Московский государственный университет им. М. В. Ломоносова
Факультет вычислительной математики и кибернетики

Отчет по заданию

Неточный метод Ньютона для ℓ_2-регуляризованной логистической регрессии

Выполнил студент 317 группы
Измаилов Павел Алексеевич

Москва, 5 ноября 2015

1 Описание проделанной работы

В данном отчете содержатся результаты экспериментов, проведенных мной в соответствии с первым заданием по спецкурсу «методы оптимизации в машинном обучении». Мной были реализованы метод Ньютона, неточный метод Ньютона, а также метод сопряженных градиентов для решения задачи квадратичного программирования. Также было проведено сравнение точного и неточного методов Ньютона, а также сравнение неточного метода с нелинейным методом сопряженных градиентов и с методом L-BFGS из библиотеки `scipy.optimize`.

2 Рассматриваемая задача

В данной работе сравниваются результаты работы различных методов оптимизации на задаче логистической регрессии. Целевая функция этой задачи имеет вид

$$F(\boldsymbol{w}) = \sum_{i=1}^{N} \ln(1 + \exp(-y_i \boldsymbol{w}^T \boldsymbol{x})) + \frac{\lambda}{2} ||\boldsymbol{w}||_2^2.$$

Здесь $\boldsymbol{w} \in \mathbb{R}^D$, $\boldsymbol{x} \in \mathbb{R}^{N \times D}$ — объекты обучающей выборки, $\boldsymbol{y} \in \{-1, 1\}^N$ — ответы на обучающей выборке, $\lambda \in \mathbb{R}$ — константа регуляризации.

3 Рассматриваемые методы

В данном разделе приводятся краткие описания методов оптимизации, рассматриваемых в данной работе.

3.1 Правило Вольфа

В данной работе как для точного, так и для неточного метода Ньютона для выбора параметра длины шага использовалось правило Вольфа, которое имеет следующий вид. Фиксируем $\varepsilon_1, \varepsilon_2 \in (0, 1)$, $\chi_1 > 1$, $\chi_2 \in (0, 1)$. Полагаем $\check{\alpha} = \hat{\alpha} = 0$, $\alpha = \alpha_k$.

1. Проверяем выполнение неравенств

$$F(\boldsymbol{w}_k + \alpha \boldsymbol{d}_k) \leq F(\boldsymbol{w_k}) + \varepsilon_1 \alpha \langle \nabla F(\boldsymbol{w}_k), d_k \rangle$$

$$\langle \nabla F(\boldsymbol{w}_k + \alpha \boldsymbol{d}_k), \boldsymbol{d}_k \rangle \geq \varepsilon_2 \langle \nabla F(\boldsymbol{w}_k), \boldsymbol{d}_k \rangle,$$

   где $\langle \boldsymbol{a}, \boldsymbol{b} \rangle$ означает скалярное произведение векторов $\boldsymbol{a}$ и $\boldsymbol{b}$.

   Если оба они выполнены, то переходим к пункту 6.

2. Если нарушено первое неравенство, то $\hat{\alpha} = \alpha$.

3. Если нарушено второе неравенство, то $\check{\alpha} = \alpha$.

4. Если $\hat{\alpha} = 0$, то выбираем новое значение $\alpha = \check{\alpha}\chi_1$ и переходим к пункту 1.

5. Выбираем новое значение $\alpha = \hat{\alpha}\chi_2 + \check{\alpha}(1 - \chi_2)$ и переходим к пункту 1.

6. Полагаем $\alpha_{k+1} = \alpha$.

Во всех экспериментах параметры полагались равными $\varepsilon_1 = 10^{-4}, \varepsilon_2 = 0.9, \chi_1 = \chi_2 = 0.5$.

