Логистическая регрессия

Машинное обучение

Наиль Баширов

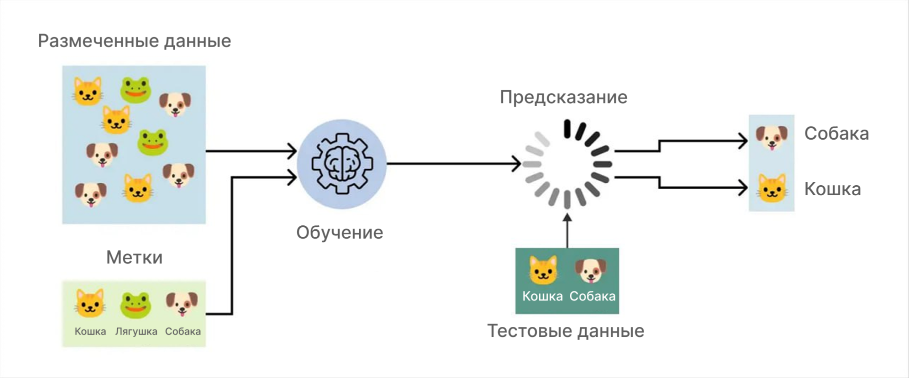
Московский физико-технический институт

21 октября 2025

ФПМИ

Примеры задач классификации (напоминание)

Классификация изображений



Постановка задачи (напоминание)

Задача. *Бинарная классификация:*

- Пусть $\mathcal{X} = \mathbb{R}^d$ пространство объектов;
- Пусть $\mathcal{Y} = \{-1, 1\}$ (либо $\{0, 1\}$) множество допустимых ответов;
- $X = \{(x^i, y^i)\}_{i=1}^n$ - обучающая выборка.

Линейная модель классификации (напоминание)

Определение

Линейная модель классификации определяется следующим образом:

$$\operatorname{sign}\left(\langle w, x\rangle + w_0\right) = \operatorname{sign}\left(\sum_{j=1}^{d} w_j x_j + w_0\right),$$

где $w \in \mathbb{R}^d$ - вектор весов, $w_0 \in \mathbb{R}$ - сдвиг.

Линейная модель классификации (напоминание)

Определение

Линейная модель классификации определяется следующим образом:

$$\operatorname{sign}\left(\langle w, x\rangle + w_0\right) = \operatorname{sign}\left(\sum_{j=1}^{d} w_j x_j + w_0\right),$$

где $w \in \mathbb{R}^d$ - вектор весов, $w_0 \in \mathbb{R}$ - сдвиг.

Замечание. Если предположить, что в данных есть $x_0 = 1$, то нет необходимости вводить сдвиг w_0, т.е. останется только $\operatorname{sign}\left(\langle w, x\rangle\right)$.

Функция потерь (напоминание)

Вопрос: *Как обучать?*

Функция потерь (напоминание)

Вопрос: *Как обучать?*
Ответ: *Максимизировать долю правильных ответов:*

Доля правильных ответов (accuracy)

$$\max_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[\text{sign}(\langle w, x^i \rangle) = y^i]. \tag{1}$$

Функция потерь (напоминание)

Вопрос: *Как обучать?*
Ответ: *Максимизировать долю правильных ответов:*

Доля правильных ответов (accuracy)

$$\max_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[\operatorname{sign}(\langle w, x^i \rangle) = y^i]. \tag{1}$$

Или эквивалентно минимизировать долю неверных ответов

$$\max_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[\operatorname{sign}(\langle w, x^i \rangle) = y^i] = 1 - \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[\operatorname{sign}(\langle w, x^i \rangle) \neq y^i]$$

$$\Rightarrow \quad \min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[\operatorname{sign}(\langle w, x^i \rangle) \neq y^i].$$

Как обучать? (напоминание)

Задача оптимизации:

$$\min_{w \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[\mathrm{sign}\left(\langle w, x^i \rangle\right) \neq y^i] \right\}.$$

Вопрос: Какие могут быть сложности при обучении в такой постановке задачи?

Как обучать? (напоминание)

Задача оптимизации:

$$\min_{w \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[\text{sign}\left(\langle w, x^i \rangle\right) \neq y^i] \right\}.$$

Вопрос: Какие могут быть сложности при обучении в такой постановке задачи?

Проблемы:

- Целевая функция дискретна относительно весов.
- Возможно наличие множества глобальных минимумов

Как обучать? (напоминание)

Задача оптимизации:

$$\min_{w \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[\mathrm{sign}\left(\langle w, x^i \rangle\right) \neq y^i] \right\}.$$

Вопрос: Какие могут быть сложности при обучении в такой постановке задачи?

Проблемы:

- Целевая функция дискретна относительно весов.
- Возможно наличие множества глобальных минимумов

Решение: Свести задачу к минимизации гладкого функционала.

Отступы (напоминание)

Задача оптимизации:

$$\min_{w \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[\text{sign}\left(\langle w, x^i \rangle\right) \neq y^i] \right\}.$$

Наблюдение: Заметим, что

$$\begin{aligned} y^i \cdot \langle w, x^i \rangle &> 0, \text{ если } y^i = \text{sign}\left(\langle w, x^i \rangle\right); \\ y^i \cdot \langle w, x^i \rangle &< 0, \text{ иначе.} \end{aligned}$$

Величина $M_i = y^i \cdot \langle w, x^i \rangle$ называется *отступом*.

Верхние оценки

Задача оптимизации

$$\min_w \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[M_i < 0], \text{ где } M_i = y^i \cdot \langle w, x^i \rangle$$

.

Замечание: Мы можем оценить негладкую функцию индикатора $h(M) = \mathbb{I}[M_i < 0]$ сверху гладкой функций $\widetilde{h}(M)$, т.е.

