УДК 519.23:616006

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ МИНИМИЗАЦИИ ФУНКЦИИ КРОСС-ЭНТРОПИИ

Гладкая Н.Н., студент

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Примичева З.Н. – канд. физ.-мат. наук, доцент

Аннотация. Работа посвящена математическому обоснованию ускорения итерационных алгоритмов минимизации функции кросс-энтропии на примере задачи бинарной классификации. Определяется принадлежность опухоли к злокачественному или доброкачественному классу по вектору числовых характеристик. Выведена функция правдоподобия, её логарифмирование приводит к функционалу кросс-энтропии. Вычислены градиент и гессиан функционала потерь. Доказана положительная определённость гессиана, устанавливающая единственность глобального минимума и сходимость итерационных методов.

Ключевые слова. Логистическая регрессия, кросс-энтропия, распределение Бернулли, классификация, градиент, гессиан, метода Ньютона, градиентный спуск.

Введение. Ранняя и точная диагностика злокачественных новообразований остаётся одной из ключевых задач современной медицины. На основе результатов биопсии, представленных в виде числовых векторов характеристик клеток (радиус, текстура, периметр, компактность и другие), необходимо принять бинарное решение: классифицировать опухоль как злокачественную или доброкачественную. Формализация данной задачи в рамках теории вероятностей и математического анализа сводит её к минимизации функционала бинарной кросс-энтропии в модели логистической регрессии. Экспериментальные данные взяты с открытой платформы Kaggle [1].

Задача сводится к минимизации функционала потерь (бинарной кросс-энтропии) как функции вектора параметров $w \in \mathbb{R}^{\mathrm{d}}$. Градиентный спуск (стандартный метод машинного обучения) использует производные первого порядка. Однако математический анализ предоставляет значительно более мощный инструмент: производные второго порядка, образующие матрицу Гессе, учитывают кривизну поверхности функционала и позволяют достичь той же точности решения значительно быстрее.

1. Постановка задачи. Рассмотрим задачу бинарной классификации. Пусть задана обучающая выборка $\{(x_i, y_i)\}_{i=1}^{N}$, где N – объём выборки, $x_i = (x_{i1}, \dots, x_{id})^{\mathrm{T}} \in \mathbb{R}^{\mathrm{d}}$ – вектор из d числовых характеристик i-го наблюдения, $y_i \in \{0, 1\}$ – соответствующая бинарная метка ($y_i = 1$ означает злокачественную опухоль, $y_i = 0$ – доброкачественную). Через $\mathbb{R}^{\mathrm{d}}$ обозначается d – мерное вещественное евклидово пространство. Требуется найти вектор параметров $w \in \mathbb{R}^{\mathrm{d}}$, при котором функция $\sigma(w^{\mathrm{T}}x)$ наилучшим образом аппроксимирует условную вероятность злокачественности. Задача нахождения $w^* = argmin\, \mathcal{L}(w)$ является центральным математическим объектом работы. Предметом исследования является сопоставление двух численных методов: метода градиентного спуска и метода Ньютона.

2. Вероятностная модель и функционал потерь.

2.1. Модель Бернулли. Предположим, что при фиксированном x_i и параметрах w метка y_i является реализацией случайной величины: $y_i \sim Bernoulli(p_i), p_i = \sigma(w^{\mathrm{T}}x_i)$ [2]. Вероятностная масс-функция для i-го наблюдения записывается в компактном виде:

$$p(y_i \mid x_i, w) = {p_i}^{y_i} \cdot (1 - p_i)^{1-y_i},$$

где показатели степени $y_i \in \{0,1\}$ действуют как бинарный селектор: при $y_i = 1$ остаётся p_i, при $y_i = 0$ – множитель $(1 - p_i)$, что позволяет охватить оба случая единой формулой.