3.2 Метод Ньютона

Итерация метода Ньютона имеет вид

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha_k \nabla^2 F(\boldsymbol{w}_k)^{-1} \cdot \nabla F(\boldsymbol{w}_k),$$

где $\nabla^2 F(\boldsymbol{w})$ и $\nabla F(\boldsymbol{w})$ — соответственно гессиан и градиент функции F в точке w, а α_k — параметр длины шага, выбираемый тем или иным образом. В данной работе для выбора параметра длины шага использовалось правило Вольфа.

3.3 Метод сопряженных градиентов

Метод сопряженных градиентов предназначен для решения задачи оптимизации

$$f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle \to \min_x,$$

с симметричной положительно определенной матрицей $A \in \mathbb{R}^{n \times n}$. Эта задача эквивалентна линейной системе $Ax = b$. Для решения этой задачи в методе сопряженных градиентов строится система векторов $d_i, i = 1 \dots n$, такая что $\langle Ad_i, d_j\rangle = 0\ \forall i \neq j$. Такая система называется сопряженной относительно матрицы A. Будем обозначать через x_k приближение, полученные методом сопряженных градиентов на k-ой итерации $g_k = \nabla f(x_k) = Ax_k - b$, $f_k = f(x_k)$. Система d_k строится в методе сопряженных градиентов следующим образом:

1. Полагается $g_0 = Ax_0 - b$, $d_0 = -g_0$, $u_0 = Ad_0$.

2. Для $k = 1 \dots \#iter$ полагается

   (a) $\alpha_k = \dfrac{g_k^T g_k}{d_k^T u_k}$;

   (b) $x_{k+1} = x_k + \alpha_k d_k$;

   (c) $g_{k+1} = g_k + \alpha_k u_k$;

   (d) Если $||g_{k+1}|| < \varepsilon$, где ε — требуемая точность, остановиться;

   (e) $\beta_k = \dfrac{g_{k+1}^T g_{k+1}}{g_k^T g_k}$;

(f) $d_{k+1} = -g_{k+1} + \beta_k d_k$;

(g) $u_{k+1} = A d_k + 1$.

Если все операции метода осуществлять точно, то решение задачи будет получено не более, чем за n итераций.

3.4 Неточный метод Ньютона

Итерации неточного метода Ньютона имеют вид

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \alpha_k d_k,$$

где d_k — направление убывания, получаемое приближенным решением системы уравнений $\nabla^2 F(\boldsymbol{w}_k) d_k = -\nabla F(\boldsymbol{w}_k)$ с помощью метода сопряженных градиентов. При этом точность, с которой решается эта система на k-ой итерации, в данной работе полагается равной $\varepsilon_k = \min\{0.5, \sqrt{||\nabla F(\boldsymbol{w}_k)||}\}$.

4 Эксперименты

В данном разделе приводятся результаты проведенных экспериментов. Для получения графиков все методы запускались на некоторое фиксированное количество итераций, а в качестве минимального значения функции f_* бралось приближение, полученное заранее запуском одного из методов на большое количество итераций.

4.1 Метод Ньютона

В данной секции приводятся результаты работы метода Ньютона на небольшой задаче логистичекой регрессии. На графиках на рисунке 1 показаны результаты работы метода Ньютона на задаче со модельными данными и на задаче с реальными данными из набора fourclass.

Из графиков видно, что сходимость носит сверхлинейный характер. На наборе данных fourclass норма градиента становится меньше машинной точности уже на пятой итерации.