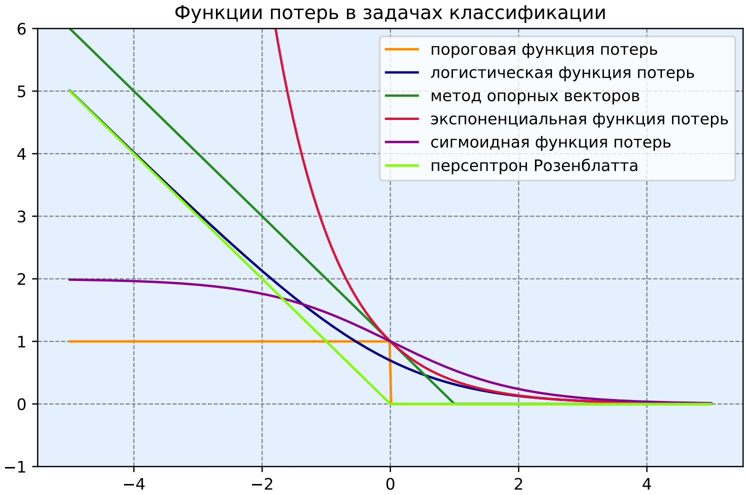
$$h(M) \ \leq \ \widetilde{h}(M).$$

Верхние оценки

Примеры верхних оценок

- $\widetilde{h}(M) = \log\left(1 + e^{-M}\right)$ - логистическая функция потерь;
- $\widetilde{h}(M) = (1 - M)_+ = \max\{0, 1 - M\}$ - кусочно-линейная функция потерь (используется в методе опорных векторов);
- $\widetilde{h}(M) = (-M)_+ = \max\{0, -M\}$ - кусочно-линейная функция потерь (соответствует персептрону Розенблатта);
- $\widetilde{h}(M) = e^{-M}$ - экспоненциальная функция потерь;
- $\widetilde{h}(M) = \frac{2}{1+e^M}$ - сигмоидная функция потерь.

Верхние оценки: визуализация примеров



Логистическая регрессия

Задача оптимизации:

$$\min_{w \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^{n} \widetilde{h}\left(y^i \langle w, x^i \rangle\right) \right\},$$

где для логистической регрессии $\widetilde{h}(M) = \log\left(1 + e^{-M}\right)$.

Логистическая регрессия

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \exp\left(-y^i \langle w, x^i \rangle\right)\right)$$

Логистическая регрессия

Задача оптимизации:

$$\min_{w \in \mathbb{R}^d} \left\{ \mathcal{L}(w) = \frac{1}{n} \sum_{i=1}^{n} \underbrace{\log\left(1 + exp\left(-y^i \langle w, x^i \rangle\right)\right)}_{=\ell_i(g(x^i, w), y^i)} \right\}.$$

Некоторые свойства:

- Каждая функция ℓ_i является выпуклой и $\frac{\|x^i\|^2}{4}$-гладкой;
- Функция $\mathcal{L}$ является выпуклой и $\frac{1}{4}\lambda_{\max}\left(\frac{1}{n}\sum\limits_{i=1}^{n} x^i (x^i)^\top\right)$-гладкой.

Оценивание вероятностей

Утверждение

Предсказание логистической регрессии можно интерпретировать как вероятность принадлежности объекта к каждому из классов.

Оценивание вероятностей

Утверждение

Предсказание логистической регрессии можно интерпретировать как вероятность принадлежности объекта к каждому из классов.

- Зафиксируем $x \in \mathcal{X}$;
- $p(y = 1|x)$ - вероятность того, что объект x будет принадлежать классу 1;
- Модель $g(x)$ возвращает числа из отрезка $[0, 1]$.

Цель: выбрать для него такую процедуру обучения, что в точке x ему будет оптимально выдавать число $p(y = 1|x)$.

Оценивание вероятностей

Если в выборке объект x встречается m раз с ответом $\{y_1, \ldots, y_m\}$, то имеем следующее требование

$$\underset{\hat{y}\in\mathbb{R}}{\operatorname{argmin}} \frac{1}{m}\sum_{i=1}^{m} \ell(y^i, \hat{y}) \approx p(y = 1|x).$$

При стремлении m к бесконечности получим, что функционал стремится к матожиданию ошибки:

$$\underset{\hat{y}\in\mathbb{R}}{\operatorname{argmin}} \mathbb{E}\left[\ell(y^i, \hat{y})|x\right] = p(y = 1|x).$$

Оценивание вероятностей

Пример 1

Покажите, что при использовании $\ell(y^i, \hat{y}) = (\mathbb{I}[y^i = +1] - \hat{y})^2$ (квадратичной функции потерь) $\hat{y}$ принимает значения от 0 до 1, которые можно интерпретировать как вероятность.

Оценивание вероятностей

Пример 1

Покажите, что при использовании $\ell(y^i, \hat{y}) = (\mathbb{I}[y^i = +1] - \hat{y})^2$ (квадратичной функции потерь) $\hat{y}$ принимает значения от 0 до 1, которые можно интерпретировать как вероятность.

Запишем матожидание функции потерь в точке x:

$$\begin{aligned} \mathbb{E}\left[\ell(y^i, \hat{y})|x\right] &= \mathbb{E}\left[\mathbb{I}[y^i = +1](1-\hat{y})^2 + \mathbb{I}[y^i = -1]\hat{y}^2|x\right] \\ &= p(y^i = +1|x)(1-\hat{y})^2 + \left(1 - p(y^i = +1|x)\right)\hat{y}^2. \end{aligned}$$

Оценивание вероятностей

Пример 1

Покажите, что при использовании $\ell(y^i, \hat{y}) = (\mathbb{I}[y^i = +1] - \hat{y})^2$ (квадратичной функции потерь) $\hat{y}$ принимает значения от 0 до 1, которые можно интерпретировать как вероятность.

Матожидание функции потерь в точке x:

$$\mathbb{E}\left[\ell(y^i, \hat{y})|x\right] = p(y^i = +1|x)(1 - \hat{y})^2 + \left(1 - p(y^i = +1|x)\right)\hat{y}^2.$$

Продифференцируем по $\hat{y}$:

$$\begin{aligned}\frac{\partial}{\partial \hat{y}}\mathbb{E}\left[\ell(y^i, \hat{y})|x\right] &= 2p(y^i = +1|x)(\hat{y} - 1) + 2\left(1 - p(y^i = +1|x)\right)\hat{y} \\ &= 2(\hat{y} - p(y^i = +1|x)) = 0.\end{aligned}$$

Оценивание вероятностей

Пример 1

Покажите, что при использовании $\ell(y^i, \hat{y}) = (\mathbb{I}[y^i = +1] - \hat{y})^2$ (квадратичной функции потерь) $\hat{y}$ принимает значения от 0 до 1, которые можно интерпретировать как вероятность.

