Линейные модели. Логистическая регрессия.

Часть 2

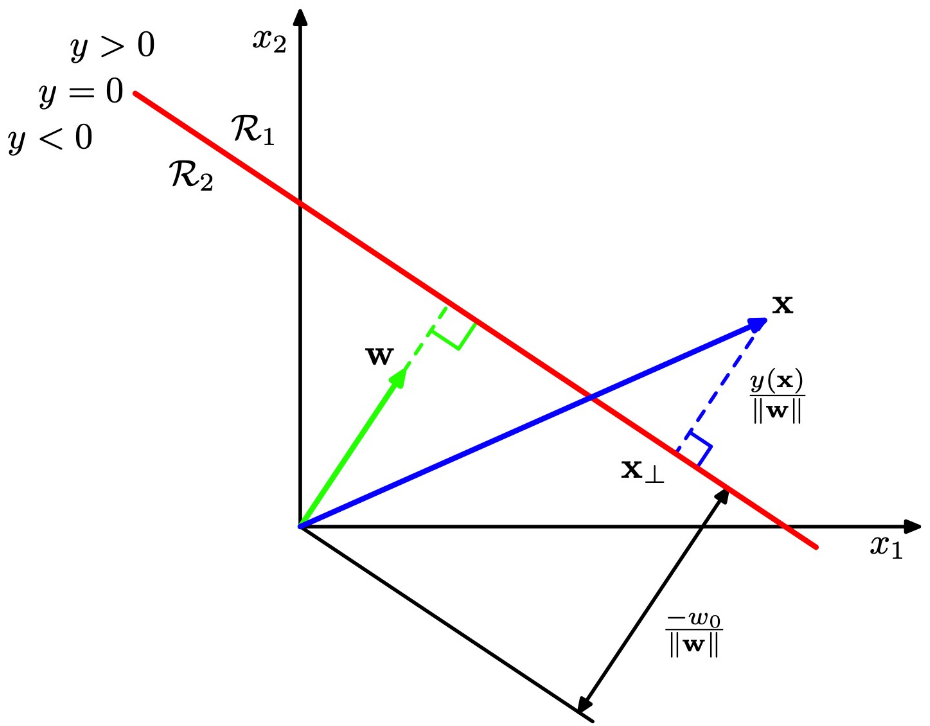
Сбертех, МФТИ

Модели классификации

Линейные модели классификации можно представить как разделяющую гиперплоскость размерностью $(D-1)$ в пространстве D

$$y(x) = \mathrm{w}_0 + \mathrm{w}_1\mathrm{x}_1 + \mathrm{w}_2\mathrm{x}_2 = w^T x$$

- w_0 – сдвиг плоскости относительно начала координат
- $w_1, \dots, w_n$ - направляющий вектор плоскости
- Расстояние от точки до разделяющей гиперплоскости - $\rho = \frac{w^T x}{||w||}$