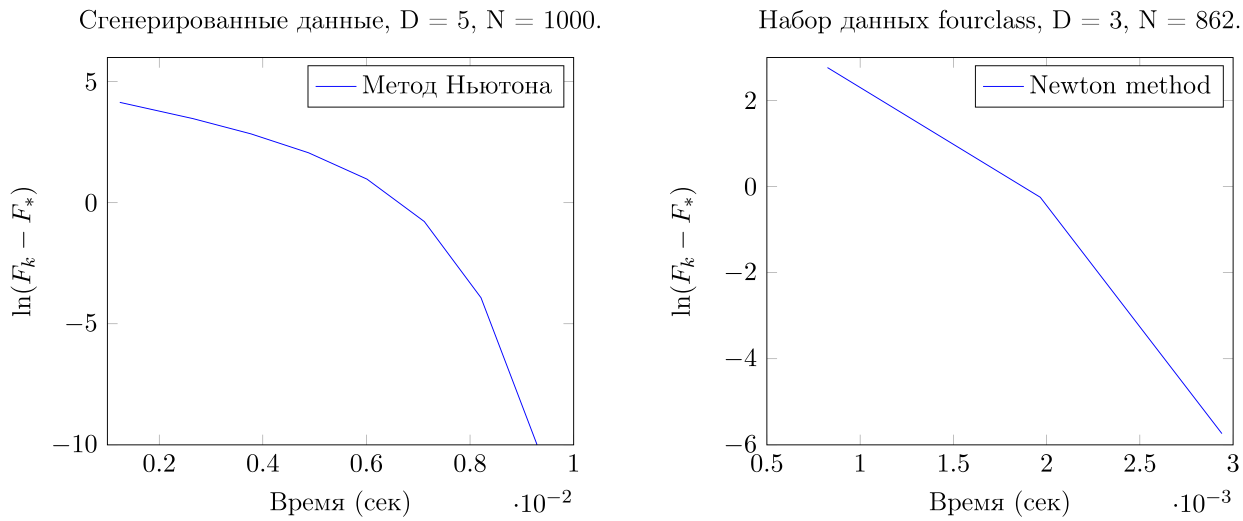

Рис. 1: Метод Ньютона на небольшой задаче логистической регрессии.

4.2 Метод сопряженных градиентов

В данном разделе приводятся результаты работы метода сопряженных градиентов. Для генерации симметричной положительно определенной матрицы с заданным набором собственных значений использовалась процедура, предложенная в задании, с учетом того, что у положительно определенной матрицы собственные значения совпадают с сингулярными.

На рисунке 2 представлен график сходимости метода сопряженных градиентов по итерациям для двух различных задач квадратичного программирования размерности 100. В первой из них собственные значения матрицы разбиваются на 50 кластеров, а во второй — на 10. Видно, что метод сходится за число итераций примерно равное числу кластеров собственных значений.

4.3 Неточный метод Ньютона

В данном разделе представлены результаты работы неточного метода Ньютона, а также его сравнение с другими методами.

На рисунке 3 показана зависимость логарифма невязки по функции от времени для обычного и неточного методов Ньютона на сгенерированных данных и на данных fourclass. Видно, что точный метод Ньютона на этих задачах работает несколько лучше, хотя оба метода сходятся очень быстро.

Далее рассмотрим результаты работы метода на больших задачах и сравним его с L-BFGS и с нелинейным методом сопряженных градиентов.

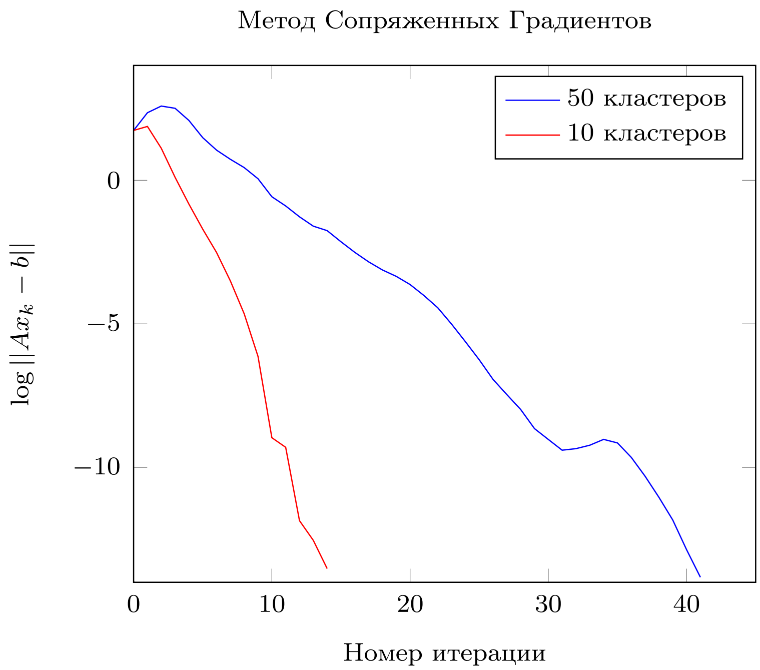

Рис. 2: Метод сопряженных градиентов для квадртаичной задачи с матрицей $A \in \mathbb{R}^{100\times 100}$ для 50 и 10 кластеров собственных значений.