Матожидание функции потерь в точке x:

$$\mathbb{E}\left[\ell(y^i, \hat{y})|x\right] = p(y^i = +1|x)(1 - \hat{y})^2 + \left(1 - p(y^i = +1|x)\right)\hat{y}^2.$$

Продифференцируем по $\hat{y}$:

$$\begin{aligned}\frac{\partial}{\partial \hat{y}}\mathbb{E}\left[\ell(y^i, \hat{y})|x\right] &= 2p(y^i = +1|x)(\hat{y} - 1) + 2\left(1 - p(y^i = +1|x)\right)\hat{y} \\ &= 2(\hat{y} - p(y^i = +1|x)) = 0.\end{aligned}$$

Легко видеть, что оптимальный ответ модели действительно равен вероятности: $\hat{y} = p(y^i = +1|x)$

Оценивание вероятностей

Пример 2

Покажите, что для абсолютной функции потерь

$$\ell(y^i, \hat{y}) = |\mathbb{I}[y^i = +1] - \hat{y}|,$$

$\hat{y} \in [0; 1]$ нельзя интерпретировать как вероятность.

Оценивание вероятностей

Пример 2

Покажите, что для абсолютной функции потерь

$$\ell(y^i, \hat{y}) = |\mathbb{I}[y^i = +1] - \hat{y}|,$$

$\hat{y} \in [0; 1]$ нельзя интерпретировать как вероятность.

Запишем матожидание функции потерь в точке x:

$$\begin{aligned}
\mathbb{E}\left[\ell(y^i, \hat{y})|x\right] &= \mathbb{E}\left[\mathbb{I}[y^i = +1]|1 - \hat{y}| + \mathbb{I}[y^i = -1]|\hat{y}| \mid x\right] \\
&= p(y^i = +1|x)(1 - \hat{y}) + \left(1 - p(y^i = +1|x)\right)\hat{y}.
\end{aligned}$$

Оценивание вероятностей

Пример 2

Покажите, что для абсолютной функции потерь

$$\ell(y^i, \hat{y}) = |\mathbb{I}[y^i = +1] - \hat{y}|,$$

$\hat{y} \in [0; 1]$ нельзя интерпретировать как вероятность.

Матожидание функции потерь в точке x:

$$\mathbb{E}\left[\ell(y^i, \hat{y})|x\right] = p(y^i = +1|x)(1 - \hat{y}) + \left(1 - p(y^i = +1|x)\right)\hat{y}.$$

Продифференцируем по $\hat{y}$:

$$\frac{\partial}{\partial \hat{y}}\mathbb{E}\left[\ell(y^i, \hat{y})|x\right] = -p(y^i = +1|x) + \left(1 - p(y^i = +1|x)\right)$$

Оценивание вероятностей

Пример 2

Покажите, что для абсолютной функции потерь

$$\ell(y^i, \hat{y}) = |\mathbb{I}[y^i = +1] - \hat{y}|,$$

$\hat{y} \in [0; 1]$ нельзя интерпретировать как вероятность.

Матожидание функции потерь в точке x:

$$\mathbb{E}\left[\ell(y^i, \hat{y})|x\right] = p(y^i = +1|x)(1 - \hat{y}) + \left(1 - p(y^i = +1|x)\right)\hat{y}.$$

Продифференцируем по $\hat{y}$:

$$\begin{aligned}\frac{\partial}{\partial \hat{y}}\mathbb{E}\left[\ell(y^i, \hat{y})|x\right] &= -p(y^i = +1|x) + \left(1 - p(y^i = +1|x)\right) \\ &= 1 - 2p(y^i = +1|x) = 0.\end{aligned}$$

Оценивание вероятностей

Пример 2

Покажите, что для абсолютной функции потерь

$$\ell(y^i, \hat{y}) = |\mathbb{I}[y^i = +1] - \hat{y}|,$$

$\hat{y} \in [0; 1]$ нельзя интерпретировать как вероятность.

Рассмотрим 2 случая:

- $p(y^i = +1|x) = \frac{1}{2} \Rightarrow$ классификатор не позволяет предсказывать корректную вероятность в точке x (Почему?);
- $p(y^i = +1|x) \neq \frac{1}{2} \Rightarrow$ классификатор также не позволяет предсказывать корректную вероятность в точке (Почему?).

Правдоподобие: напоминание

Пусть $X \times Y$ — в.п. с плотностью $p(x, y|w) = \mathbb{P}(y|x, w)p(x)$.
Пусть X^n — простая (i.i.d.) выборка: $(x^i, y^i)_{i=1}^n \sim p(x, y|w)$
Оценка максимального правдоподобия для w:

$$\prod_{i=1}^n p(x^i, y^i|w) = \prod_{i=1}^n \mathbb{P}(y^i|x^i, w)p(x^i) \to \max_w$$

Логарифм правдоподобия (log-likelihood, log-loss):

$$L(w) = \sum_{i=1}^n \log \mathbb{P}(y^i|x^i, w) \to \max_w$$

В случае двух классов $y^i \in Y = \{0, 1\}$, можно рассмотреть $\mathbb{P}(y = 1|x, w)$
Тогда логарифм правдоподобия принимает вид:

$$L(w) = \sum_{i=1}^n y^i \log \mathbb{P}(y^i = 1|x^i, w) + (1 - y^i) \log\left(1 - \mathbb{P}(y^i = 1|x^i, w)\right) \to \max_w$$

Правдоподобие

Пример 3

Покажите, что при минимизации $-\log \mathrm{likelihood}$:

$$\min_w \left\{ -\sum_{i=1}^{n} \left[\mathbb{I}[y^i = +1] \log g(x^i, w) + \mathbb{I}[y^i = -1] \log \left(1 - g(x^i, w)\right) \right] \right\}$$

$\hat{y} = g(x^i, w)$ принимает значения от 0 до 1, которые можно интерпретировать как вероятность.

Правдоподобие

Пример 3

Покажите, что при минимизации $-\log \mathrm{likelihood}$:

$$\min_w \left\{ -\sum_{i=1}^{n} \left[\mathbb{I}[y^i = +1] \log g(x^i, w) + \mathbb{I}[y^i = -1] \log \left(1 - g(x^i, w)\right) \right] \right\}$$

$\hat{y} = g(x^i, w)$ принимает значения от 0 до 1, которые можно интерпретировать как вероятность.