2.2. Логит-связь. Линейный предиктор $\eta_i = w^{\mathrm{T}}x_i$ принимает значения на всей прямой $\mathbb{R}$, тогда как вероятность $p_i \in (0,1)$. Для согласования областей значений вводится логит-связь [2]: $\ln(p_i/(1 - p_i)) = w^{\mathrm{T}}x_i$. Потенцируя и выражая p, получаем $p_{\mathrm{i}} = \sigma(w^{\mathrm{T}}x_i)$. Производная сигмоиды вычисляется дифференцированием сложной функции [3]:

$$\sigma'(z) = \frac{e^{-z}}{(1 + e^{-z})^2} = \sigma(z) \cdot (1 - \sigma(z)).$$

Это тождество является ключевым: правая часть выражена только через значения сигмоиды в данной точке, без экспонент и дробей, что существенно упрощает последующее дифференцирование функции.

2.3. Кросс-энтропия. При предположении о независимости наблюдений совместное правдоподобие факторизуется: $L(w) = \prod_{i=1}^{N} \sigma_i^{y_i}(1-\sigma_i)^{1-y_i}$. Поскольку логарифм является строго монотонным преобразованием, он не меняет точку максимума, но переводит произведение в сумму:

$$\ell(w) = \sum_{i=1}^{N}[\ y_i \ \ln\sigma_i + (1 - y_i)\ \ln(1 - \sigma_i)\ ].$$

Совершим переход от максимизации логарифма правдоподобия к минимизации его отрицательной величины и нормируем на N. Получаем функционал бинарной кросс-энтропии [4]:

$$\mathcal{L}(w) = -\frac{1}{N} \cdot \sum_{i=1}^{N}[\ y_i \ \ln\sigma(w^{\mathrm{T}}x_i) + (1 - y_i)\ \ln(1 - \sigma(w^{\mathrm{T}}x_i))\ ].$$

3. Градиент функционала кросс-энтропии.

Обозначим $z_i = w^{\mathrm{T}}x_i$. Применяя цепное правило [3], вычислим компоненту градиента $\partial\mathcal{L}/\partial w_j$. Дифференцируем i-ое слагаемое.

Первое слагаемое $y_i \ \ln\sigma(z_i)$:

$$\partial/\partial w_j\ [y_i \ \ln\sigma_i] = y_i \cdot (1/\sigma_i) \cdot \sigma_i(1-\sigma_i) \cdot x_{ij} = y_i(1-\sigma_i)x_{ij}.$$

Второе слагаемое $(1-y_i)\ln(1-\sigma_i)$:

$$\partial/\partial w_j\ [(1-y_i)\ln(1-\sigma_i)] = -(1-y_i)\sigma_i x_{ij}.$$

Складывая оба слагаемых, члены $\pm y_i\sigma_i x_{ij}$ взаимно уничтожаются, имеем:

$$\partial\ell_i/\partial w_j = y_i x_{ij} - y_i\sigma_i x_{ij} - \sigma_i x_{ij} + y_i\sigma_i x_{ij} = (y_i - \sigma_i)x_{ij}.$$

Суммируя по всем наблюдениям и учитывая знак минус перехода от ℓ к $\mathcal{L}$, получим

$$\partial\mathcal{L}/\partial w_j = \frac{1}{N} \cdot \sum_{i=0}^{N}(\sigma_i - y_i)\ x_{ij}\ .$$

В матричной форме, вводя матрицу признаков $X \in \mathbb{R}^{\mathrm{Nxd}}$ и вектор предсказанных вероятностей $\hat{p} = \sigma(Xw) \in \mathbb{R}^{\mathrm{N}}$, имеем

$$\nabla w\ \mathcal{L}(w) = \frac{1}{N} \cdot X^{\mathrm{T}}(\hat{p} - y).$$

Геометрическая интерпретация: градиент есть взвешенная сумма строк матрицы X, где вес i-ой строки равен ошибке предсказания $(\hat{p}_i - y_i)$. Формула градиентного спуска с шагом $\alpha > 0$ запишется как

$$w_{\{k+1\}} = w_k - \alpha \cdot \frac{1}{N} \cdot X^{\mathrm{T}}(\hat{p}_k - y).$$

