Логистическая регрессия

Дано

- Множество пар объект-ответ $(x_i, y_i) = X^m \ i = \overline{1, m}$
- Объекты представляют собой множество действительных чисел $x_i \in R^n$
- Ответы представляют собой бинарное множество вида: $y_i \in \{0 \quad 1\}$

Задача: необходимо найти классификатор, который оптимально разделяет выборку, где в качестве результата представляется вероятность.

Вспомним, что из выборок строят гистограммы, которые аппроксимируют плотности. Это подсказывает нам о том, что выборку можно представить в виде элементов плотности вероятности. Также мы можем представить условную вероятность появления класса +1 как $F\left(\widehat{y_i} = 1 \mid X = x_i\right)$ - вероятность появления класса 1 при условии x_i. Класс 0 $F\left(\widehat{y_i} = 0 \mid X = x_i\right)$

Положим, что существует такая переменная – элемент плотности вероятности, представленная в виде прямой проходящей через точку ε_i, параллельной прямой разделяющей классы ($f(x) = \beta_0 + \beta_1 x$): $\hat{y}_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

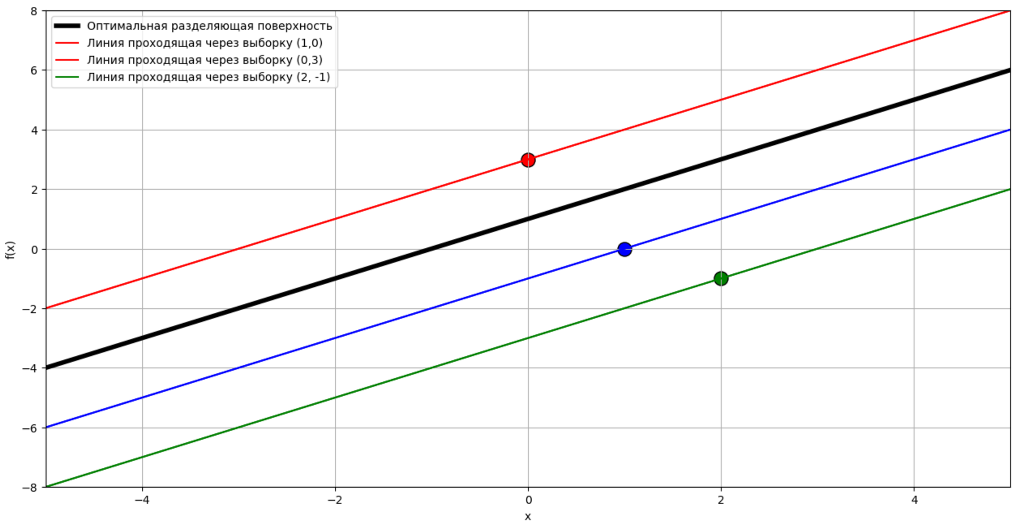

Рисунок 1 - Отображение прямых, проходящих через выборки, параллельных $f(x_i) = \beta_0 + \beta_1 x_i$ и разделяющая поверхность

Снова видно, что смещение такой прямой, можно задать величиной ε_i. Предполагается что ε_i распределена согласно логит-распределению. (предполагаю, что выбрана именно такая функция плотности, потому что она легко интегрируема и при некоторых параметрах аппроксимирует нормальную функцию плотности.

Из определения функции распределения мы знаем, что $F(x) = p\left(X \le x\right)$. Выполним замену:

$$p\left(\hat{y}_i \ge 0\right) = p\left(\beta_0 + \beta_1 x_i + \varepsilon_i \ge 0\right) = p\left(-\varepsilon_i \le \beta_0 + \beta_1 x_i\right) = p\left(\varepsilon_i \le \beta_0 + \beta_1 x_i\right) = F\left(\beta_0 + \beta_1 x_i\right)$$

Проинтегрируем логит-функцию плотности:

$$F(t) = \int_{-\infty}^{t} \frac{e^{-x}}{\left(1 + e^{-x}\right)^2} dx = \frac{e^t}{1 + e^t}$$

Выполним подстановку и получим сигмоиду:

$$F\left(\beta_0 + \beta_1 x_i\right) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