Покажем, что она также позволяет получить значения, котороые можно интерпретировать как вероятности:
Функция потерь:

$$\ell(y^i, \hat{y}) = -\mathbb{I}[y^i = +1] \log \hat{y} - \mathbb{I}[y^i = -1] \log (1 - \hat{y})$$

Правдоподобие

Пример 3

Покажите, что при минимизации $-\log \mathrm{likelihood}$:

$$\min_w \left\{ -\sum_{i=1}^{n} \left[\mathbb{I}[y^i = +1] \log g(x^i, w) + \mathbb{I}[y^i = -1] \log \left(1 - g(x^i, w)\right) \right] \right\}$$

$\hat{y} = g(x^i, w)$ принимает значения от 0 до 1, которые можно интерпретировать как вероятность.

Матожидание функции потерь в точке x:

$$\mathbb{E}\left[\ell(y^i, \hat{y}) | x^i\right] = -\mathbb{P}(y^i = +1 | x^i) \log \hat{y} - \left(1 - \mathbb{P}(y^i = +1 | x^i)\right) \log(1 - \hat{y}).$$

Продифференцируем по $\hat{y}$:

$$\frac{\partial}{\partial \hat{y}} \mathbb{E}\left[\ell(y^i, \hat{y}) | x^i\right] = -\frac{\mathbb{P}(y^i = +1 | x^i)}{\hat{y}} + \frac{1 - \mathbb{P}(y^i = +1 | x^i)}{1 - \hat{y}} = 0.$$

Правдоподобие

Пример 3

Покажите, что при минимизации $-\log \mathrm{likelihood}$:

$$\min_w \left\{ -\sum_{i=1}^{n} \left[\mathbb{I}[y^i = +1] \log g(x^i, w) + \mathbb{I}[y^i = -1] \log \left(1 - g(x^i, w)\right) \right] \right\}$$

$\hat{y} = g(x^i, w)$ принимает значения от 0 до 1, которые можно интерпретировать как вероятность.

Матожидание функции потерь в точке x:

$$\mathbb{E}\left[\ell(y^i, \hat{y}) | x^i\right] \quad = \quad -\mathbb{P}(y^i = +1 | x^i) \log \hat{y} - \left(1 - \mathbb{P}(y^i = +1 | x^i)\right) \log(1 - \hat{y}).$$

Продифференцируем по $\hat{y}$:

$$\frac{\partial}{\partial \hat{y}} \mathbb{E}\left[\ell(y^i, \hat{y}) | x^i\right] \quad = \quad -\frac{\mathbb{P}(y^i = +1 | x^i)}{\hat{y}} + \frac{1 - \mathbb{P}(y^i = +1 | x^i)}{1 - \hat{y}} = 0.$$

Оптимальный ответ модели - вероятность класса «+1»: $\hat{y} = \mathbb{P}(y^i = +1 | x^i)$.

Связь правдоподобия и аппроксимации эмпирического риска

Пусть $X \times Y$ — вероятностное пространство с плотностью $p(x, y|w) = \mathbb{P}(y|x, w)p(x)$.

Пусть X^n — простая (i.i.d.) выборка: $(x^i, y^i)_{i=1}^n \sim p(x, y|w)$

- Максимизация правдоподобия (Maximum Likelihood, ML):

$$L(w) = \sum_{i=1}^{n} \log P(y^i|x^i, w) \to \max_{w}$$

Связь правдоподобия и аппроксимации эмпирического риска

Пусть $X \times Y$ — вероятностное пространство с плотностью $p(x, y|w) = \mathbb{P}(y|x, w)p(x)$.

Пусть X^n — простая (i.i.d.) выборка: $(x^i, y^i)_{i=1}^n \sim p(x, y|w)$

- Максимизация правдоподобия (Maximum Likelihood, ML):

$$L(w) = \sum_{i=1}^{n} \log P(y^i|x^i, w) \to \max_w$$

- Минимизация аппроксимированного эмпирического риска:

$$\mathcal{L}(w) = \sum_{i=1}^{n} \ell(y^i, g(x^i, w)) \to \min_w$$

Связь правдоподобия и аппроксимации эмпирического риска

Пусть $X \times Y$ — вероятностное пространство с плотностью $p(x, y|w) = \mathbb{P}(y|x, w)p(x)$.
Пусть X^n — простая (i.i.d.) выборка: $(x^i, y^i)_{i=1}^n \sim p(x, y|w)$

- Максимизация правдоподобия (Maximum Likelihood, ML):

$$L(w) = \sum_{i=1}^{n} \log P(y^i|x^i, w) \to \max_w$$

- Минимизация аппроксимированного эмпирического риска:

$$\mathcal{L}(w) = \sum_{i=1}^{n} \ell(y^i, g(x^i, w)) \to \min_w$$

Эти два принципа эквивалентны, если $-\log P(y^i|x^i, w) = \ell(y^i, g(x^i, w))$.

Вероятностный смысл регуляризации

$P(y|x, w)$ - вероятностная модель данных;

$p(w; \gamma)$ - априорное распределение параметров модели;

γ - вектор *гиперпараметров*;

Теперь не только появление выборки X^ℓ, но и появление модели w также полагается стохастическим.
Совместное правдоподобие данных и модели:

$$p(X^\ell, w) = p(X^\ell|w)\, p(w; \gamma).$$

Принцип максимума апостериорной вероятности (Maximum a Posteriori Probability, MAP):

$$L(w) = \ln p(X^\ell, w) = \sum_{i=1}^{\ell} \log P(y_i|x_i, w) + \underbrace{\log p(w; \gamma)}_{\text{регуляризатор}} \to \max_w .$$

Логистическая регрессия

• Чтобы модель $g(w)$ возвращала числа из отрезка $[0, 1]$, можно положить

$$g(w) = \sigma\left(w_0 + \langle w, x\rangle\right),$$

где σ - некая непрерывная слева неубывающая функция со значениями в отрезке $[0, 1]$.