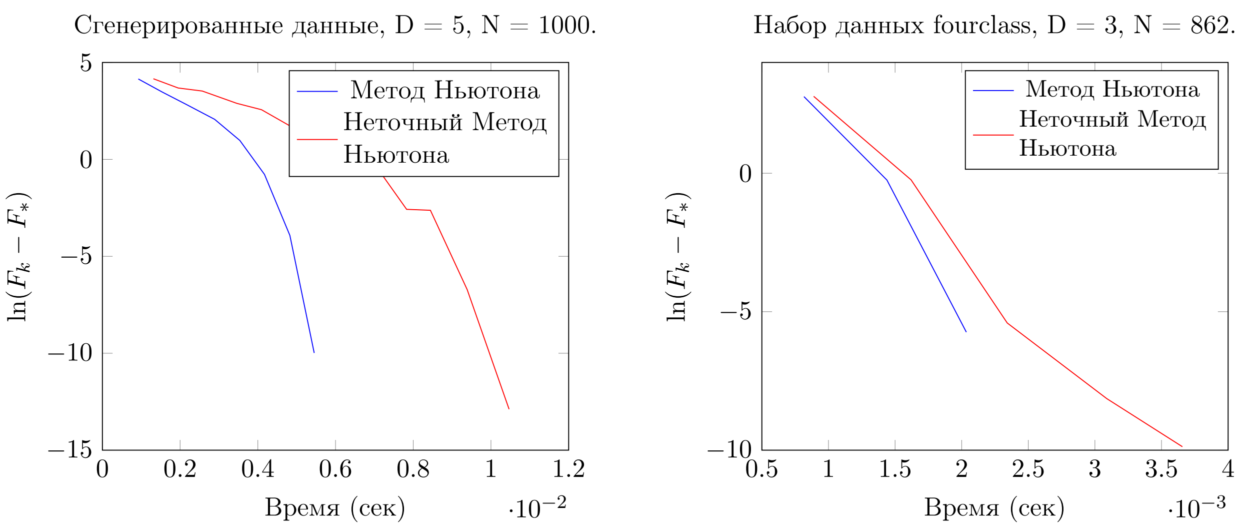

Рис. 3: Точный и неточный методы Ньютона на небольшой задаче логистической регрессии.

4.4 Наборы данных leukemia и duke breast cancer

Набор данных leukemia состоит из $N = 38$ объектов, имеющих $D = 7129$ признаков. Для экспериментов с этим набором данных использовалась плотная матрица объектов. Набор данных duke breast cancer состоит из $N = 44$ объектов, обладающих $D = 7129$ признаками. Для эксперименотов с этим набором использовалась разреженная матрица объектов. Результаты работы методов приведены на рисунке 4.