Уравнение 1 - Вероятность получить класс $y_i = 0$

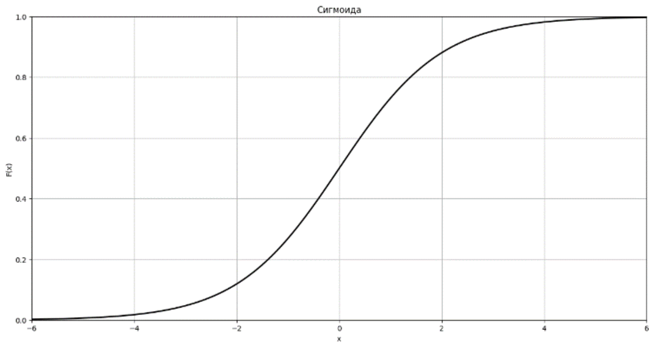

Рисунок 2 - Сигмоида

Положим, что классу $+1$ принадлежат только те точки, которые находятся выше разделяющей прямой и наоборот.

Введем отношение шансов. Это максимально удобно, поскольку это поможет разделить выборку на две части (бинарно классифицировать), сравнивая с 0.5.

$$\frac{F(y_i = 1)}{F(y_i = 0)}$$

Уравнение 2 - Логарифм отношения шансов

Если получена вероятность того, что классы находятся ниже этой прямой $F(y_i = 1)$, что есть $F\left(\beta_0 + \beta_1 x_i\right) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$, из теоремы о сумме вероятностей найдем:

$$F(y_i = 0) = 1 - \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

$$F(y_i = 0) = \frac{1}{1 + e^{\beta_0 + \beta_1 x_i}}$$

Подставим в уравнения отношения шансов:

$$\frac{F(y_i = 1)}{F(y_i = 0)} = \frac{\dfrac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}}{\dfrac{1}{1 + e^{\beta_0 + \beta_1 x_i}}}$$

$$\frac{F(y_i = 1)}{1 - F(y_i = 0)} = e^{\beta_0 + \beta_1 x_i}$$

$$\ln\left(\frac{F(y_i = 1)}{1 - F(y_i = 0)}\right) = \beta_0 + \beta_1 x_i$$

Уравнение 3- Логарифм отношения шансов

Также можно увидеть, что вероятности для плоскости представляют собой прямые, что можно увидеть на следующем рисунке

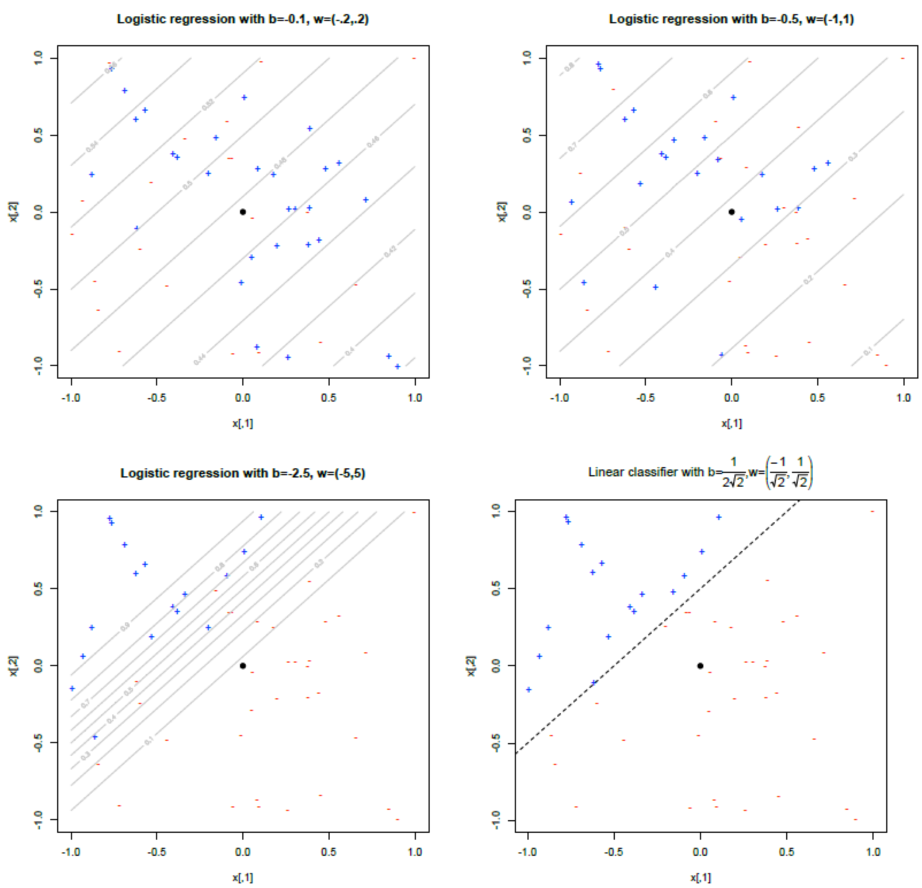

Рисунок 3 - Связь вероятности и параллельных прямых к оптимальной разделяющей плоскости