Схема сходится при $\alpha < 2/\lambda_{max}(H)$ со скоростью линейной сходимости [5]; число итераций до ε-точности определяется числом обусловленности $k(H) = \frac{\lambda_{max}}{\lambda_{min}}$ гессиана, где λ_{max} и λ_{min} – наибольшее и наименьшее собственные значения матрицы H. Физически $k(H)$ показывает, насколько вытянуты уровневые множества функционала: при $k(H) >> 1$, то есть когда наибольшее собственное значение на несколько порядков превышает наименьшее, эллипсоиды уровня сильно вытянуты, и градиентный спуск совершает зигзагообразную траекторию в «овраге», сходясь крайне медленно.

4. Гессиан и строгая выпуклость.

4.1. Вычисление гессиана. Гессиан – матрица вторых частных производных [3]:

$$H_{jk} = \partial^2\mathcal{L}/(\partial w_j\ \partial w_k).$$

Дифференцируем $\partial\mathcal{L}/\partial w_j$ по w_k; единственный множитель, зависящий от w_k, – это σ_i:

$$\partial\sigma_i/\partial w_k = \sigma'(z_i) \cdot x_{ik} = \sigma_i(1 - \sigma_i) \cdot x_{ik},$$

$$H_{jk} = \frac{1}{N} \cdot \sum_{i=1}^{N} \sigma_i(1 - \sigma_i) \cdot x_{ij} \cdot x_{ik}.$$

Веса $\sigma_i(1-\sigma_i)$ зависят только от номера наблюдения i. Введём диагональную матрицу $S = diag(\sigma_1(1-\sigma_1), \dots, \sigma_n(1-\sigma_n)) \in \mathbb{R}^{\mathrm{NxN}}$. Тогда сумма факторизуется:

$$H(w) = \frac{1}{N} \cdot X^{\mathrm{T}} S X.$$

Матрица H симметрична ($H = H^{\mathrm{T}}$) и зависит от w через диагональные элементы S, что отличает нелинейную задачу от линейных, где гессиан постоянен. H допускает разложение в сумму N матриц ранга 1:

$$H = \sum_{i=1}^{N} \sigma_i(1-\sigma_i) x_i x_i^{\mathrm{T}}.$$

4.2. Теорема (положительная определённость). При условии $rank(X) = d$ матрица $H(w)$ положительно определена для любого $w \in \mathbb{R}^{\mathrm{d}}$ [6].

Доказательство. Для произвольного ненулевого вектора $v \in \mathbb{R}^{\mathrm{d}}$ вычислим квадратичную форму:

$$v^{\mathrm{T}} H v = \frac{1}{N} \cdot \sum_{i=1}^{N} \sigma_i(1-\sigma_i) \cdot (v^{\mathrm{T}} x_i)^2 .$$

Поскольку $\sigma_i \in (0,1)$, все коэффициенты $\sigma_i(1-\sigma_i) > 0$; квадраты $(v^{\mathrm{T}} x_{\mathrm{i}})^2 \geq 0$, следовательно, указанная сумма является оетрицательной. Строгое неравенство обеспечивается условием $rank(X) = d$: если бы $v^{\mathrm{T}} x_i = 0$ для всех i, вектор v лежал бы в ядре X^{T}, что противоречит полноранговости. Значит, $(v^{\mathrm{T}} x_i)^2 > 0$ хотя бы для одного i, и $v^{\mathrm{T}} H v > 0$ для всех $v \neq 0$. Теорема доказана.

Следствие. Положительная определённость гессиана равносильна строгой выпуклости $\mathcal{L}(w)$. Строго выпуклая функция имеет единственную стационарную точку; локальных минимумов, кроме глобального, не существует. Оба алгоритма при выполнении условий сходимости гарантированно приходят к единственному w^*.