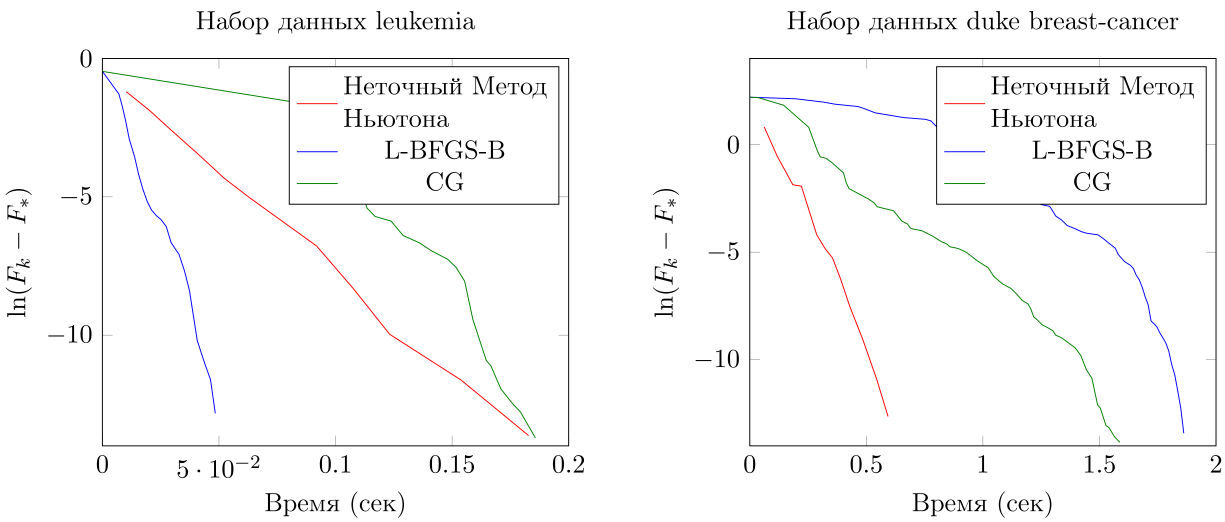

Рис. 4: Наборы данных leukemia ($N = 38$, $D = 7129$) и duke breast-cancer ($N = 44$, $D = 7129$).

В обоих случаях все методы сошлись очень быстро. Разницу в результатах можно объяснить тем, что расположение начальной точки по отношению к оптимуму в этих двух экспериментах было разным.

4.4.1 Набор данных gisette

Набор данных gisette состоит из $N = 6000$ объектов, число признаков которых равно $D = 5000$. В этом эксперименте использовалась плотная матрица объектов. На рисунке 5 представлены графики сходимости методов.

Из графиков видно, что итерации неточного метода Ньютона на этой задаче занимают значительно больше времени, чем итерации других методов. Тем не менее, эти итерации значительно эффективнее и метод очень быстро достигает очень низкой нормы градиента и останавливается, обогнав остальные методы. Метод CG на данной задаче существенно уступает двум другим методам.

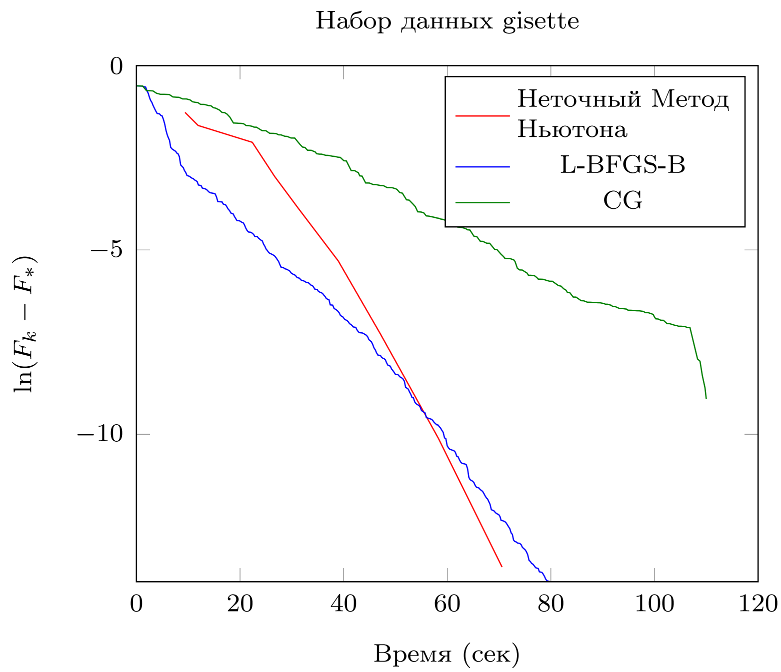

Рис. 5: Набор данных gisette, $N = 6000$, $D = 5000$.

