Градиентный спуск и метод Ньютона

Артём Заварзин

Ноябрь 2021

1 Матричные производные логистической регрессии

1. Значение функции

$$f(x) = \frac{< \log\left(1 + e^{-b \cdot Ax}\right), 1_m >}{m} + \frac{\lambda}{2} < x, x >$$

2. Производная

$$\nabla f(x) = -\frac{1}{m} A^T \left(\frac{b}{1_m + e^b}\right)$$

3. Гессиан

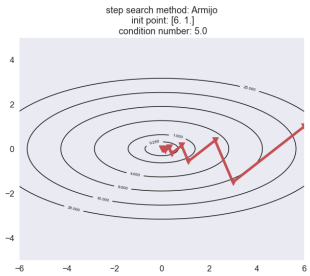
$$sigm = \frac{b}{1_m + e^b}$$

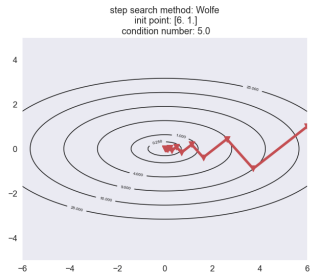
$$\nabla^2 f(x) = \frac{1}{m} A^T ((sigm \cdot (1_m - sigm)) I_m) A + \lambda I_n$$

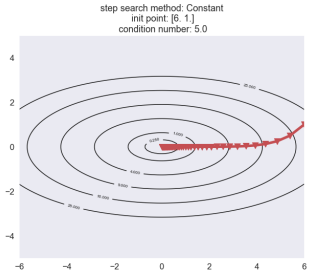
2 Траектория градиентного спуска на квадратичной функции

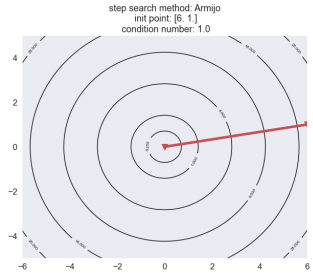
2.1 Зависимость от числа обусловленности

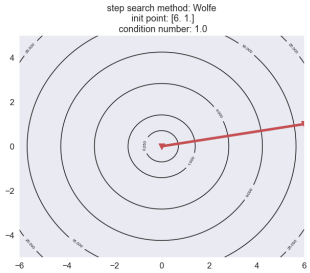
Чем больше число обусловленности тем хуже сходится функция, так как оси эллипса не пропорциональны и для данных методов находится не оптимальный шаг.