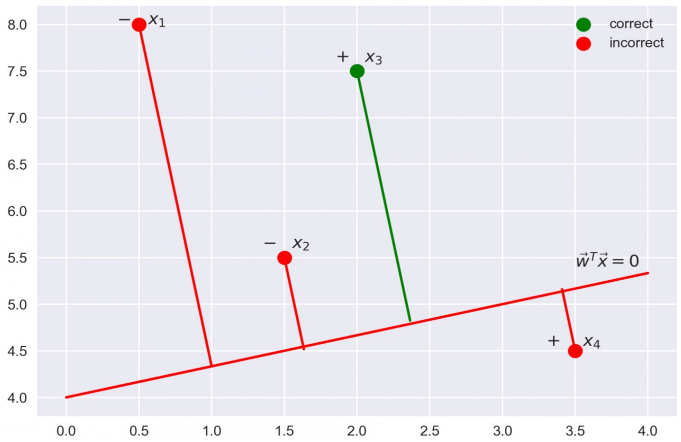


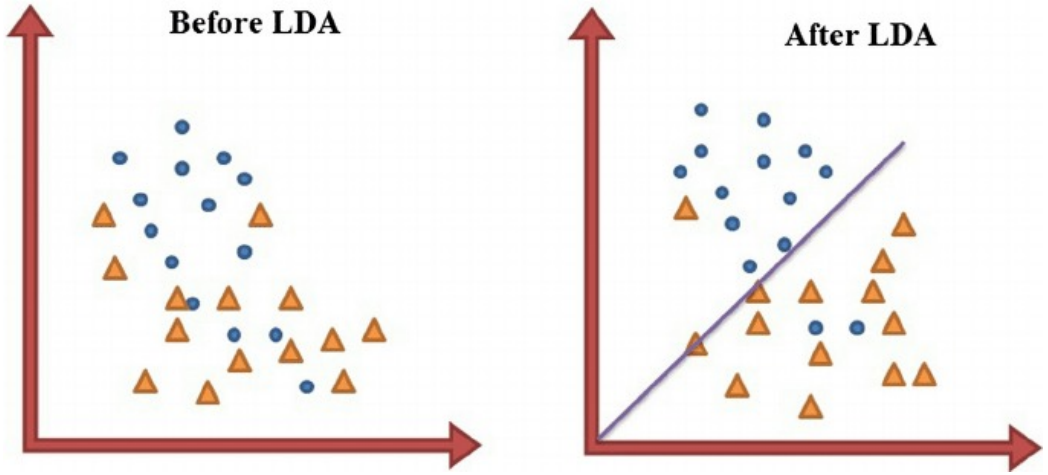
Модели классификации

Есть задача классификации – $y_i \in \{-1; +1\}$. Тогда расстояние между разделяющей прямой и задается уравнением

$$M_i = y_i w^T x$$

- **Если M_i > 0 - значит предсказание верное**
  1) $w^T x$ положительное, y_i = +1
  2) $w^T x$ отрицательное, y_i = -1
- **Если M_i < 0 – предсказание ошибочное**
  1) говорим что $w^T x$ > 0, а на самом деле лейбл меньше
  2) Говорим что $w^T x$ < 0, а на самом деле лейбл больше
- Чем больше M_i– тем более уверены мы в своем решении
- **Задача - максимизировать расстояние от разделяющей гиперплоскости размерности $(D-1)$ до каждой точки из обучающей выборки в пространстве - $\sum M_i \rightarrow max$**





Положительный класс – $\{+1\}$, отрицательный класс – $\{-1\}$

$$f(x, w) = w^T x = \begin{bmatrix} 3 \\ 0.75 \\ 1 \end{bmatrix} [1 \quad x_1 \quad x_2] = 3 + 0.75x_1 + x_2$$

- Если $f(x_i) > 0$, значит предсказываемый объект над разделяющей прямой
  - Если $y_i f(x_i) > 0$ – знак класса такой же как и предсказание, значит решение верное(x_3)
- Если $f(x_i) < 0$, значит предсказываемый объект под разделяющей прямой
  - Если $y_i f(x_i) < 0$ – знак класса не соответствует знаку предсказания, значит решение неверное(x_1, x_2, x_4)
- Чем больше расстояние точки от прямой – тем выше уверенность

Построение функции потерь для оценки вероятности

Пускай есть модель a с некоторым константным предсказанием на всех объектах. Необходимо выбрать такую функцию ошибки, которая была бы наименьшей при значении a равной вероятности отношения количества объектов положительного класса ко всем объектам выборки.

$$\underset{a \in R}{\operatorname{argmin}} \frac{1}{n} \sum L(y_i, a) \approx \frac{1}{n} \sum [y_i = +1]$$

$$\underset{a \in R}{\operatorname{argmin}}\, E(L(y_i, a)|x) = p(y = +1|x)$$

Требование корректного оценивания вероятности моделью

- Если объект x встречается N раз в выборке с классом $+1$ и M раз с классом -1 – $a(x)$ оценивает вероятность положительного класса – $p(y = +1|x) = \frac{N}{N+M}$
- Если много разных объектов в выборке и нет ограничений на прогноз каждого из них – мы выдаем действительное число
- Если много разных объектов в выборке и прогнозы на них ограничены видом модели – $m(x_1, x_2) < eps \Rightarrow a(x_1) \approx a(x_2)$ – нужно чтобы модель оценивала вероятность положительного класса.