5. Метод Ньютона: аппроксимация второго порядка.

5.1. Разложение Тейлора второго порядка. Идея метода – локальная замена $\mathcal{L}$ квадратичной аппроксимацией в окрестности текущего приближения w_k. В одномерном случае имеем:

$$f(x + \Delta x) \approx f(x) + f'(x)\Delta x + \frac{1}{2} f''(x)(\Delta x)^2.$$

В многомерном случае скаляры заменяются на градиент и гессиан:

$$\mathcal{L}(w_k + \Delta w) \approx Q(\Delta w) = \mathcal{L}(w_k) + \nabla \mathcal{L}^{\mathrm{T}} \Delta w + \frac{1}{2} \cdot \Delta w^{\mathrm{T}} H \cdot \Delta w.$$

Слагаемые $Q(\Delta w)$: $\mathcal{L}(w_k)$ – значение в текущей точке (константа); $\nabla \mathcal{L}^{\mathrm{T}} \Delta w = \sum_{j=1}^{d} (\partial \mathcal{L} / \partial w_j) \cdot \Delta w_j$ – линейная поправка, учитывающая наклон вдоль каждой оси; $\frac{1}{2} \cdot \Delta w^{\mathrm{T}} H \cdot \Delta w$ – квадратичная поправка, учитывающая кривизну, в том числе перекрёстную (недиагональные элементы H_{jk}). Именно последнее слагаемое отличает метод Ньютона от градиентного спуска, полностью игнорирующего кривизну.

5.2. Минимизация квадратичной аппроксимации. Берём производную $Q(\Delta w)$ по Δw и приравниваем к нулю:

$$\partial Q / \partial \Delta w = \nabla \mathcal{L} + H \cdot \Delta w = 0.$$

Из обратимости H (следует из $H \succ 0$):

$$\Delta w^* = -H^{-1} \cdot \nabla \mathcal{L}.$$

5.3. Шаг Ньютона. Подставляя оптимальный сдвиг в правило обновления, имеем:

$$w_{k+1} = w_k - H(w_k)^{-1} \cdot \nabla \mathcal{L}(w_k).$$

Для задачи кросс-энтропии, подставляя явные формулы, найдем:

$$w_{k+1} = w_k - \left[\frac{1}{N} \cdot X^{\mathrm{T}} S_k X \right]^{-1} \cdot \frac{1}{N} \cdot X^{\mathrm{T}} (\hat{p}_k - y).$$

5.4. Квадратичная сходимость.

Вблизи строго выпуклого минимума w^* метод Ньютона обладает квадратичной сходимостью [7]: $\|w_{k+1} - w^*\| \leq C \cdot \|w_k - w^*\|^2$. Число верных знаков удваивается на каждой итерации. Градиентный спуск сходится линейно: $\|w_{k+1} - w^*\| \leq \rho \cdot \|w_k - w^*\|$ при $\rho < 1$, зависящем от $k(H)$; при $k(H) \gg 1$ линейная сходимость крайне медленна.

6. Сравнительный анализ алгоритмов.

Таблица 1 – Сравнение алгоритмов первого и второго порядка

Характеристика	**Градиентный спуск**	**Метод Ньютона**
Используемые производные	1-й порядок: $\nabla\mathcal{L}$	2-й порядок: $\nabla\mathcal{L}, H$
Шаг обновления	$w \leftarrow w - \alpha \cdot \nabla\mathcal{L}$	$w \leftarrow w - H^{-1}\nabla\mathcal{L}$
Учёт кривизны	Нет	Да (гессиан)
Скорость сходимости	Линейная: $O\left(k \cdot \ln\left(\frac{1}{\varepsilon}\right)\right)$	Квадратичная: $O\left(\left(\ln\left(\frac{1}{\varepsilon}\right)\right)\right)$
Стоимость итерации	$O(Nd)$	$O(Nd^2 + d^3)$
Выбор шага α	Обязателен	Не требуется