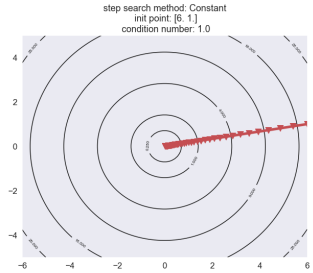

step search method: Armijo
init point: [6. 1.]
condition number: 5.0


step search method: Wolfe
init point: [6. 1.]
condition number: 5.0


step search method: Constant
init point: [6. 1.]
condition number: 5.0

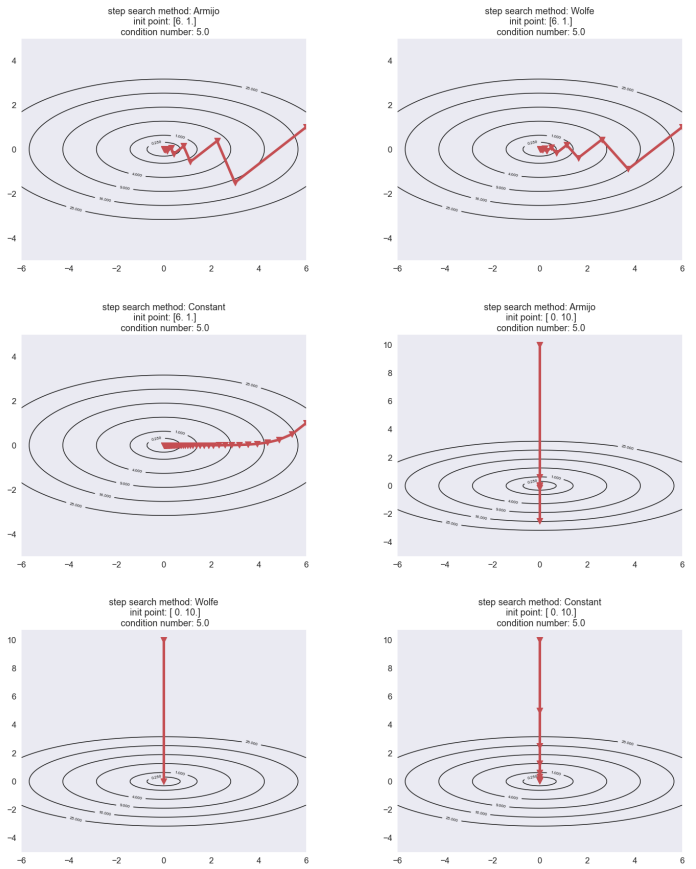

step search method: Armijo
init point: [6. 1.]
condition number: 1.0


step search method: Wolfe
init point: [6. 1.]
condition number: 1.0


step search method: Constant
init point: [6. 1.]
condition number: 1.0

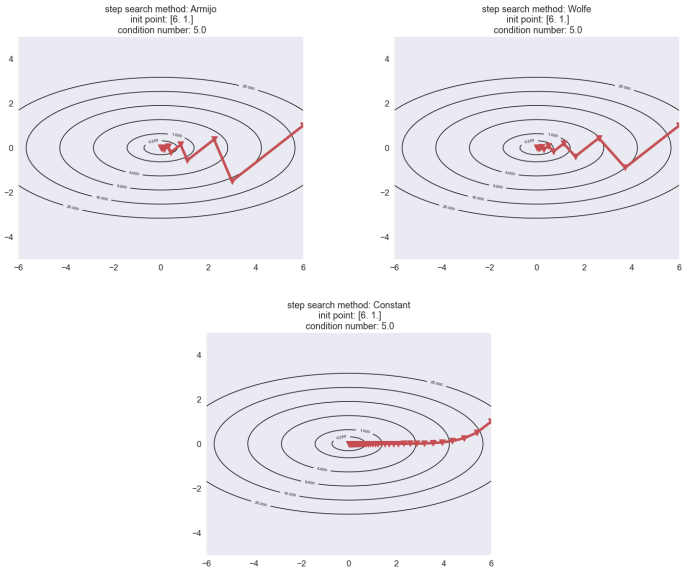
2.2 Зависимость от выбора начального шага

Если точка находится на полюсе эллипса то независимо от других параметров метод будет сходиться оптимально. Если точка находится вне полюса, то если число обусловленности не 1, то сходимость не оптимальна и градиент скачет по линиям уровней.