Построение функции потерь для оценки вероятности

$$E(L(y_i, a)|x) = p(y = +1|x) * L(+1, a) + p(y = -1|x) * L(-1, a)$$

Рассмотрим на примере MSE(перейдем от $\{-1; +1\} \rightarrow \{1; 0\}$):

$$E((y - a)^2|x) = p(y = 1|x)(1 - a)^2 + p(y = 0|x)(0 - a)^2$$

$$\frac{dE((y - a)^2|x)}{da} = -2p(y = 1|x) * (1 - \mathrm{a}) - 2\big(1 - p(y = 1|x)\big)a$$

$$\frac{dE((y - a)^2|x)}{da} = -(2pa - 2p + 2a - 2pa) = 0$$

$$\mathrm{a} = p(y = 1|x)$$

Требования к модели и функции ошибки

1) Модель выдает вероятность от [0;1]
2) Если мы выбрали константную модель, то минимум функции ошибки достигался бы в точке вероятности отнесения объекта к положительному классу – **можно использовать MSE**.
3) Нам нужно моделировать вероятность с помощью весов модели $w \in R$

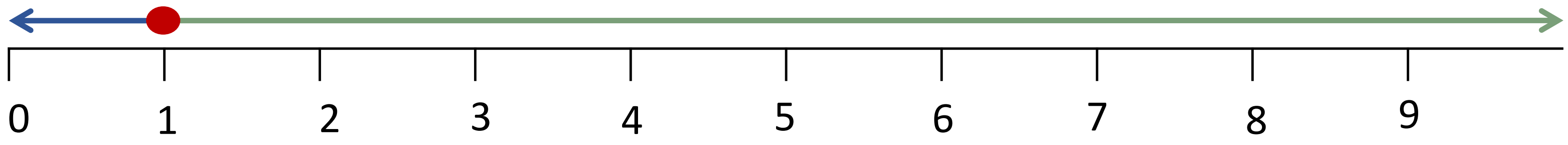
Как нам моделировать $p(y = 1|x)$?

Log Odds

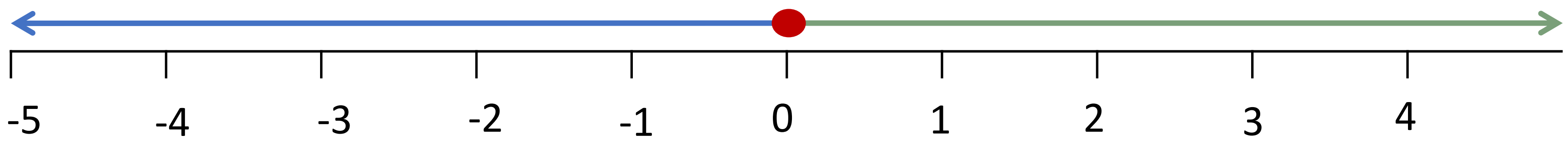
Пускай соотношение победы команды X над командой Y составляет 1:4 в игре x. Обозначим вероятность выигрыша $X - p_X(x) \in [0;1]$, а вероятность выигрыша $Y - p_Y(x) \in [0;1]$. Отношение шансов победы X и Y и вероятность происшествия события X обладает одинаковой информацией. Выразим с помощью вероятности отношение шансов:

$$odds = \frac{\text{вероятность выигрыша } X}{\text{вероятность выигрыша } Y} = \frac{p_X(x)}{p_y(x)} = \frac{0.2}{0.8} = 0.25$$

$$odds\ \mathrm{p_X(X)} = \frac{p_X(X)}{1 - p_X(X)} \in [0; +\infty)$$



$$logodds\ \mathrm{p_X(X)} = \log\left(\frac{p_X(X)}{1 - p_X(X)}\right) \in R$$



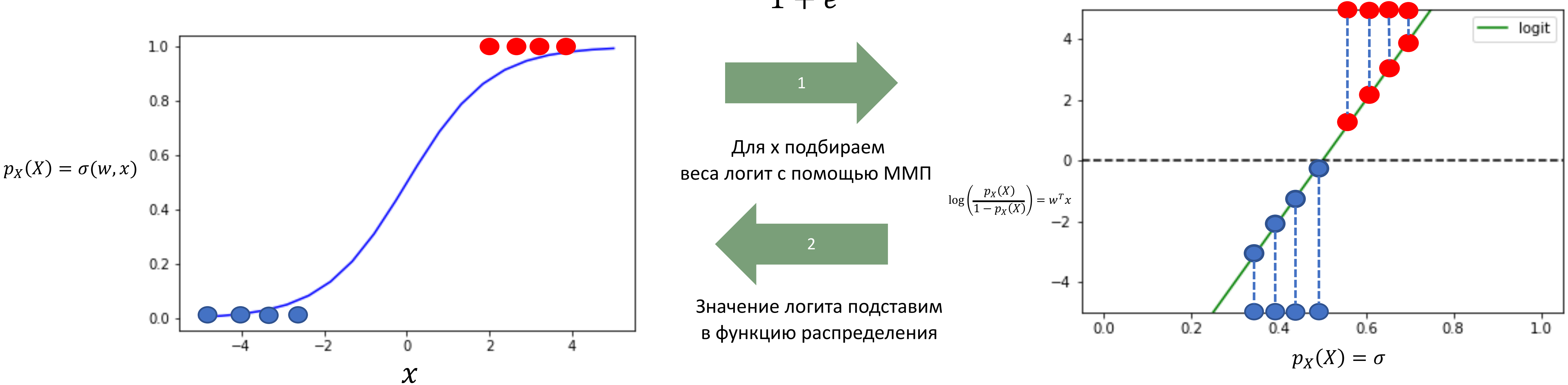
Выбор функции распределения вероятностей

Задача классификации – научить модель определять отношения вероятности положительного класса к отрицательному с помощью функции распределения значений.

$$Log(odds) = \log(\frac{p_X(X)}{p_Y(X)}) = \log(\frac{p_X(X)}{1 - p_X(X)})$$

$$\log\left(\frac{p_X(X)}{1 - p_X(X)}\right) = w^T x$$

$$p_X(X) = \frac{1}{1 + e^{-w^T x}} = \sigma(w, x)$$



Sigmoid

Logit

Построение функции ошибки

$$p_X(x) = \sigma(w^T x)$$

$$p_Y(x) = 1 - p_X(x) = \sigma(-w^T x)$$

1) $p(x) = \sigma(y_i w^T x), y_i \in \{-1; +1\}$

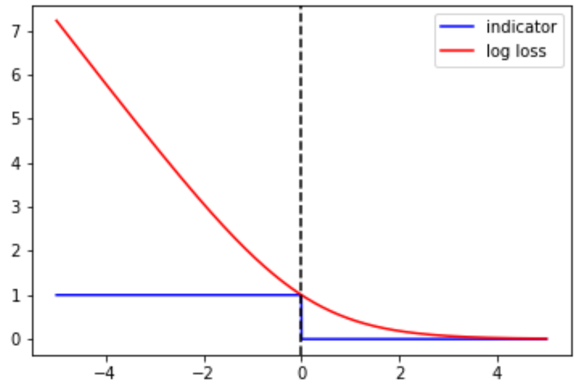
2) $p(x) = \sigma(w^T x)^{y_i} * \left(1 - \sigma(w^T x)\right)^{1-y_i}, y_i \in \{0; 1\}$

