Лабороторная работа №1

Роман Положиев

Формулы логистической регрессии

$$f(x) = \frac{1}{m}\sum_{i=1}^{m} \ln(1+\exp(-b_i\langle a_i, x\rangle)) + \frac{\lambda}{2}\|x\|_2^2$$

$$df(x) = \frac{1}{m}\sum_{i=1}^{m} d\Big(\ln(1+\exp(-b_i\langle a_i, x\rangle))\Big) + \frac{\lambda}{2}d\|x\|_2^2 = \frac{1}{m}\sum_{i=1}^{m}\frac{-\exp(-b_i\langle a_i, x\rangle)b_i\langle a_i, dx\rangle}{1+\exp(-b_i\langle a_i, x\rangle)} + \frac{\lambda}{2}d\langle x, x\rangle =$$
$$= \frac{1}{m}\sum_{i=1}^{m}\frac{-b_i\langle a_i, dx\rangle}{1+\exp(b_i\langle a_i, x\rangle)} + \lambda\langle x, dx\rangle$$

$$\nabla f(x) = \frac{1}{m}\sum_{i=1}^{m}\frac{-b_i a_i}{1+\exp(b_i\langle a_i, x\rangle)} + \lambda x$$

В матричной форме:

$$A = (a_1, \ldots, a_m)^T, \quad A \in \mathbb{R}^{m\times n}, \quad b = (b_1, \ldots, b_m)^T$$
$$\boxed{\nabla f(x) = -\tfrac{1}{m}A^T b \odot \big(1+\exp(b\odot Ax)\big)^{-1} + \lambda x}$$

где для $v = (v_1, \ldots, v_k)^T$: $v^{-1} = (1/v_1, \ldots, 1/v_k)^T$

$$d^2 f(x) = -\frac{1}{m}\sum_{i=1}^{m} d\Big(\frac{b_i\langle a_i, dx_1\rangle}{1+\exp(b_i\langle a_i, x\rangle)}\Big) + d\lambda\langle x, dx_1\rangle = \frac{1}{m}\sum_{i=1}^{m}\frac{b_i\exp(b_i\langle a_i, x\rangle)b_i\langle a_i, dx_2\rangle\langle a_i, dx_1\rangle}{\big(1+\exp(b_i\langle a_i, x\rangle)\big)^2} + \langle\lambda dx_1, dx_2\rangle =$$
$$= \frac{1}{m}\sum_{i=1}^{m}\frac{\exp(b_i\langle a_i, x\rangle)a_i a_i^T\langle dx_1, dx_2\rangle}{\big(1+\exp(b_i\langle a_i, x\rangle)\big)^2} + \langle\lambda dx_1, dx_2\rangle$$

$$\nabla^2 f(x) = \frac{1}{m}\sum_{i=1}^{m}\frac{\exp(b_i\langle a_i, x\rangle)}{\big(1+\exp(b_i\langle a_i, x\rangle)\big)^2}a_i a_i^T + \lambda$$

В матричной форме:

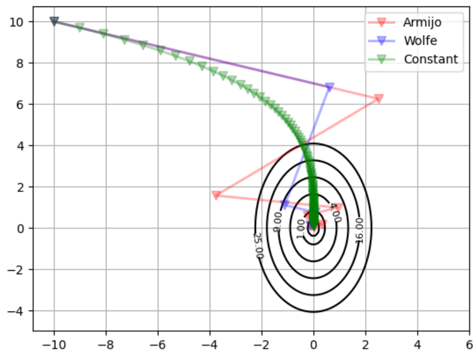
$$\boxed{\nabla^2 f(x) = \tfrac{1}{m}A^T \operatorname{diag}\big(\sigma\odot(1-\sigma)\big)A + \lambda I_n, \quad \text{где} \quad \sigma = \exp(-b\odot Ax)}$$