Далее необходимо, собственно, найти параметры разделяющей β. Это решается по методу максимального правдоподобия. Метод максимального правдоподобия подразумевает решение задачи о поиске таких параметров вероятности, при которых заданная выборка максимально будет похожа на генеральную совокупность. Поскольку у нас кроме наших данных нет, мы можем считать, что выборка и есть наилучшее представление генеральной совокупности и нам ничего не остается как найти искомые параметры. Для простоты, вероятности(точки выборки) будем далее обозначать, как $p(x)$ и будем считать, что они статистически независимы, тогда совместная вероятность возникновения всех событий может быть описана функцией правдоподобия:

$$L(\beta) = \prod_{i=1}^{n} p(x_i)^{y_i} \left(1 - p(x_i)\right)^{1-y_i}$$

Уравнение 4 - Функция правдоподобия

Очевидно, что нужно эту функцию максимизировать, в таком случае вероятность возникновения всех совместных событий максимальна. Максимизация осуществляется методом производных. Логарифмируя $L(\beta)$, мы избавляемся от нахождения производной по методу: $(uv)' = u'v + v'u$, что упрощает задачу:

$$\log L(\beta) = \log \prod_{i=1}^{n} p(x_i)^{y_i} \left(1 - p(x_i)\right)^{1-y_i}$$

$$\log L(\beta) = \log \prod_{i=1}^{n} p(x_i)^{y_i} + \log \prod_{i=1}^{n} \left(1 - p(x_i)\right)^{1-y_i}$$

$$\log L(\beta) = \sum_{i=1}^{n} \left\{ y_i \log p(x_i) + (1 - y_i) \log\left[1 - p(x_i)\right] \right\}$$

Проанализируем логарифм отношения правдоподобия.

Допустим $y_i = 0$, тогда выражение с заданным индексом представляет собой функцию: $\log L(\beta) = -\log\left[1 - p(x_i)\right]$(знак минус, потому что $\max(x) = \min(-x)$).