4.4.2 Набор данных Real-sim

Набор данных real-sim состоит из $N = 72309$ объектов, обладающих $D = 20958$ признаками. Однако для экспериментов были использованы только первые 10000 объектов, так как иначе методы работали слишком долго. В этом эксперименте использовалась разреженная матрица объектов. Результаты работы методов приведены на рисунке 6.

На этой задаче методы L-BFGS и неточный метод Ньютона показали близкие результаты. Метод CG снова проиграл двум другим методам.

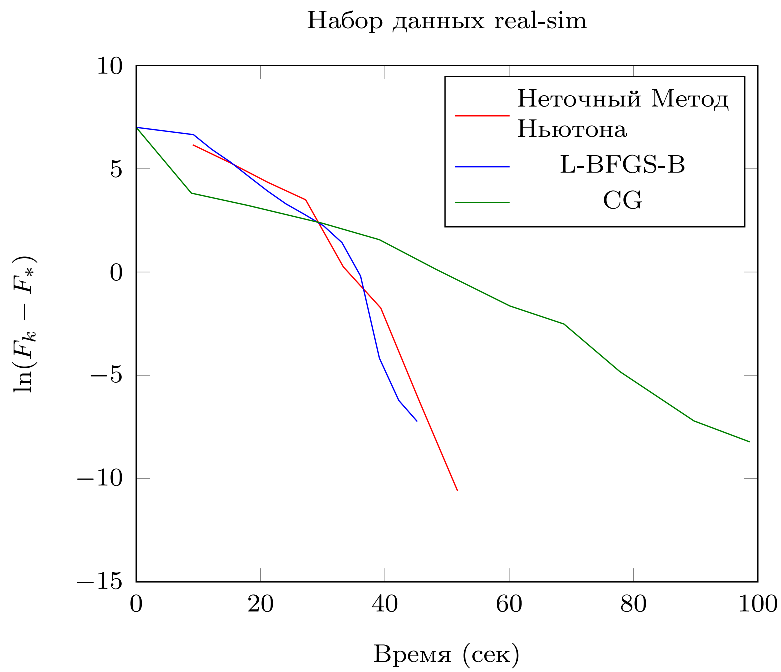

Рис. 6: Набор данных real-sim, $N = 72309$, $D = 20958$.

5 Выводы

По результатам экспериментов можно сделать ряд выводов.

Во-первых, на небольших задачах (когда гессиан вычисляется быстро) метод Ньютона ожидаемо сходится несколько быстрее неточного метода Ньютона и поэтому в таких задачах имеет смысл использовать точный метод.

Во-вторых экспериментально было проверено утверждение о том, что метод сопряженных градиентов для квадратичной задачи $\langle Ax, x\rangle + \langle b, x\rangle \to \min\limits_x$ сходится за число итераций примерно равное числу кластеров собственных значений матрицы A.

В-третьих было проведено сравнение неточного метода Ньютона с методами CG и L-BFGS из библиотеки `scipy.optimize` на задачах в пространствах размерности $D \geq 5000$. Как и следовало ожидать, неточный метод Ньютона обладает лучшей скоростью сходимости по итерациям, однако сами его итерации существенно дольше, чем у других двух методов. На наборах данных leukemia и duke breast-cancer методы показали близкие результаты. Эти задачи можно отнести к небольшим, все три метода сходятся на них быстро, и пользоваться можно любым. На больших наборах данных наборах данных gisette и real-sim методы L-BFGS-B и неточный метод Ньютона показали близкие результаты. При этом на ранних итерациях на наборе данных gisette метод L-BFGS-B показывает лучшую точность, поэтому на больших задачах, если достаточна не слишком высокая точность, следует отдавать преимущество ему. Метод CG в экспериментах на этих двух наборах данных показал себя существенно хуже других двух методов.