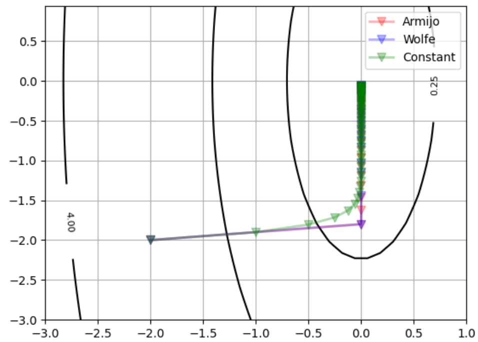
где $\sigma = \exp(-b\odot Ax)$

Сама функция в матричной форме:

$$f(x) = \frac{1}{m}1_m^T \times \ln\big(1+\exp(-b\odot Ax)\big) + \frac{\lambda}{2}\|x\|_2^2$$

Эксперимент 1



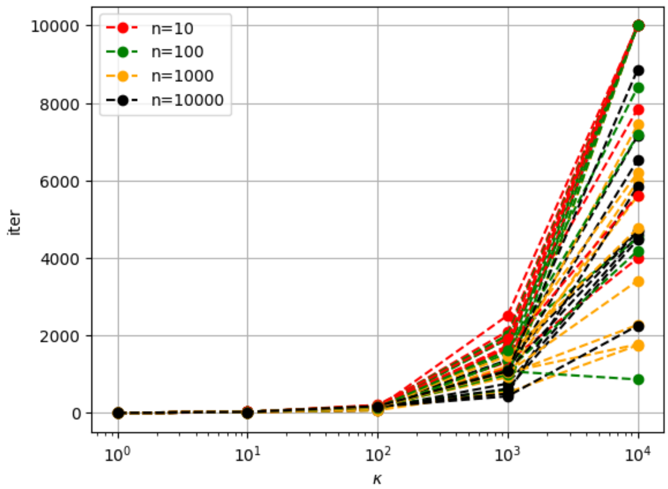


Чем больше число обусловленности функции, тем более вытянутые линии уровня и тем дольше сходится метод градиентного спуска.

Выбор начальной точки может по-разному влиять на поведение метода. Банально: если выбрать точку далеко от точки оптимума, то метод будет работать дольше, чем если бы выбрали начальную точку рядом. Также выбор начальной точки может влиять на то, куда мы сойдемся. Пример: гиперболический параболоид, если выбрать начальную точку $(x_0, 0)$, то метод сойдется к «седлу», а если же выбрать вторую координату отличную от нуля, то метод «выскочит из седла» и пойдет искать оптимум на бесконечность (как и должно быть).

Константная стратегия выбора шага как правило требует больше шагов для сходимости, особенно если начальная точка находится на отдалении от точки оптимума.

Эксперимент 2



Чем больше число обусловленности, тем больше число итераций. При этом, чем больше размерность пространства, тем меньше требуется итераций при одинаковом числе обусловленности. Я считаю, что такой результат получается из-за того, что в многомерных пространствах меньше вероятность попасть на вытянутые оси эллипсоида, вдоль которых метод будет медленно сходиться.