Если $y_i = 1$ то $\log L(\beta) = -\log p(x_i)$

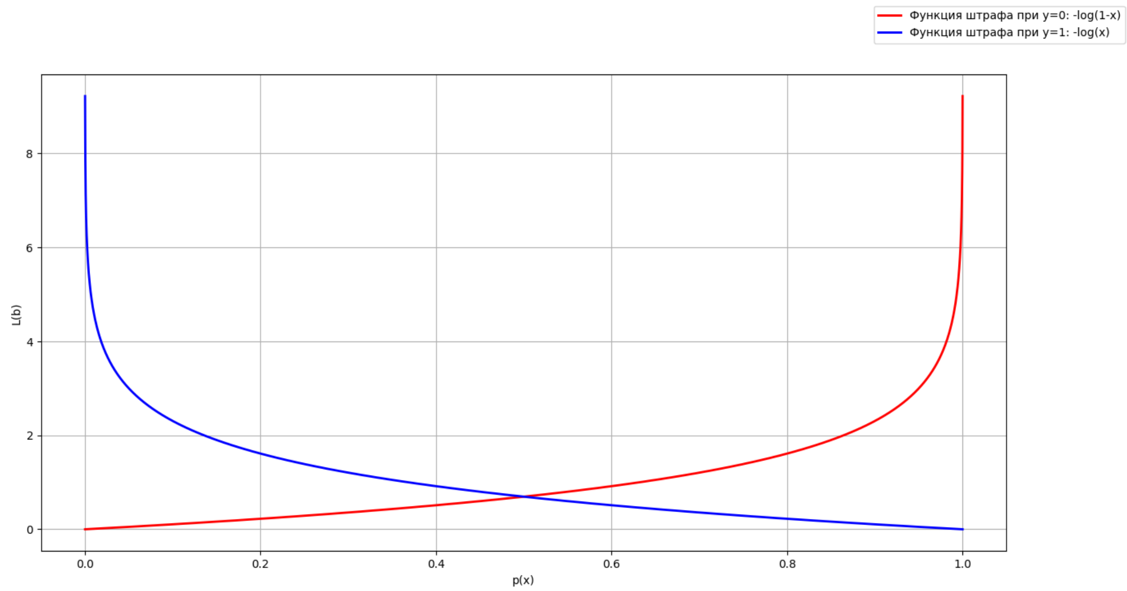

Рисунок 4 – Функция штрафа при различных значениях y_i

Таким образом чем больше $p(x_i)$ при $y_i = 0$ тем больше потери, что говорит о том, что шансы возникновения $p(x_i)$ близкой к 1 малы. При $y_i = 1$ аналогично. Кстати, сумма таких кривых с обратным знаком (логарифм отношения правдоподобия) представляет собой следующий вид:

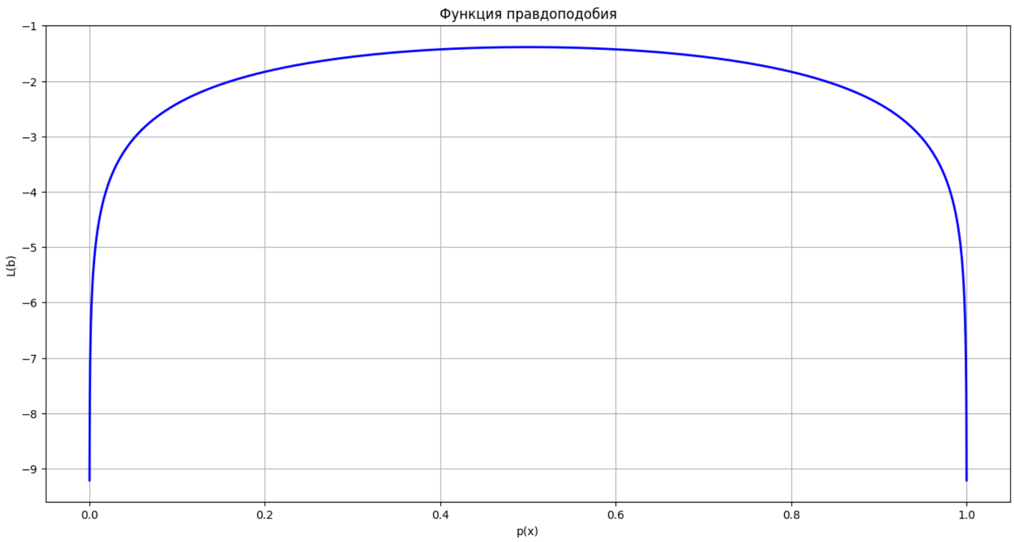

Рисунок 5 - Максимизируемая функция

Выполним замену $p\left(\beta_0 + \beta_1 x_i\right) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$ Найдем точку максимума:

$$\log L(\beta) = \sum_{i=1}^{n} \left\{ y_i \log \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} + (1-y_i) \log \left[1 - \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \right] \right\}$$

$$\log L(\beta) = \sum_{i=1}^{n} \left\{ y_i \log \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} + (1-y_i) \log \left[\frac{1}{1+e^{\beta_0+\beta_1 x_i}} \right] \right\}$$

$$\log L(\beta) = \sum_{i=1}^{n} \left\{ y_i \left(\log e^{\beta_0+\beta_1 x_i} - \log\left(1+e^{\beta_0+\beta_1 x_i}\right) \right) + (1-y_i) \left(\log 1 - \log\left(1+e^{\beta_0+\beta_1 x_i}\right) \right) \right\}$$

$$\log L(\beta) = \sum_{i=1}^{n} \left\{ y_i \left(\log e^{\beta_0+\beta_1 x_i} - \log\left(1+e^{\beta_0+\beta_1 x_i}\right) \right) + (y_i - 1) \log\left(1+e^{\beta_0+\beta_1 x_i}\right) \right\}$$