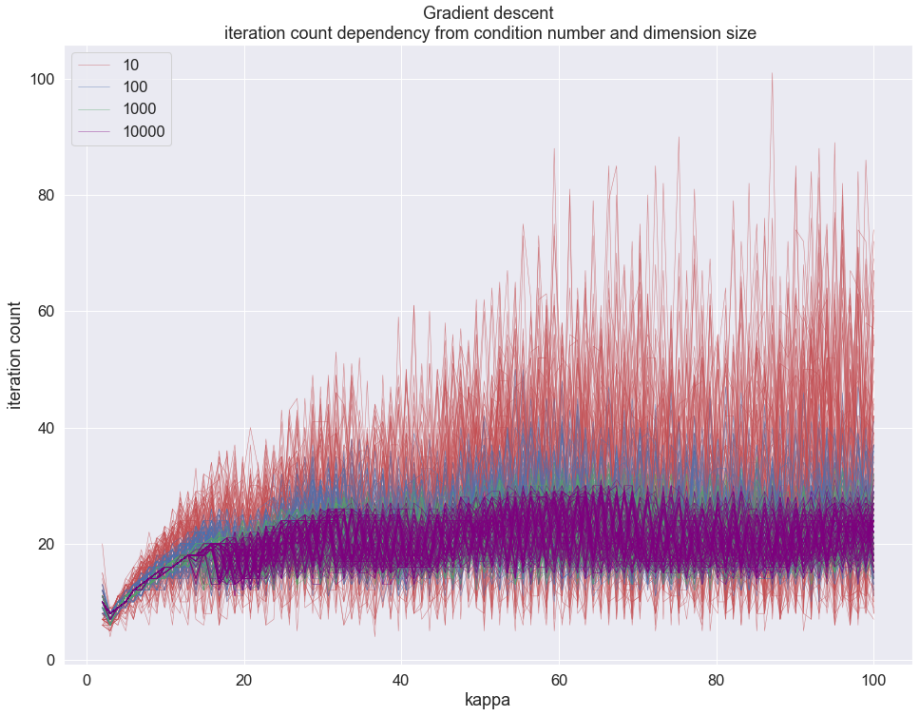
2.3 Зависимость от выбранной стратегии

Константный метод сходится достаточно оптимально, но долго при маленькой константе, при большой константе метод может не сходиться в окрестности оптимума. Методы Армихо и Вульфа сходятся быстро, но не так оптимально и скачут по линиям уровня, Можно заметить, что для метода Вульфа скачки чуть менее разбросаны и направлены ближе к центру.



3 Зависимость числа итераций градиентного спуска от числа обусловленности и размерности пространства

В эксперименте мы провели несколько экспериментов для нахождения зависимости числа итераций от числа обусловленности при разных значениях размерности наблюдения.



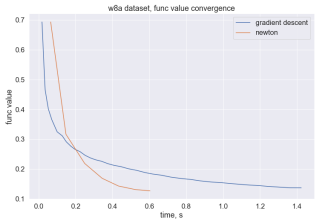
На графике видно, что число итераций монотонно растёт при росте числа обусловленности до 20. Кроме того, можно увидеть, что для размера выборки 10 среднее число итераций сильно выше чем при остальных n. По графику можно так же сказать, что при росте размерности уменьшается дисперсия числа итераций.

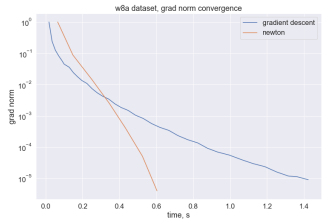
4 Сравнение методов градиентного спуска и Ньютона на реальной задаче логистической регрессии

Сравним методы на 3 датасетах.

4.1 w8a

Размер (49749, 300).

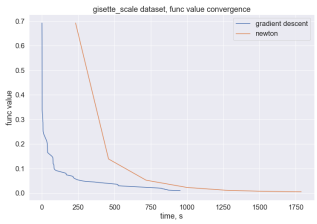


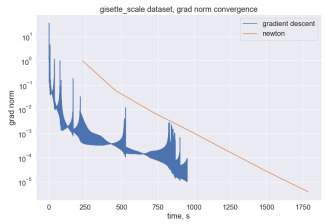


Для данного датасета с небольшой размерностью метод Ньютона сходится быстрее, линейно по норме градиента.

4.2 gisette$_{s}cale$

Размер (6000, 5000)

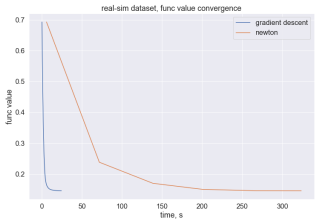


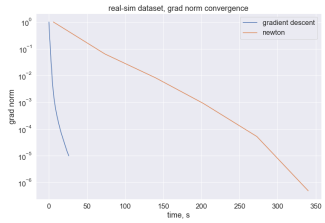


Сходимость по норме градиента по-прежнему линейна для Ньютона, однако градиентный спуск даёт лучшие результаты за счёт того, что стоимость 1 итерации Ньютона сильно возрасла по сравнению с предыдущим результатом.

4.3 real-sim

Размер (72309, 20958)





В этом датасете очень большая размерность пространства, поэтому стоимость 1 итерации Ньютона очень большая, поэтому он проигрывает градиентному спуску.

4.4 Вывод

Метод Ньютона имеет линейную сходимость, но при этом большую стоимость итерации для большой размерности пространства. Поэтому он сходится за малое число итераций и даёт хорошие результаты на датасетах с небольшой размерностью. Однако при большой размерности время сильно увеличивается по сравнению с градиентным спуском.

4.4.1 Ньютон

1. Сложность 1 итерации $O(n^3)$
2. Память для 1 итерации $O(n^2)$

4.4.2 Градиентный спуск

1. Сложность 1 итерации $O(n)$
2. Память для 1 итерации $O(n)$