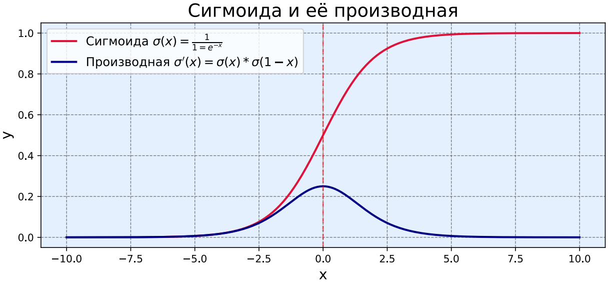
Логистическая регрессия

- Чтобы модель $g(w)$ возвращала числа из отрезка $[0, 1]$, можно положить

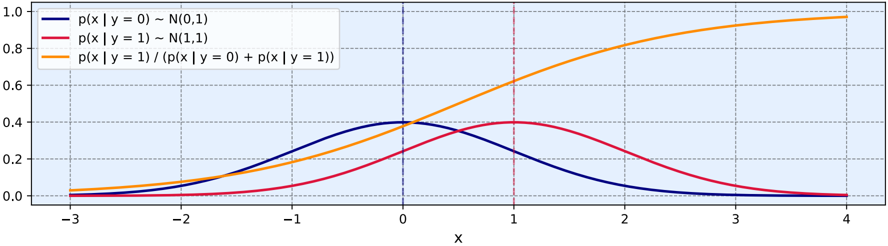
$$g(w) = \sigma\left(w_0 + \langle w, x \rangle\right),$$

где σ - некая непрерывная слева неубывающая функция со значениями в отрезке $[0, 1]$.

- Мы можем использовать **сигмоидную функцию**: $\sigma(z) = \frac{1}{1+e^{-z}}$
- Ее производная: $\sigma'(z) = \sigma(z)(1 - \sigma(z))$.



Откуда берётся сигмоида?



$$p(x \mid y = t) = \frac{1}{(2\pi)^{n/2} \left|\sum\right|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_t)^T \sum^{-1}(x - \mu_t)\right)$$

Нормальное распределение с одинаковыми матрицами ковариации

$$p(y = t \mid x) = \frac{p(x \mid y = t)p(y = t)}{p(x \mid y = 0)p(y = 0) + p(x \mid y = 1)p(y = 1)}$$

Откуда берётся сигмоида

$$p(y=t|x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_t)^T\Sigma^{-1}(x-\mu_t)\right)$$

$$= \sum_i \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_t)^T\Sigma^{-1}(x-\mu_t)\right)$$

$$= \frac{1}{1+\exp\left(+\frac{1}{2}(x-\mu_t)^T\Sigma^{-1}(x-\mu_t) - \frac{1}{2}(x-\mu_{1-i})^T\Sigma^{-1}(x-\mu_{1-i})\right)}$$

$$= \frac{\mathbf{1}}{\mathbf{1}+\exp\left(-\frac{\mathbf{1}}{\mathbf{2}}\mu_t^T\Sigma^{-\mathbf{1}}x - \frac{\mathbf{1}}{\mathbf{2}}x^T\Sigma^{-\mathbf{1}}\mu_t + \frac{\mathbf{1}}{\mathbf{2}}\mu_t^T\Sigma^{-\mathbf{1}}\mu_t + \frac{\mathbf{1}}{\mathbf{2}}\mu_{\mathbf{1}-i}^T\Sigma^{-\mathbf{1}}x + \frac{\mathbf{1}}{\mathbf{2}}x^T\Sigma^{-\mathbf{1}}\mu_{\mathbf{1}-i} - \frac{\mathbf{1}}{\mathbf{2}}\mu_{\mathbf{1}-i}^T\Sigma^{-\mathbf{1}}\mu_{\mathbf{1}-i}\right)}$$

$$= \sigma(w^Tx + w_{\mathbf{0}})$$

Логистическая регрессия

Тогда мы имеем:

$$p(y^i = +1|x) = \frac{1}{1 + e^{-\langle w, x \rangle}}.$$

Подставим трансформированный ответ линейной модели в логарифмическую функцию потерь:

$$\mathcal{L}(w, Xy) = -\sum_{i=1}^{n} \left(\mathbb{I}[y^i = +1] \log \frac{1}{1 + e^{-\langle w, x^i \rangle}} + \mathbb{I}[y^i = -1] \log \frac{e^{-\langle w, x^i \rangle}}{1 + e^{-\langle w, x^i \rangle}} \right)$$

Логистическая регрессия

Тогда мы имеем:

$$p(y^i = +1|x) = \frac{1}{1 + e^{-\langle w, x \rangle}}.$$

Подставим трансформированный ответ линейной модели в логарифмическую функцию потерь:

$$\begin{aligned}
\mathcal{L}(w, Xy) &= -\sum_{i=1}^{n} \left(\mathbb{I}[y^i = +1] \log \frac{1}{1 + e^{-\langle w, x^i \rangle}} + \mathbb{I}[y^i = -1] \log \frac{e^{-\langle w, x^i \rangle}}{1 + e^{-\langle w, x^i \rangle}} \right) \\
&= -\sum_{i=1}^{n} \left(\mathbb{I}[y^i = +1] \log \frac{1}{1 + e^{-\langle w, x^i \rangle}} + \mathbb{I}[y^i = -1] \log \frac{1}{1 + e^{\langle w, x^i \rangle}} \right) \\
&= \sum_{i=1}^{n} \log \left(1 + \exp\left(-y^i \langle w, x^i \rangle \right) \right)
\end{aligned}$$

Градиент и гессиан логистической регрессии

Логистическая регрессия – прекрасная модель, ведь для нее легко можно найти значения градиента и гессиана.

Градиент и гессиан логистической регрессии

Логистическая регрессия – прекрасная модель, ведь для нее легко можно найти значения градиента и гессиана.

- **Градиент:**

$$\nabla f(\mathrm{w}) = \frac{1}{n}\mathrm{X}^\top(\boldsymbol{\sigma} - \mathrm{y}) + \lambda \mathrm{w},$$

где $\boldsymbol{\sigma} = (\sigma((x^1)^\top w), \ldots, \sigma((x^n)^\top w)^\top$.

Градиент и гессиан логистической регрессии

Логистическая регрессия – прекрасная модель, ведь для нее легко можно найти значения градиента и гессиана.