$$\log L(\beta) = \sum_{i=1}^{n} \left\{ y_i \log e^{\beta_0+\beta_1 x_i} - y_i \log\left(1+e^{\beta_0+\beta_1 x_i}\right) + y_i \log\left(1+e^{\beta_0+\beta_1 x_i}\right) - \log\left(1+e^{\beta_0+\beta_1 x_i}\right) \right\}$$

$$\log L(\beta) = \sum_{i=1}^{n} \left\{ y_i \left(\beta_0 + \beta_1 x_i\right) - \log\left(1+e^{\beta_0+\beta_1 x_i}\right) \right\}$$

$$\log L(\beta) = \sum_{i=1}^{n} y_i \left(\beta_0 + \beta_1 x_i\right) - \sum_{i=1}^{n} \log\left(1+e^{\beta_0+\beta_1 x_i}\right)$$

$$\log L(\beta) = \sum_{i=1}^{n} y_i \left(\beta^T x_i\right) - \sum_{i=1}^{n} \log\left(1+e^{\beta^T x_i}\right)$$

$$\frac{\partial \log L(\beta)}{\partial \beta} = \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \frac{x}{1+e^{\beta^T x_i}}$$

$$\frac{\partial \log L(\beta)}{\partial \beta} = \sum_{i=1}^{n} x_i \left(y_i - \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}} \right)$$

Уравнение 5 – Найденный градиент

Пусть β^T - вектор и X_i многомерна, в этом случае, решение не изменится.

Метод Ньютона-Рафсона

Идея метода Ньютона-Рафсона базируется на идее касательных: $tg(\alpha) = f'(x)$

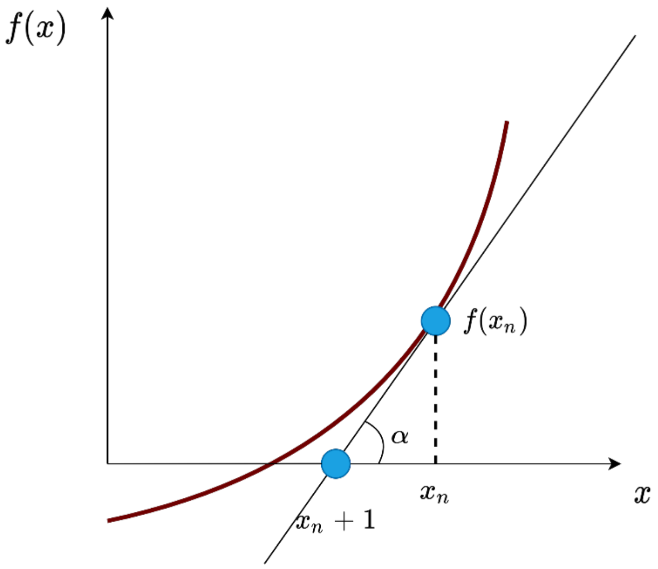

Рисунок 6 - Геометрическая интерпретация - метода Ньютона-Рафсона

$$f'(x_n) = \frac{f(x_n)}{x_{n+1} - x_n}$$

Уравнение 6 - Тангенс угла наклона касательной

Полагая, что $f(x_{n+1}) = 0$, то $f'(x_n) = \dfrac{f(x_n)}{x_n - x_{n+1}}$

$$x_n - x_{n+1} = \frac{f(x_n)}{f'(x_n)}$$

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

Уравнение 7 – Поиск решение для метода-Ньютона-Рафсона

Далее, поскольку непосредственный поиск производной логарифма отношения правдоподобия – задача сложная, мы можем разложить его в ряд Тейлора до 3 порядка, извлекая таким образом оптимизируемый аргумент β^T

$$f(x) \approx f(x_k) + \nabla f(x_k)(x - x_k)^T + \frac{1}{2}(x - x_k)^T H(x_k)(x - x_k)$$

Уравнение 8 - Многомерное разложение в ряд Тейлора

где

$H(x_k)$ - Гессиан; $\nabla f(x_k)$ - Градиент

Найдем производную по $(x - x_0)$

$$\nabla f(x_k) + H(x_k)(x - x_k) = 0$$
$$\nabla f(x_k) + H(x_k)x - H(x_k)x_k = 0$$
$$x_{k+1} = x_k - H^{-1}(x_k)\nabla f(x_k)$$