1 - функция правдоподобия, значит можем найти оптимальные параметры с помощью ММП

2 – функция правдоподобия Бернулли

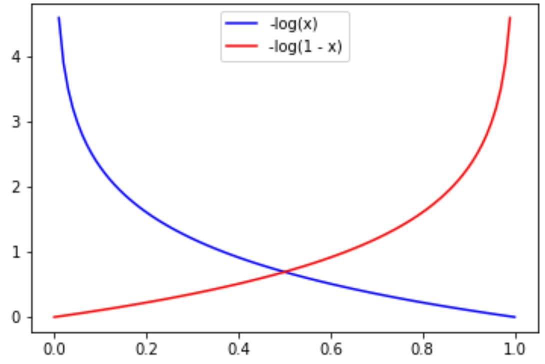
$$\boxed{p(X|w^T, x) = \prod_{i=1}^{n} p(y = X|w^T, x) = \sum_{i=0}^{n} \log(p(y = X|w^T, x))}$$

$$\sum_{i=0}^{n} \log(p(y = X|w^T, x)) = \sum_{i=0}^{n} \log(\frac{1}{1 + e^{-y^i w^T x}}) = -\sum_{i=0}^{n} \log\left(1 + e^{-y^i w^T x}\right)$$



Задача – максимизация отступа от разделяющей прямой на каждого класса

$$\sum_{i=0}^{n} \log(p(y = X|w^T, x)) = -\sum_{i=0}^{n} y_i \log(\sigma(w^T x)) + (1 - y_i) log(1 - (\sigma(w^T x))$$



На положительных объектах предсказать высокую вероятность отнесения к полож. классу

На отрицательных объектах – вероятность отнесения к положительному классу должна быть низкой

Итоговая функция оптимизации называется Log Loss - минимизация логарифма функции правдоподобия - Сигмоиды и Бернулли

Обучение логистической регрессии

$$\sigma_\theta(x) = \frac{1}{1+e^{-\theta x}}$$

$$\frac{d\sigma}{d\theta} = -\frac{1}{(1+e^{-\theta x})^2} e^{-\theta x} * (-x) = x\frac{1}{1+e^{-\theta x}} * \frac{e^{-\theta x}}{1+e^{-\theta x}}$$

$$\frac{e^{-\theta x}}{1+e^{-\theta x}} = 1 - \frac{1}{1+e^{-\theta x}} \Rightarrow \boldsymbol{\frac{d\sigma}{d\theta} = x\sigma(1-\sigma)}$$

$$L_\theta(x,y) = -\sum_{i=0}^{n} y_i \log(\sigma(x)) + (1-y_i)\log(1-\sigma(x)))$$

$$\frac{dL}{d\sigma} = \frac{d\big(-ylog(\sigma)\big)}{d\sigma} + \frac{d(-(1-y)\log(1-\sigma))}{d\sigma}$$

$$\frac{dL}{d\sigma} = \frac{y}{\sigma} + \frac{1-y}{1-\sigma} = \frac{\sigma - y}{\sigma(1-\sigma)}$$

$$\frac{dL}{d\theta} = \frac{dL}{d\sigma} * \frac{d\sigma}{d\theta} = \frac{\sigma - y}{\sigma(1-\sigma)}\big(x\sigma(1-\sigma)\big) = x(\sigma - y)$$

$$\theta_{i+1} = \theta_i - \eta * \frac{1}{n}\sum \nabla L_\theta(x,y)$$

Много-классовая логистическая регрессия

$$\log\left(\frac{p(C_1|x)}{p(C_2|x)}\right) = \log\left(\frac{p(C_1|x)}{1-p(C_1|x)}\right), \text{где } p(C_1|x) = \frac{1}{1+e^{-w^T x}}$$