- **Градиент:**

$$\nabla f(\mathrm{w}) = \frac{1}{n}\mathrm{X}^\top(\boldsymbol{\sigma} - \mathrm{y}) + \lambda \mathrm{w},$$

где $\boldsymbol{\sigma} = (\sigma((x^1)^\top w), \dots, \sigma((x^n)^\top w)^\top$.

- **Гессиан:**

$$\nabla^2 f(\mathrm{w}) = \frac{1}{n}\mathrm{X}^\top \mathrm{D}\mathrm{X} + \lambda \mathrm{I},$$

где
$\mathrm{D} = \operatorname{diag}\left(\sigma((x^1)^\top w)(1 - \sigma((x^1)^\top w)), \dots, \sigma((x^1)^\top w)(1 - \sigma((x^1)^\top w))\right)$
— диагональная матрица весов.

От градиентного спуска до метода Ньютона

- Градиентный спуск работает с линейной аппроксимацией в текущей точке, метод Ньютона — с квадратичной:

$$f(x) \approx f(w^k) + \langle \nabla f(w^k), x - w^k \rangle + \frac{1}{2}\langle x - w^k, \nabla^2 f(w^k)(x - w^k) \rangle.$$

От градиентного спуска до метода Ньютона

- Градиентный спуск работает с линейной аппроксимацией в текущей точке, метод Ньютона — с квадратичной:

$$f(x) \approx f(w^k) + \langle \nabla f(w^k), x - w^k \rangle + \frac{1}{2} \langle x - w^k, \nabla^2 f(w^k)(x - w^k) \rangle.$$

Минимизируем квадратичную аппроксимацию по x:

От градиентного спуска до метода Ньютона

- Градиентный спуск работает с линейной аппроксимацией в текущей точке, метод Ньютона — с квадратичной:

$$f(x) \approx f(w^k) + \langle \nabla f(w^k), x - w^k \rangle + \frac{1}{2} \langle x - w^k, \nabla^2 f(w^k)(x - w^k) \rangle.$$

Минимизируем квадратичную аппроксимацию по x:
$\nabla f(w^k) + \nabla^2 f(w^k)(x - w^k) = 0.$

От градиентного спуска до метода Ньютона

- Градиентный спуск работает с линейной аппроксимацией в текущей точке, метод Ньютона — с квадратичной:

$$f(x) \approx f(w^k) + \langle \nabla f(w^k), x - w^k \rangle + \frac{1}{2} \langle x - w^k, \nabla^2 f(w^k)(x - w^k) \rangle.$$

Минимизируем квадратичную аппроксимацию по x: $\nabla f(w^k) + \nabla^2 f(w^k)(x - w^k) = 0$. Откуда получаем следующую точку метода:

$$w^{k+1} = w^k - \left(\nabla^2 f(w^k) \right)^{-1} \nabla f(w^k).$$

От градиентного спуска до метода Ньютона

- Градиентный спуск работает с линейной аппроксимацией в текущей точке, метод Ньютона — с квадратичной:

$$f(x) \approx f(w^k) + \langle \nabla f(w^k), x - w^k \rangle + \frac{1}{2} \langle x - w^k, \nabla^2 f(w^k)(x - w^k) \rangle.$$

  Минимизируем квадратичную аппроксимацию по x: $\nabla f(w^k) + \nabla^2 f(w^k)(x - w^k) = 0$. Откуда получаем следующую точку метода:
$$w^{k+1} = w^k - \left(\nabla^2 f(w^k) \right)^{-1} \nabla f(w^k).$$

- Метод Ньютона использует оракул второго порядка: требует вычисление гессиана.

От градиентного спуска до метода Ньютона

- Градиентный спуск работает с линейной аппроксимацией в текущей точке, метод Ньютона — с квадратичной:

$$f(x) \approx f(w^k) + \langle \nabla f(w^k), x - w^k \rangle + \frac{1}{2}\langle x - w^k, \nabla^2 f(w^k)(x - w^k)\rangle.$$

  Минимизируем квадратичную аппроксимацию по x: $\nabla f(w^k) + \nabla^2 f(w^k)(x - w^k) = 0$. Откуда получаем следующую точку метода:
$$w^{k+1} = w^k - \left(\nabla^2 f(w^k)\right)^{-1} \nabla f(w^k).$$
- Метод Ньютона использует оракул второго порядка: требует вычисление гессиана.
- Стоимость итерации значительно возрастает (по сравнению с градиентным спуском) не только из-за гессиана, но и его обращения.

От градиентного спуска до метода Ньютона

- Градиентный спуск работает с линейной аппроксимацией в текущей точке, метод Ньютона — с квадратичной:

$$f(x) \approx f(w^k) + \langle \nabla f(w^k), x - w^k \rangle + \frac{1}{2}\langle x - w^k, \nabla^2 f(w^k)(x - w^k) \rangle.$$

Минимизируем квадратичную аппроксимацию по x: $\nabla f(w^k) + \nabla^2 f(w^k)(x - w^k) = 0$. Откуда получаем следующую точку метода:

$$w^{k+1} = w^k - \left(\nabla^2 f(w^k)\right)^{-1} \nabla f(w^k).$$

- Метод Ньютона использует оракул второго порядка: требует вычисление гессиана.
- Стоимость итерации значительно возрастает (по сравнению с градиентным спуском) не только из-за гессиана, но и его обращения. **Вопрос:** за сколько итераций метод Ньютона сойдется для квадратичной задачи с положительно определенной матрицей?

От градиентного спуска до метода Ньютона

- Градиентный спуск работает с линейной аппроксимацией в текущей точке, метод Ньютона — с квадратичной:
$$f(x) \approx f(w^k) + \langle \nabla f(w^k), x - w^k \rangle + \frac{1}{2}\langle x - w^k, \nabla^2 f(w^k)(x - w^k)\rangle.$$
Минимизируем квадратичную аппроксимацию по x: $\nabla f(w^k) + \nabla^2 f(w^k)(x - w^k) = 0$. Откуда получаем следующую точку метода: $$w^{k+1} = w^k - \left(\nabla^2 f(w^k)\right)^{-1} \nabla f(w^k).$$
- Метод Ньютона использует оракул второго порядка: требует вычисление гессиана.
- Стоимость итерации значительно возрастает (по сравнению с градиентным спуском) не только из-за гессиана, но и его обращения. **Вопрос:** за сколько итераций метод Ньютона сойдется для квадратичной задачи с положительно определенной матрицей? за 1 (но дорогую).

