Исследование логистической регрессии

Петров Егор

МФТИ

2 ноября 2024 г.

План

Постановка задачи

- Постановка задачи логистической регрессии
  Итак, ставим задачу бинарной классификации $\{x_i, y_i\}_{i=1}^n$, $x_i \in \mathbb{R}^d, y_i \in \{0, 1\}$

Сама модель логистической регрессии имеет вид:

$$\hat{y}_i = \begin{cases} 0, \text{если } \sigma(\theta^T x) < threshold \\ 1, \text{иначе} \end{cases} \tag{1}$$

В таком случае, мы пользуемся предположением, о том, что сигмоиду $\sigma(\theta^T x) = \frac{1}{1-\exp(-\theta^T x)}$ можно интерпретировать как вероятность

LogLoss

Теперь займемся выводом оптимизируемого функционала ошибки
Поскольку $y_i \in \{0, 1\}$, то будем интерпретировать задачу как поиск оптимального параметра для распределения $Bern(p)$, то есть
$y_i \sim Bern(p)$
Теперь перепишем условную вероятноть в виде:

$$\mathbb{P}_p(y_i) = p^{y_i} \cdot (1-p)^{1-y_i} \tag{2}$$

Тогда правдоподобие примет вид:

$$\mathbb{L}_{y_i}(p) = p^{\sum_{i=1}^n y_i} \cdot (1-p)^{n-\sum_{i=1}^n y_i} \tag{3}$$

LogLoss

Рассмотрим логарифм правдоподобия (здесь стоит заметить, что логарифм монотонная возрастающая функция, а значит с точки зрения оптимизации данный переход корректен)

$$l_{y_i}(p) = (\sum_{i=1}^{n} y_i)\log(p) + (n - \sum_{i=1}^{n} y_i)\log(1-p) \tag{4}$$

Теперь заменим p на нашу оценку истинной вероятности $\sigma(\theta^T x_i)$ и и записываем под одной суммой

$$l_{y_i}(\theta) = \sum_{i=1}^{n} y_i \log(\sigma(\theta^T x_i)) + (1 - y_i)\log(1 - \sigma(\theta^T x_i)) \to \max_{\theta} \tag{5}$$

LogLoss

Тогда сама функция LogLoss(θ) приме вид (заменяем для минимизациии знак в предыдущем выражении)

$$LogLoss(\theta) = -\left(\sum_{i=1}^{n} y_i \log(\sigma(\theta^T x_i)) + (1 - y_i) \log(1 - \sigma(\theta^T x_i))\right) \to \min_{\theta} \tag{6}$$

Таким образом, получили оптимизируемый функционал
Теперь убедимся, что для нахождения оптимума полученного выражения можно пользоваться итерационными методами, основанными на градиентном спуске, то есть докажем выпуклость полученного функционала

Выпуклость

Для доказательства этого факта воспользуемся критерием выпуклости второго порядка, который гласит

$$f(x) - \text{выпукла} \Leftrightarrow dom(f) - \text{выпуклое и } \nabla^2 f \succcurlyeq 0 \tag{7}$$

В данном случае, выпуклость множества $dom(f)$ очевидна, поскольку следует из вида нашей функции, так как здесь $dom(f) = R^d$, что по определению является выпуклым множеством (посольку значения LogLoss по модулю близкие к $\pm\infty$ достигаются при аналогичных значениях $\theta \to \pm\infty$, которые входят только в расширенное пространство $\overline{\mathbb{R}}$)
Теперь остановимся на положительной полуопределенности гессиана, для начала найдем его

Выпуклость

Градиент:

$$\nabla_\theta LogLoss(\theta) = -\nabla_\theta \left(\sum_{i=1}^{n} y_i \log(\sigma(\theta^T x_i)) + (1 - y_i) \log(1 - \sigma(\theta^T x_i)) \right) =$$

$= \text{-}\sum_{i=1}^{n} \left(y_i \nabla_\theta \log(\sigma(\theta^T x_i)) + (1 - y_i) \nabla_\theta \log(1 - \sigma(\theta^T x_i)) \right)$ (8)
Далее следует заметить, что $\sigma'(x) = \sigma(x)(1 - \sigma(x))$ $\nabla_\theta LogLoss(\theta) =$
$- \sum_{i=1}^{n} \left(y_i \frac{\sigma(\theta^T x_i)(1-\sigma(\theta^T x_i)) \cdot x_i}{\sigma(\theta^T x_i)} - (1 - y_i) \frac{\sigma(\theta^T x_i)(1-\sigma(\theta^T x_i)) \cdot x_i}{1-\sigma(\theta^T x_i)} \right) =$
$= \text{-}\sum_{i=1}^{n} \left(y_i (1 - \sigma(\theta^T x_i)) x_i - (1 - y_i) \sigma(\theta^T x_i) \cdot x_i \right) =$
$= \text{-}\sum_{i=1}^{n} \left(y_i - \sigma(\theta^T x_i) \right) x_i = \sum_{i=1}^{n} \left(\sigma(\theta^T x_i) - y_i \right) x_i$
Что можно несложно переписать в матричном виде:

$$\nabla_\theta LogLoss(\theta) = X^T (S(\theta) - Y) \tag{9}$$

где $X_i = x_i, S(\theta)_i = \sigma(\theta^T x_i), Y_i = y_i$

Выпуклость

Теперь вычислим непосредственно гессиан:

$$\nabla^2_\theta LogLoss(\theta) = \sum_{i=1}^{n} \big(\sigma(\theta^T x_i)(1 - \sigma(\theta^T x_i)\big) x_i x_i^T \tag{10}$$

Что также несложно переписывается в матричном виде:

$$\nabla^2_\theta LogLoss(\theta) = X^T V(\theta) X \tag{11}$$

где $V(\theta) = diag(\sigma(\theta^T x_i)(1 - \sigma(\theta^T x_i)))$
Заметим, что $V(\theta) \succcurlyeq 0$, поскольку матрица диагональна и при этом на самой диагонали стоят положительные элементы, а поскольку эта матрица умножается на X^T и X, то и в произведении получим положительно полуопределенную матрицу
Таким образом, доказали, что

$$\nabla^2_\theta LogLoss(\theta) \succcurlyeq 0 \tag{12}$$

А значит, сама функция $LogLoss(\theta)$ - выпуклая

Оптимизация, поиск оптимума

Теперь выпишем итеративные способы нахождения оптимального параметра θ (нетрудно показать, что при добавлении регуляризации формулы примут следующий вид):
Итеративная формула для GD:

$$\theta_{k+1} = \theta_k - \eta \cdot (-X^T(Y - S(\theta_k))) + 2\eta\lambda\theta_k \tag{13}$$

Итеративная формула для SGD:

$$\theta_{k+1} = \theta_k - \eta \frac{n}{|I|} \sum_{i \in I} (-X_i(Y_i - S_i(\theta_k))) + 2\eta\lambda\theta_k \tag{14}$$

Оптимизация, поиск оптимума

Согласно формулам полученным ранее матрица Гессе имеет вид:

$$\nabla^2 F(\theta) = X^T V(\theta) X + 2\lambda E, \text{ где} \tag{15}$$

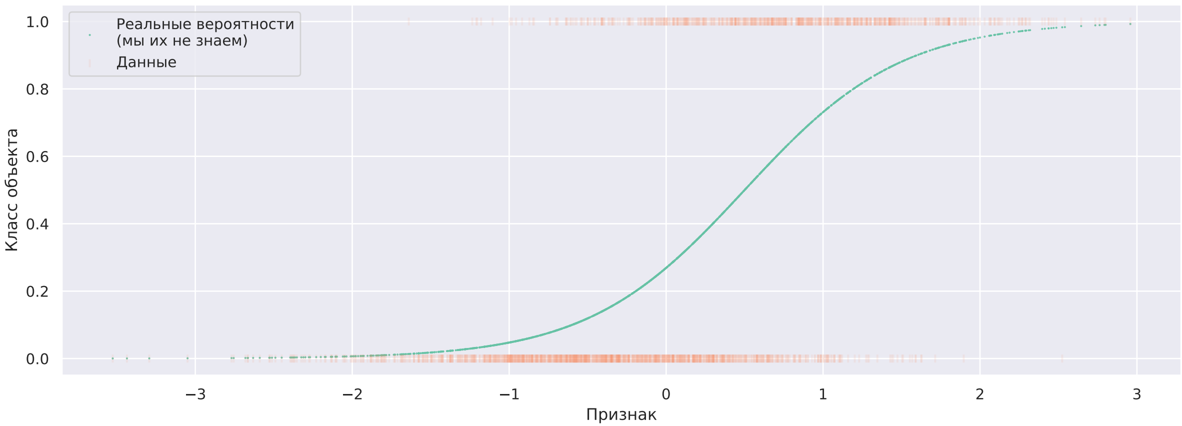
$V(\theta) = diag(\sigma(X_i^T\theta)(1 - \sigma(X_i^T\theta)))$, $E-$единичная матрица Тогда итеративная формула для IRLS:

$$\theta_{k+1} = \theta_k - (X^T V(\theta_k) X + 2\lambda E)^{-1} \cdot (-X^T (Y - S(\theta_k)) + 2\lambda\theta_k), \text{ где} \tag{16}$$