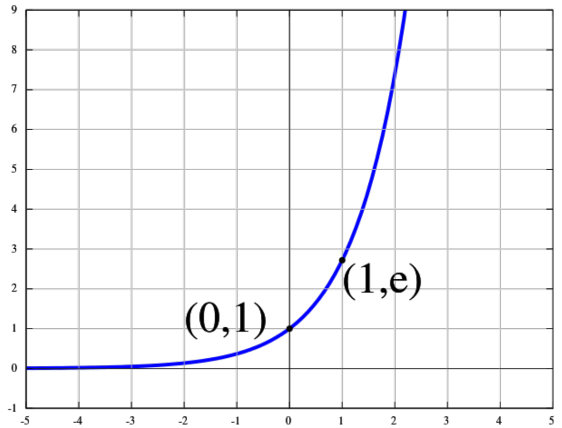
1-class: $p(C_1|x) = \dfrac{p(x|C_1)p(C_1)}{\sum_{k=1}^{2} p(x|C_k)p(C_k)}$ $\quad p(C_2|x) = 1 - p(C_1|x)$ $\quad\Rightarrow\quad$ $logodds\; p(C_1|x) = \log\left(\dfrac{p(C_1|x)}{1 - p(C_1|x)}\right)$

N-class: $p(C_k|x) = \dfrac{p(x|C_k)p(C_k)}{\sum_k p(x|C_k)p(C_k)}$ $\quad\Rightarrow\quad$ $logodds\; p(C_k|x) = \log(p(x|C_k)p(C_k))$

Какую функцию распределения выбрать?

Нам нужно преобразовывать выход модели в некоторое распределение вероятностей по классам. Требование к преобразованию:

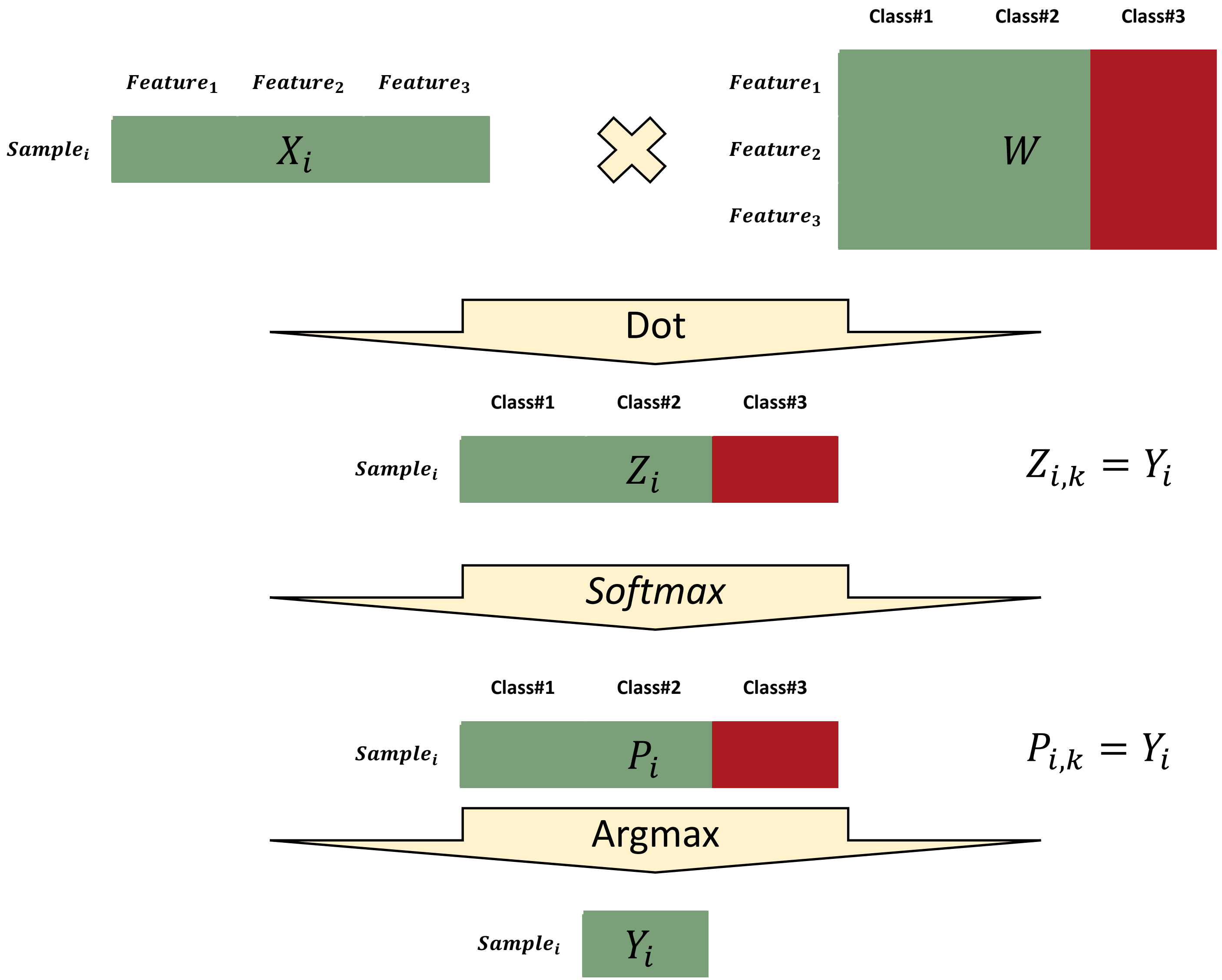
- Функция должна быть дифференцируема
- Высокие значения логита соответствовали высоким значениям вероятностей – наша функция должна быть похожа на argmax