Метод Ньютона

Алгоритм Метод Ньютона

Вход: стартовая точка $w^0 \in \mathbb{R}^d$, количество итераций K

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: Вычислить $\nabla f(w^k)$, $\nabla^2 f(w^k)$
3: $w^{k+1} = w^k - \left(\nabla^2 f(w^k)\right)^{-1} \nabla f(w^k)$
4: **end for**

Выход: w^K

Метод Ньютона: сходимость

Теорема об оценке сходимости метода Ньютона для μ-сильно выпуклых функций с M-Липшецевым гессианом

Пусть задача безусловной оптимизации с μ-сильно выпуклой целевой функцией f с M-Липшецевыми гессианом решается методом Ньютона. Тогда справедлива следующая оценка сходимости за 1 итерацию

$$\|w^{k+1} - w^*\|_2 \leq \frac{M}{2\mu}\|w^k - w^*\|_2^2.$$

Метод Ньютона: сходимость

Теорема об оценке сходимости метода Ньютона для μ-сильно выпуклых функций с M-Липшецевым гессианом

Пусть задача безусловной оптимизации с μ-сильно выпуклой целевой функцией f с M-Липшецевыми гессианом решается методом Ньютона. Тогда справедлива следующая оценка сходимости за 1 итерацию

$$\|w^{k+1} - w^*\|_2 \leq \frac{M}{2\mu}\|w^k - w^*\|_2^2.$$

Мы уже знаем, что такого рода оценки дают квадратичную скорость сходимости.

Метод Ньютона: сходимость

- Сходимость, как и в случае первор覆одного метода Ньютона, является локальной.

Метод Ньютона: сходимость

- Сходимость, как и в случае первородного метода Ньютона, является локальной. А именно, чтобы гарантировать $\|w^1 - w^*\|_2 < \|w^0 - w^*\|_2$, нужно предположить, что

$$\|w^0 - w^*\|_2 < \frac{2\mu}{M}.$$

Метод Ньютона: сходимость

- Сходимость, как и в случае первородного метода Ньютона, является локальной. А именно, чтобы гарантировать $\|w^1 - w^*\|_2 < \|w^0 - w^*\|_2$, нужно предположить, что
$$\|w^0 - w^*\|_2 < \frac{2\mu}{M}.$$
- Поймем насколько быстро сходится метод. Пусть $M = 2$, $\mu = 1$, а $\|w^0 - w^*\|_2 = \frac{1}{2}$.

Метод Ньютона: сходимость

- Сходимость, как и в случае первородного метода Ньютона, является локальной. А именно, чтобы гарантировать $\|w^1 - w^*\|_2 < \|w^0 - w^*\|_2$, нужно предположить, что
$$\|w^0 - w^*\|_2 < \frac{2\mu}{M}.$$
- Поймем насколько быстро сходится метод. Пусть $M = 2$, $\mu = 1$, а $\|w^0 - w^*\|_2 = \frac{1}{2}$. Тогда мы можем гарантировать, что $\|w^1 - w^*\|_2 \leq \frac{1}{2^2}$,

Метод Ньютона: сходимость

- Сходимость, как и в случае первородного метода Ньютона, является локальной. А именно, чтобы гарантировать $\|w^1 - w^*\|_2 < \|w^0 - w^*\|_2$, нужно предположить, что

$$\|w^0 - w^*\|_2 < \frac{2\mu}{M}.$$

- Поймем насколько быстро сходится метод. Пусть $M = 2$, $\mu = 1$, а $\|w^0 - w^*\|_2 = \frac{1}{2}$. Тогда мы можем гарантировать, что $\|w^1 - w^*\|_2 \leq \frac{1}{2^2}$, $\|w^2 - w^*\|_2 \leq \frac{1}{(2^2)^2}$ и так далее.

Метод Ньютона: модификации

- Пытаемся решить проблему локальной сходимости. Действуем по аналогии с градиентным спуском. **Вопрос:** идеи?

Метод Ньютона: модификации

- Пытаемся решить проблему локальной сходимости. Действуем по аналогии с градиентным спуском. **Вопрос:** идеи?
- Идея первая – шаг:

$$w^{k+1} = w^k - \gamma_k \left(\nabla^2 f(w^k)\right)^{-1} \nabla f(w^k).$$

Такой метод называется демпфированный метод Ньютона.

Метод Ньютона: модификации

- Пытаемся решить проблему локальной сходимости. Действуем по аналогии с градиентным спуском. **Вопрос:** идеи?
- Идея первая – шаг:

$$w^{k+1} = w^k - \gamma_k \left(\nabla^2 f(w^k)\right)^{-1} \nabla f(w^k).$$

Такой метод называется демпфированный метод Ньютона.
Вопрос: как выбирать шаг?

Метод Ньютона: модификации

- Пытаемся решить проблему локальной сходимости. Действуем по аналогии с градиентным спуском. **Вопрос:** идеи?
- Идея первая – шаг:

$$w^{k+1} = w^k - \gamma_k \left(\nabla^2 f(w^k)\right)^{-1} \nabla f(w^k).$$

Такой метод называется демпфированный метод Ньютона.
Вопрос: как выбирать шаг? Много разных способов, например, на прошлой лекции обсуждали линейный поиск:
$\arg\min_\gamma f(w^k + \gamma p_k)$, где $p_k = -\left(\nabla^2 f(w^k)\right)^{-1} \nabla f(w^k)$.

Метод Ньютона: модификации

- Идея вторая – «оценки сверху». В основе анализа градиентного спуска лежала оптимизация «оценки сверху» на функцию:

$$w^{k+1} = \arg\min_{x \in \mathbb{R}^d} \left(f(w^k) + \langle \nabla f(w^k), x - w^k \rangle + \frac{L}{2} \|x - w^k\|_2^2 \right).$$

Метод Ньютона: модификации

- Идея вторая – «оценки сверху». В основе анализа градиентного спуска лежала оптимизация «оценки сверху» на функцию:

$$w^{k+1} = \arg\min_{x \in \mathbb{R}^d} \left(f(w^k) + \langle \nabla f(w^k), x - w^k \rangle + \frac{L}{2} \|x - w^k\|_2^2 \right).$$

Вопрос: чему равно w^{k+1}?