$V(\theta) = diag(\sigma(X_i^T\theta)(1 - \sigma(X_i^T\theta)))$, $E-$единичная матрица

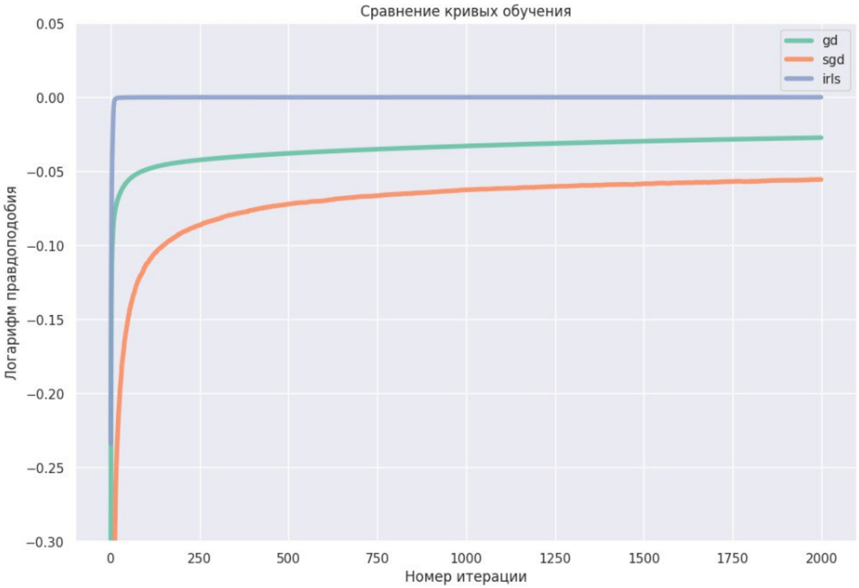
Распределение Бернулли

Для начала изобразим выборку из распределения Бернулли (Данные), которые по оси X распределены нормально, а также отрисуем вероятности принадлежности классам 0 и 1, которые мы получим при аппроксимации вероятности сигмоидой



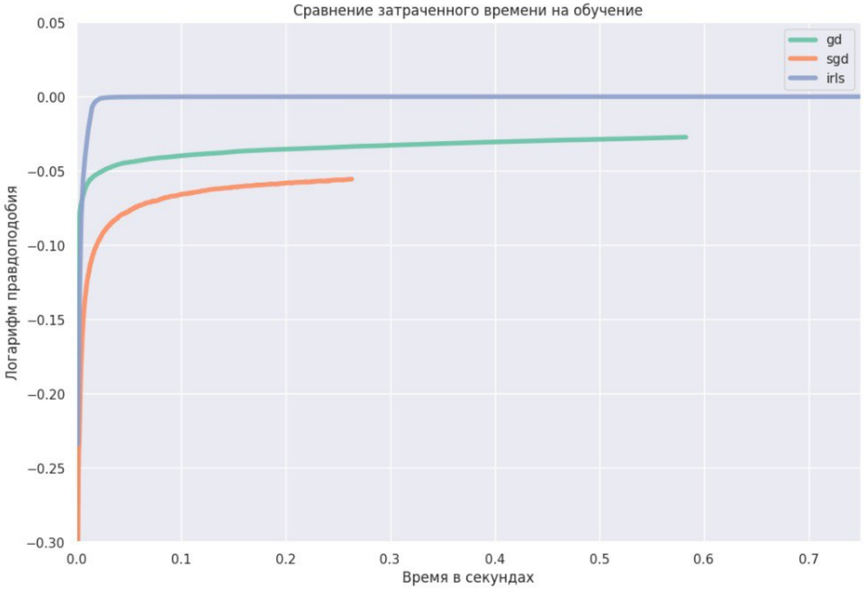
Сравнение методов оптимизации

Реализацию логистической регрессии можно посмотреть в прикрепленном файле. В ней для оптимизации функционала ошибки дано на выбор три фукнции, описанные выше. Сравним их



