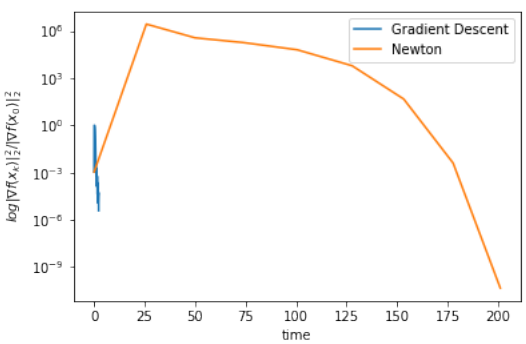
На датасете $gisette\_scale$ интересная сходимость GD, возможно, связано со структурой датасета. Ньютон так же сходится квадратично.

Датасет $real - sim$



По оси абсцисс – число время, по оси ординат – значения лосс функции.

Не знаю, с чем связано такое поведение функции :(



По оси абсцисс – число время, по оси ординат – норма невязки.

Не знаю, с чем связано такое поведение функции :(

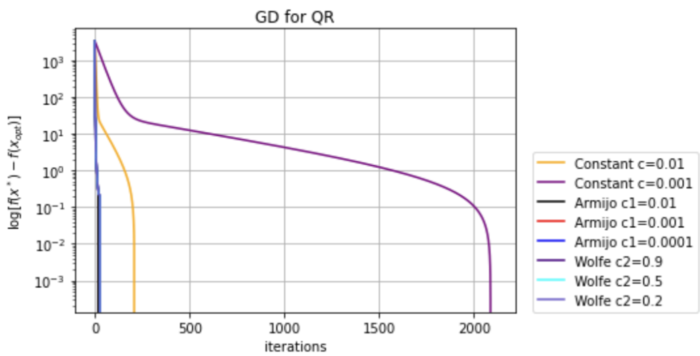
Эксперимент: Стратегия выбора длины шага в градиентном спуске

Необходимо исследовать поведение метода в зависимости от стратегии подбора шага: перебрать различные значения констант.

Использовать для теста квадратичную функцию и логистическую регрессию.

1. коэффициент обусловленности – 20
2. Начальные точки генерируются случайно из нормального распределния.
3. Матрицы подбирались произвольно.

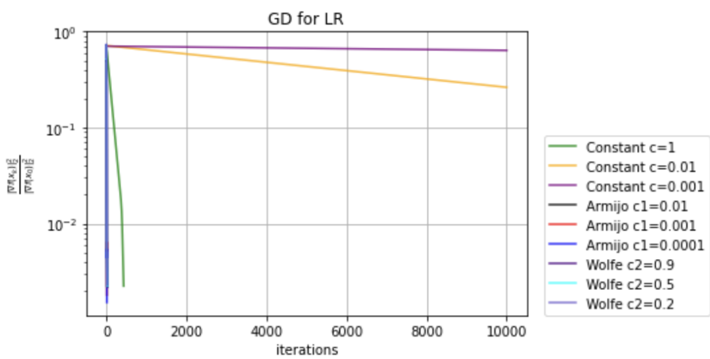
Перебор параметров для градиентного спуска для квадратичной функции



По оси абсцисс – кол-во итераций, по оси ординат – относительная невязка по функции в логарифмической шкале.

Можно сделать вывод, что процедуры бэктрекинга / wolfe_line_search в зависимости от итераций дают примерно одинаковое уменьшение нормы относительной невязки. Константная стратегия сильно уступает, причем есть зависимость от значения константы. Слишком большое значение константы не дает оптимизатору сделать маленький шаг в близи точки оптимума.

Перебор параметров для градиентного спуска для логистической регрессии

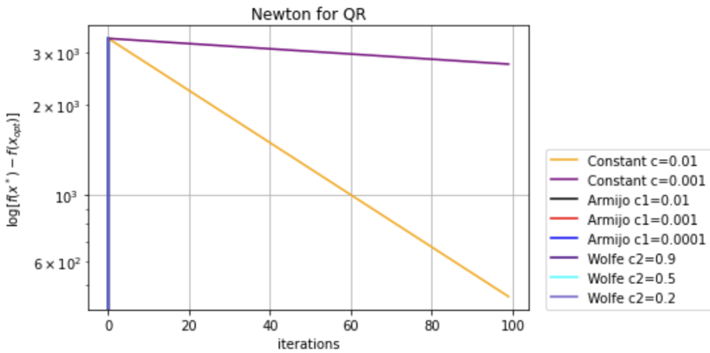


По оси абсцисс – кол-во итераций, по оси ординат – относительный квадрат нормы градиента в логарифмической шкале.

Методы бэктрекинга / wolfe_line_search сходятся быстро, константа – медленно.

Эксперимент: Стратегия выбора длины шага в методе Ньютона

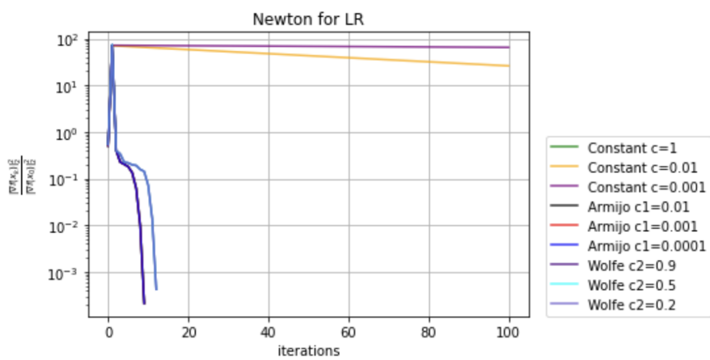
Перебор параметров для метода Ньютона для квадратичной функции



По оси абсцисс – кол-во итераций, по оси ординат – относительная невязка по функции в логарифмической шкале.

Теория говорит, что для квадартичной функции метод Ньютона должен сходиться за 1 шаг, т.к. мы приближаем целевую функцию квадратичной, у них оптимумы должны совпасть, и мы должны попасть в оптимум за 1 шаг.

Перебор параметров для метода Ньютона для логистической регрессии



По оси абсцисс – число обусловленности, по оси ординат – относительный квадрат нормы градиента в логарифмической шкале.

Теория говорит, что для метода ньютона мы должны выбрать изначально единичный c_2, для того, чтобы обеспечить квадратичную сходимость в окрестности оптимума, и если необходимо, то алгоритм подбора шага, уменьшит наш шаг, поэтому нет особого смысла что-то перебирать.

Приложение

Получил ценные наставления по выводу градиента логистической регрессии от Артемия Мосейчука. Вывод гессиана был на лекции :)