$$\Rightarrow \quad p(C_i|x) = \frac{e^{-w^T x}}{\sum_k e^{-w^T x}}$$

softmax

Много-классовая логистическая регрессия



$$W_k = Y_i$$

$$Y_i \ - onehot$$

$$Z_{i,k} = Y_i$$

$$P_{i,k} = Y_i$$

Функция ошибки и производная

$$p(X|w^T, x) = -\sum_{i=0}^{n} \log(p(y = X|w^T, x))$$

$$L = -\frac{1}{N}\sum_{i=0}^{N} \log\left(\frac{e^{-w^T x}}{\sum_k e^{-w^T x}}\right) = -\frac{1}{N}\sum_{i=0}^{N}(w^T x + \log\left(\sum_k e^{-w^T x}\right))$$

Функция ошибки

$$L_W(X, Y) = \frac{1}{N}(\sum_{i=1}^{N}(X_i W_{k=Y_i} + \log(\sum_{k=0}^{C} e^{-X_i W_k})))$$

$$= \frac{1}{N}(\sum_{i=1}^{N} X_i W Y_{i_{oh}}^T + \sum_{i=1}^{N} log \sum_{k=0}^{C} e^{-X_i W_k})$$

$$\boxed{= \frac{\mathbf{1}}{\boldsymbol{N}}(\boldsymbol{Tr}\left(\boldsymbol{XWY}_{\boldsymbol{i_{oh}}}^{\boldsymbol{T}}\right) + \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{log} \sum_{\boldsymbol{i=1}}^{\boldsymbol{C}} \mathbf{e}^{(\mathbf{-XW})_{\mathbf{ik}}})}$$