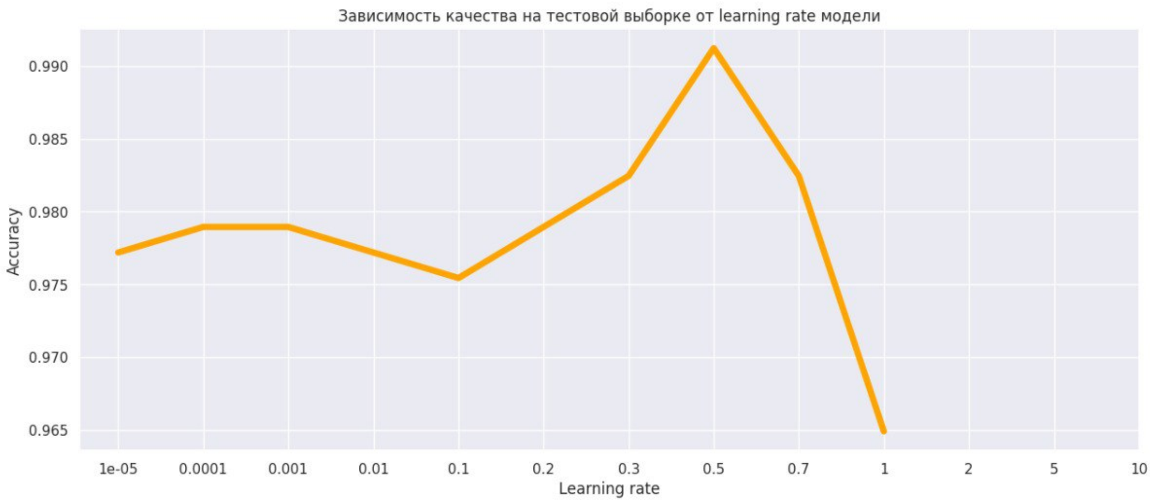
Сравнение методов оптимизации

Теперь посмотрим на затраченное время для каждого из методов



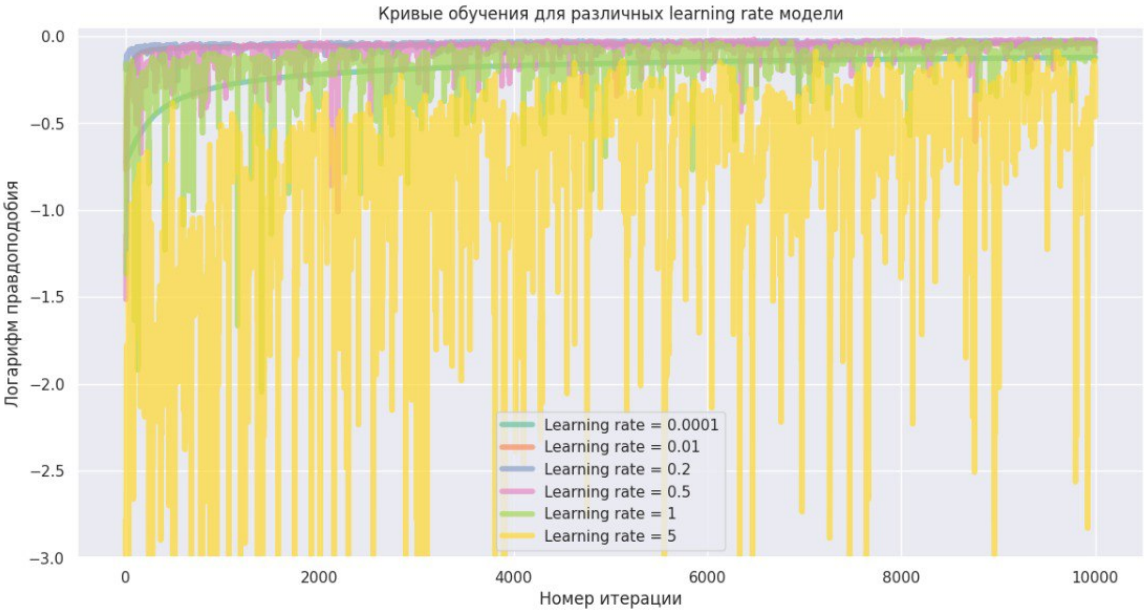
Зависимость от гиперпараметров

В данной секции исследуем зависимость качества логистической регресии (по метрике Accuracy и методу оптимизации gd) от learning rate и коэффициента регуляризации
Начнем с learning rate