Метод Ньютона: модификации

- Идея вторая – «оценки сверху». В основе анализа градиентного спуска лежала оптимизация «оценки сверху» на функцию:

$$w^{k+1} = \arg\min_{x \in \mathbb{R}^d} \left(f(w^k) + \langle \nabla f(w^k), x - w^k \rangle + \frac{L}{2} \|x - w^k\|_2^2 \right).$$

Вопрос: чему равно w^{k+1}? $w^{k+1} = w^k - \frac{1}{L} \nabla f(w^k)$.

Метод Ньютона: модификации

- Идея вторая – «оценки сверху». В основе анализа градиентного спуска лежала оптимизация «оценки сверху» на функцию:

$$w^{k+1} = \arg\min_{x \in \mathbb{R}^d} \left(f(w^k) + \langle \nabla f(w^k), x - w^k \rangle + \frac{L}{2} \| x - w^k \|_2^2 \right).$$

Вопрос: чему равно w^{k+1}? $w^{k+1} = w^k - \frac{1}{L} \nabla f(w^k)$. Запишем, похожее для аппроксимации 2-го порядка:

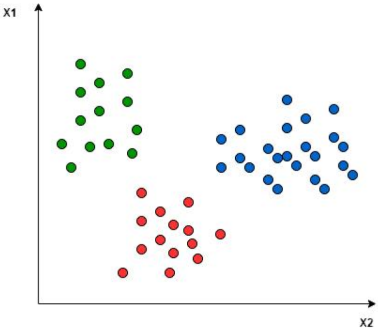
$$\begin{aligned} w^{k+1} = \arg\min_{x \in \mathbb{R}^d} & \left(f(w^k) + \langle \nabla f(w^k), x - w^k \rangle \right. \\ & \left. + \frac{1}{2} \langle x - w^k, \nabla^2 f(w^k)(x - w^k) \rangle + \frac{M}{6} \| x - w^k \|_2^3 \right). \end{aligned}$$

Здесь M – константа Липшица гессиана. Такой метод называется кубический метод Ньютона.

Постановка задачи

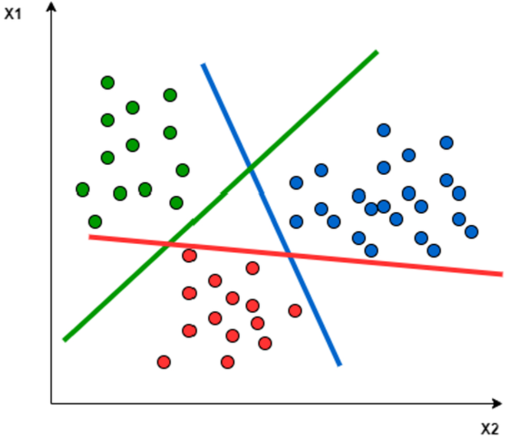
Задача *Многоклассовая классификация*

- Пусть $\mathcal{X} = \mathbb{R}^d$ пространство объектов;
- Пусть $\mathcal{Y} = \{1, \cdot, K\}$ множество допустимых ответов;
- $X = \{(x^i, y^i)\}_{i=1}^n$ - обучающая выборка.



Один против всех (one-versus-all)

- Построим K линейных моделей: $g_k(x) = \langle w_k, x^k \rangle + w_{0,k}$;
- $a_{i,j}$ будем обучать по выборке $\{(x^i, 2\mathbb{I}[y^i = k] - 1)\}_{i=1}^n$;
- Итоговый классификатор: $a(x) = \text{argmax}_{k \in \{1,\dots,K\}}\, g_k(x)$.



Все против всех (all-versus-all)

- Построим C_K^2 линейных моделей: $a_{i,j}(x) = \langle w_{i,j}, x\rangle + w_{0,i,j}$, где $\forall i, j \in \{1, \dots, K\} : \ i \neq j$;
- g_k будем обучать по подвыборке $X_{i,j} = \{(x^m, y^m) \in X \mid \mathbb{I}[y^m = i]$ или $\mathbb{I}[y^m = j]\}$;
- Итоговый классификатор: $a(x) = \mathrm{argmax}_{k\in\{1,\dots,K\}} \sum\limits_{i,j: i\neq j}^{K} \mathbb{I}[a_{i,j} = k]$.

Многоклассовая логистическая регрессия

Бинарная логистическая регрессия:

- Построили линейную модель: $g(x) = \langle w, x \rangle + w_0$;
- Перевели прогноз в вероятность с помощью сигмоидной функции;

Многоклассовая логистическая регрессия:

- Построим K линейных моделей: $g_k(x) = \langle w_k, x \rangle + w_{0,k}$;

Вопрос: Как преобразовывать вектор оценок в многоклассовой логистической регрессии в вектор вероятностей?

SoftMax

Определение

$$\mathrm{SoftMax}\left(z_1, \ldots, z_K\right) = \left(\frac{\exp(z_1)}{\sum_{k=1}^{K} \exp(z_k)}, \ldots, \frac{\exp(z_K)}{\sum_{k=1}^{K} \exp(z_k)}\right)$$

В этом случае вероятность k-го класса будет выражаться как

$$P\left(y = k | x, w\right) = \frac{\exp(\langle w_k, x\rangle + w_{0,k})}{\sum_{j=1}^{K} \exp(\langle w_j, x\rangle + w_{0,j})}.$$

SoftMax

Определение

$$\mathrm{SoftMax}\left(z_1, \ldots, z_K\right) = \left(\frac{\exp(z_1)}{\sum_{k=1}^{K}\exp(z_k)}, \ldots, \frac{\exp(z_K)}{\sum_{k=1}^{K}\exp(z_k)}\right)$$

В этом случае вероятность k-го класса будет выражаться как

$$P\left(y = k | x, w\right) = \frac{\exp(\langle w_k, x\rangle + w_{0,k})}{\sum_{j=1}^{K}\exp(\langle w_j, x\rangle + w_{0,j})}.$$

Обучать эти веса предлагается с помощью метода максимального правдоподобия:

$$\max_{w_1,\ldots,w_K}\sum_{i=1}^{n} P\left(y = y_i | x_i, w\right).$$