Уравнение 9 - Метод Ньютона-Рафсона

Найдем Гессиан логафрима правдоподобия

$$\frac{\partial^2 \log L(\beta)}{\partial\beta\partial\beta^T} = \frac{\partial}{\partial\beta^T}\left(\sum_{i=1}^{n} x_i\left(y_i - \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}\right)\right)$$

$$\frac{\partial^2 \log L(\beta)}{\partial\beta\partial\beta^T} = \frac{\partial}{\partial\beta^T}\left(\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i x_i \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}\right)$$

$$\frac{\partial^2 \log L(\beta)}{\partial\beta\partial\beta^T} = -\sum_{i=1}^{n} y_i x_i \left(\frac{x_i e^{\beta^T x_i}\left(1+e^{\beta^T x_i}\right) - x_i e^{2\beta^T x_i}}{\left(1+e^{\beta^T x_i}\right)^2}\right)$$

$$\frac{\partial^2 \log L(\beta)}{\partial\beta\partial\beta^T} = -\sum_{i=1}^{n} y_i x_i \left(\frac{x_i e^{\beta^T x_i} + x_i e^{2\beta^T x_i} - x_i e^{2\beta^T x_i}}{\left(1+e^{\beta^T x_i}\right)^2}\right)$$

$$\frac{\partial^2 \log L(\beta)}{\partial\beta\partial\beta^T} = -\sum_{i=1}^{n} y_i x_i \left(\frac{x_i e^{\beta^T x_i}}{\left(1+e^{\beta^T x_i}\right)} \times \frac{1}{\left(1+e^{\beta^T x_i}\right)}\right)$$

$$\frac{\partial^2 \log L(\beta)}{\partial\beta\partial\beta^T} = -\sum_{i=1}^{n} y_i x_i \left(\frac{x_i e^{\beta^T x_i}}{\left(1+e^{\beta^T x_i}\right)} \times \frac{1}{\left(1+e^{\beta^T x_i}\right)}\right)$$

$$\frac{\partial^2 \log L(\beta)}{\partial\beta\partial\beta^T} = -\sum_{i=1}^{n} y_i x_i x_i^T p\left(\beta^T x_i\right)\left(1 - p\left(\beta^T x_i\right)\right)$$

Выведем алгоритм Ньютона-Рафсона для логистической регрессии:

Обозначим размерности векторов

$X^T_{n\times(p+1)}$ - матрица признаков

y - вектор ответов

p - вектор вероятностей

$W_{n\times n}$ - вектор $p\left(\beta^T x_i\right)\left(1 - p\left(\beta^T x_i\right)\right)$ с диагональным элементом

$$\frac{\partial \log L(\beta)}{\partial\beta} = X^T\left(y - p\right)$$

$$\frac{\partial^2 \log L(\beta)}{\partial\beta\partial\beta^T} = -X^T W X$$

$$x_{k+1} = x_k - H^{-1}\left(x_k\right)\nabla f\left(x_k\right)(замена)$$

$$\beta_{k+1} = \beta_k + \left[X^T W X\right]^{-1} X^T\left(y-p\right)(вынесем\_за\_скобки\_\left[X^T W X\right]^{-1})$$

$$\beta_{k+1} = \left[X^T W X\right]^{-1}\left[X^T W X\right]\beta_k + \left[X^T W X\right]^{-1} X^T\left(y-p\right)(вынесем\_за\_скобки\_\left[X^T W X\right]^{-1})$$

$$\beta_{k+1} = \left[X^T W X\right]^{-1}\left\{\left[X^T W X\right]\beta_k + X^T\left(y-p\right)\right\}(умножим\_на\_WW^{-1})$$

$$\beta_{k+1} = \left[X^T W X\right]^{-1}\left\{X^T W X\beta_k + X^T W W^{-1}\left(y-p\right)\right\}(вынесем\_за\_скобки\_X^T W)$$

$$\beta_{k+1} = \left[X^T W X\right]^{-1} X^T W\left[X\beta_k + W^{-1}\left(y-p\right)\right](замена)$$

$$\beta_{k+1} = \left[X^T W X\right]^{-1} X^T W z$$

Таким образом алгоритм Ньютона-Рафсона состоит из двух этапов:

$$z_k = \left[X\beta_k + W^{-1}\left(y-p\right)\right]$$
$$\beta_{k+1} = \left[X^T W X\right]^{-1} X^T W z_k$$

Уравнение 10 - Алгоритм Ньютона-Рафсона для логистической регрессии