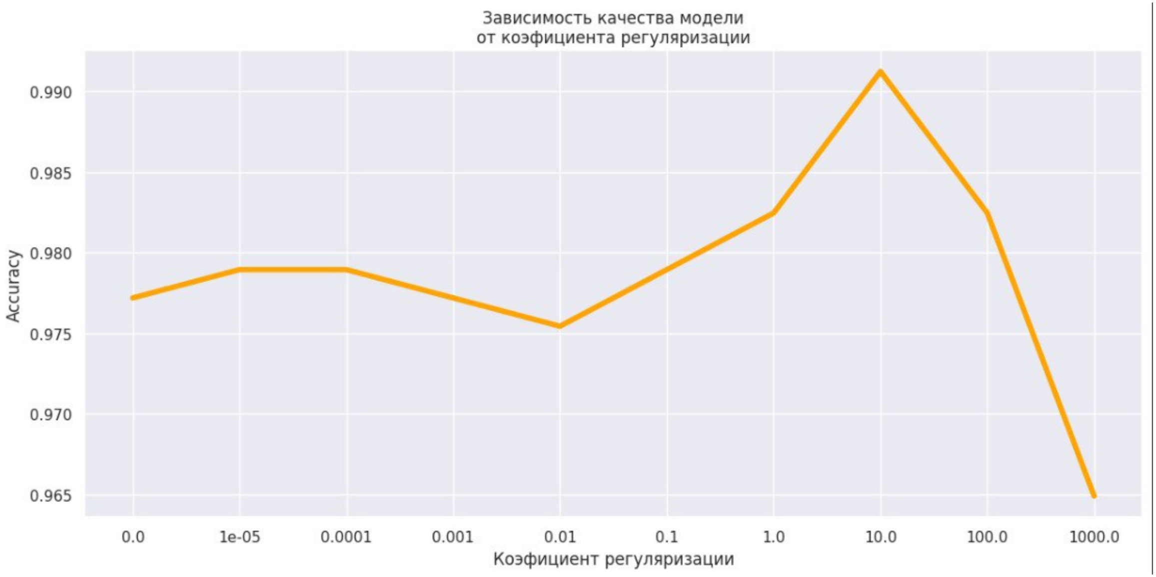
Зависимость от гиперпараметров

В данном случае важно отразить и различия в обучении для перебираемых значений lr



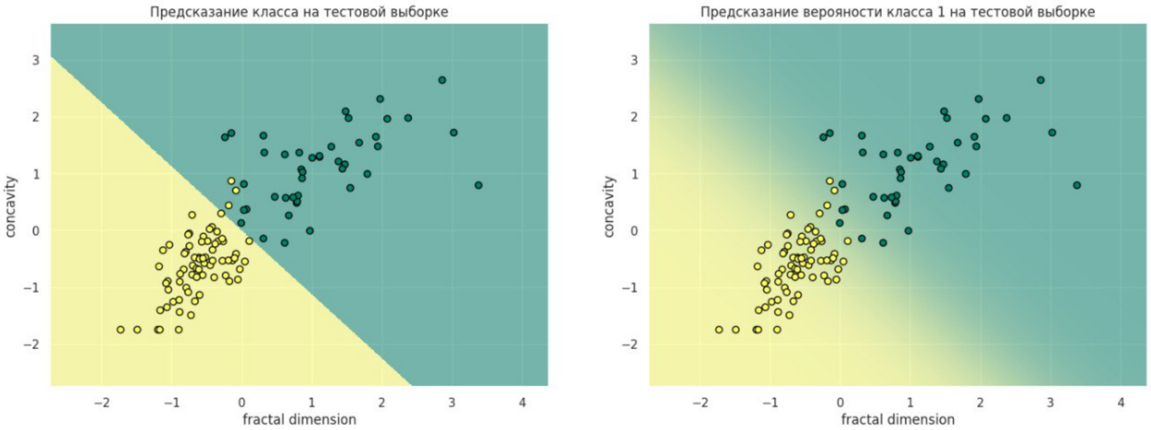
Зависимость от гиперпараметров

Теперь изучим модель при изменении параметра регуляризации



Визуализация предсказаний

Теперь непосредственно изобразим предсказания нашей модели, визуализировав как непосредственно предсказания, так и распределения вероятностей



Итог

Особенности логистической регрессии

- Логистическая регрессия может отлично справляться с задачей бинарной классификации, когда классы хорошо разделимы при помощи гиперплоскости, поскольку в её основе лежит линейная модель
- Важное предположение, которые мы делаем - вероятность хорошо аппроксимируется сигмоидой (само это утверждение конечно требует доказательство, которое строится, основываясь на Байесовском подходе)
- Нами были выведен функционал ошибки для решения задачи бинарной классификации, а также доказана его выпуклость, позволяющая решать задачу градиентными методами