Принципиальное преимущество метода Ньютона состоит в инвариантности шага к аффинным преобразованиям пространства параметров. Шаг $\Delta w^* = -H^{-1}\nabla\mathcal{L}$ автоматически масштабируется по направлениям с учётом кривизны [6], устраняя зигзагообразную траекторию в «оврагах». Градиентный спуск лишён этого механизма: шаг α требует ручной настройки, а при $k(H) \gg 1$ оптимальный α мал и число итераций велико.

Ограничение метода Ньютона – стоимость $O(Nd^2 + d^3)$ на итерацию. При $d \sim 10^4 - 10^5$ хранение и инверсия матрицы $d \times d$ становятся затратными.

7. Численный эксперимент. Рассматривается задача бинарной классификации: по вектору $x \in \mathbb{R}^{30}$ числовых характеристик опухоли предсказывается принадлежность к злокачественному или доброкачественному классу. Данные получены с платформы Kaggle ($N = 569$ наблюдений, $d = 30$ признаков); модель построена автором. Разбиение: 64% – обучение (364 наблюдений), 16% – валидация 91 наблюдение), 20% – контрольная выборка (114 наблюдений). Нормировка: каждый признак центрируется и масштабируется по обучающей выборке. Инициализация: $w_0 = 0$. Критерий остановки: $\|\Delta w\| < 10^{-6}$, не более 3000 итераций.

Результаты. Для получения численных результатов (таблица 1) была построена модель логистической регрессии на языке Python с использованием библиотеки NumPy для реализации матричных вычислений. Разбиение выборки и нормировка признаков выполнены средствами библиотеки scikit-learn, графики сходимости на рисунке 1 построены с помощью Matplotlib.

Параметры подбирались по минимуму функции потерь на валидационной выборке. Для градиентного спуска: шаг $\alpha = 0{,}1$, регуляризация $\lambda = 0$.

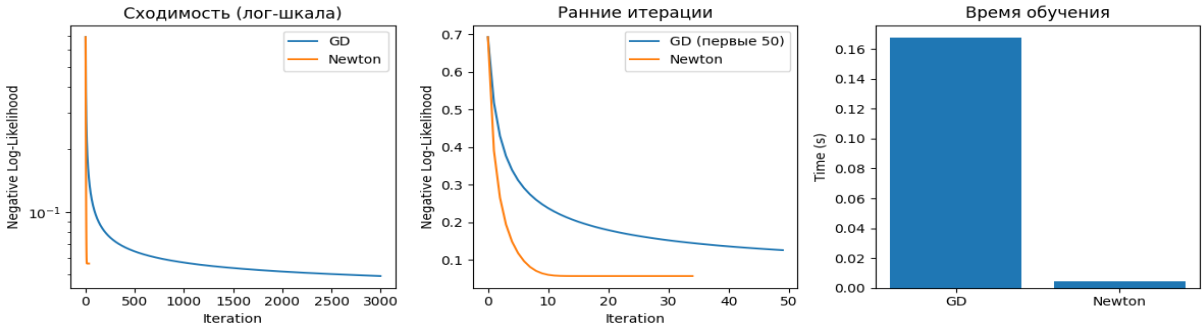

Рисунок 1 – Динамика функционала потерь: график слева – полная история обучения в логарифмической шкале; в центре – первые 50 итераций; справа – сравнение машинного времени

На рисунке отчётливо проявляется квадратичная сходимость метода Ньютона: уже к седьмой итерации значение функционала опускается ниже $0{,}08$, тогда как градиентный спуск за первые 50 итераций остаётся в диапазоне $0{,}1 - 0{,}4$. Оба метода приходят к одному качеству классификации – что теоретически гарантировано единственностью минимума.