Эксперимент 3

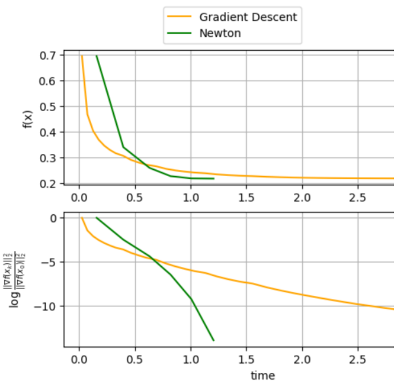


Рис. 1: w8a

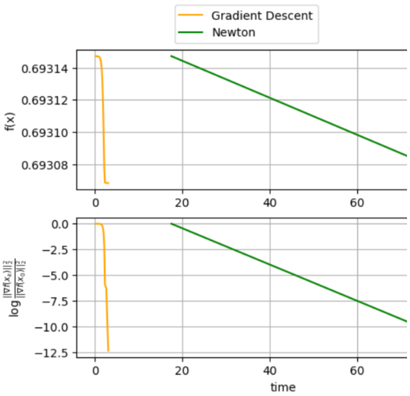


Рис. 2: real-sim

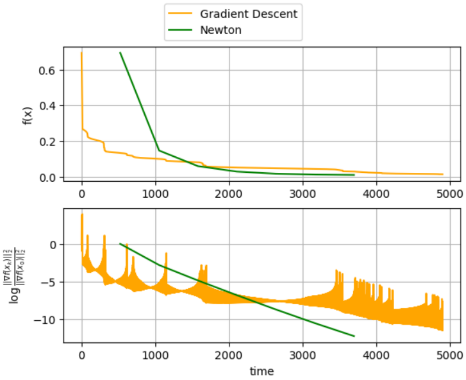


Рис. 3: gisette

Пусть $O(q)$ — стоимость нахождения $f(x)$ или $\nabla f(x)$

Метод	Память	Стоимость итерации
Градиентный спуск	$O(n)$	$O(q) + O(n)$
Ньютон	$O(n^2)$	$O(n^3) + O(qn^2)$

Метод Ньютона вычислительно эффективен только для пространств малой размерности, где он сходится быстрее чем метод градиентного спуска.

Эксперимент 4

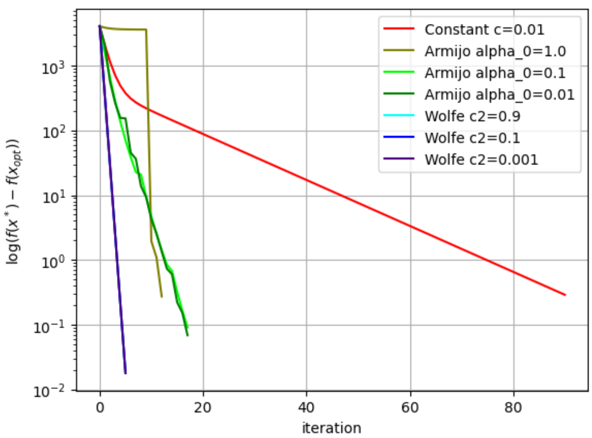


Рис. 4: Квадратичная функция

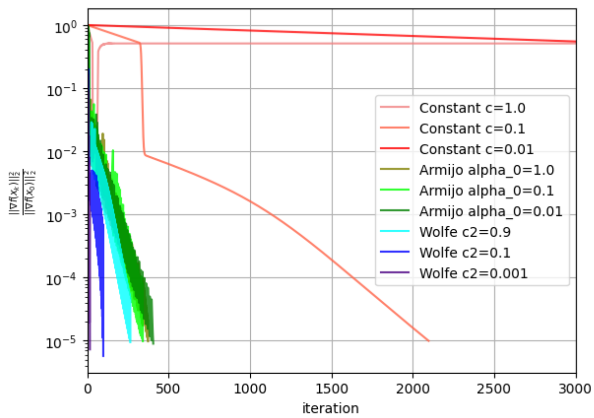
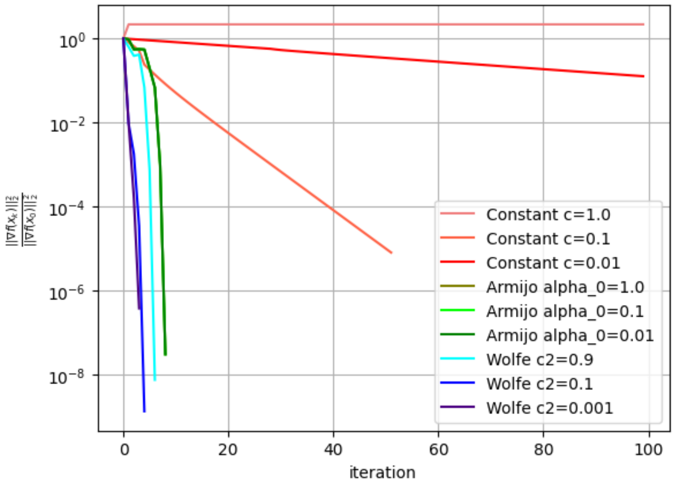


Рис. 5: Логистическая регрессия

Лучше всего работает стратегия с условиями Вульфа. И интересно, что лучше всего себя показывает стратегия с маленькой константой c_2.

Эксперимент 5



Аналогично, лучше всего работает стратегия с условиями Вульфа.