Стократное ускорение по итерациям обусловлено различием типов сходимости: линейная против квадратичной. Ускорение по машинному времени (в 24 раза) несколько меньше, поскольку каждая итерация метода Ньютона стоит дороже – $O(Nd^2 + d^3)$ против $O(Nd)$.

Заключение. В работе рассмотрена задача минимизации функционала кросс-энтропии. Из вероятностной модели Бернулли выведена функция правдоподобия; логарифмирование дало функционал $\mathcal{L}(w)$. Цепное правило дифференцирования позволило получить градиент $\nabla\mathcal{L} = \frac{1}{N} \cdot X^{\mathrm{T}}(\hat{p} - y)$. Гессиан доказан положительно определённым, что равносильно строгой выпуклости $\mathcal{L}$ и единственности глобального минимума w^*.

Разложение Тейлора второго порядка дало квадратичную аппроксимацию $Q(\Delta w)$. Её минимизация по Δw – шаг Ньютона $w \leftarrow w - H^{-1}\nabla\mathcal{L}$, обладающий квадратичной сходимостью и инвариантностью к аффинным преобразованиям пространства параметров, в отличие от градиентного спуска, чья линейная скорость определяется числом обусловленности $k(H)$.

Численный эксперимент на задаче бинарной классификации опухолей подтверждает теоретические выводы: метод Ньютона сошёлся за ~ 30 итераций против 3000 у градиентного спуска, затратив $0{,}007$ секунд против $0{,}168$ секунд при одинаковом качестве классификации – 111 верных ответов из 114 и 3 ошибки у обоих методов.

Полученные результаты демонстрируют, что теория производных, теорема Тейлора, критерий выпуклости через гессиан непосредственно применимы к задачам, имеющим социальную значимость: ускорение диагностических алгоритмов снижает энергопотребление медицинских информационных систем без какого-либо ущерба для точности постановки диагноза.

Список использованных источников:

1. Kaggle – URL: https://www.kaggle.com (дата доступа: 10.03.2026).

2. Probabilistic Machine Learning: An Introduction / K. P. Murphy – MIT Press, 2022. – с. 258-260.

3. Курс дифференциального и интегрального исчисления: в 3 т. / Г. М. Фихтенгольц. – 8-е изд. – М.: Физматлит, 2003. – Т. 1. – 680 с.; Т. 2. – 864 с.

4. Numerical Optimization / J. Nocedal, S. Wright. – 2nd ed. – Springer, 2006. – с. 44-46.

5. An Overview of Gradient Descent Optimization Algorithms / S. Ruder – arXiv, 2016. – URL: https://arxiv.org/abs/1609.04747 (дата доступа: 10.03.2026).

6. Pattern Recognition and Machine Learning / C. M. Bishop – Springer, 2006. – с. 204-206.

7. Convex Optimization / S. Boyd, L. Vandenberghe – Cambridge University Press, 2004. – с. 67-69.

UDC 519.23:616006

COMPARATIVE ANALYSIS OF METHODS FOR MINIMIZING THE CROSS-ENTROPY FUNCTION

Gladkaya Natalia Nikolaevna, student

Belarusian State University of Informatics and Radioelectronics Minsk, Republic of Belarus

Primicheva Z.N. – PhD in Physics and Mathematics

Abstract. The work is devoted to the mathematical justification of accelerating iterative algorithms for minimizing the cross-entropy function using the example of a binary classification problem. The classification of a tumor as malignant or benign is determined by a vector of numerical characteristics. The likelihood function is derived, and its logarithm leads to the cross-entropy functional. The gradient and Hessian of the loss functional are computed. The positive definiteness of the Hessian is proven, establishing the uniqueness of the global minimum and the convergence of iterative methods.

Keywords. Logistic regression, cross-entropy, Bernoulli distribution, classification, gradient, Hessian, Newton's method, gradient descent.