Градиент

$$\nabla L_{W_k}(X, Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i^T I_{[Y_i=k]} - \frac{X_i^T e^{-X_i W_k}}{\sum_{k=0}^{C} e^{-X_i W_k}})$$

$$= \frac{1}{N}(\sum_{i=1}^{N} X_i^T I_{[Y_i=k]} - \sum_{i=1}^{N} X_i^T P_i)$$

$$\boxed{= \frac{\mathbf{1}}{\boldsymbol{N}}(\boldsymbol{X}^{\boldsymbol{T}}(\boldsymbol{Y}_{\boldsymbol{oh}} - \boldsymbol{P}))}$$

Эквивалентность МНК и ММП

Пускай задана зависимость

$$y \; = \; f(x,w) + \epsilon = Xw + N(0,\sigma^2 \;) = N(Xw,\sigma^2)$$

Поэтому мы можем задать нашу зависимость как условную вероятность

$$p(y|x,w,\beta) = N(y|Xw,\sigma^2)$$

Объекты независимы друг от друга, значит наша функция правдоподобия

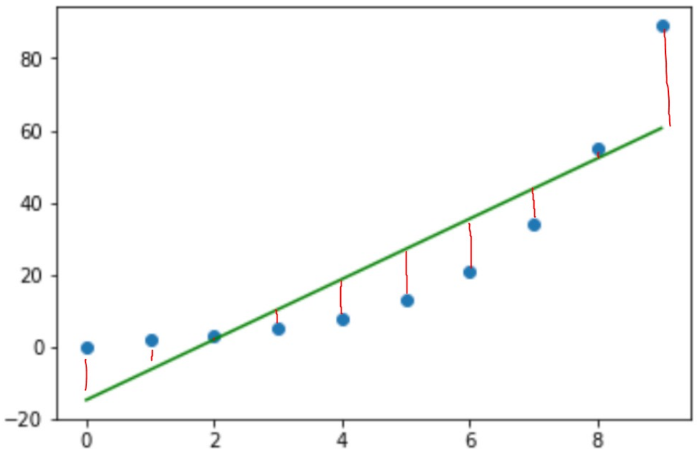
$$p(y|X,w,\sigma^2) = \; \prod_{i=1}^{N} N(y_i|Xw,\sigma^2) = \sum_{i=1}^{N} \log(N(y_i|Xw,\sigma^2)) = \sum_{n=1}^{N} \log(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-Xw)})$$

Максимизация метода максимального правдоподобия эквивалентна методу наименьших квадратов:

$$p(y|X,w,\sigma^2) = -\left(\frac{N}{2} log(2\pi\sigma^2) + \boldsymbol{\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i \; - w^T x_i\right)^2}\right)$$

Следовательно, предпосылки для выборка MSE лежат за нормальным распределением ошибки модели.

Оценка логистической регрессии



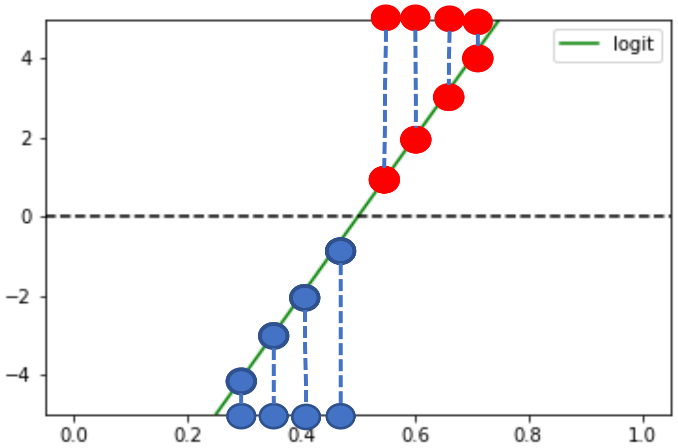
$$SS_{model} = \frac{1}{\mathrm{n}}\sum(y_i - \hat{y}_i)^2$$

$$\bar{x} = \frac{\text{количество положительных объектов}}{\text{общее число объектов}}$$

$$LL_{mean} = \sum_{i=0}^{n} \log(p(y = X|\bar{x}))$$

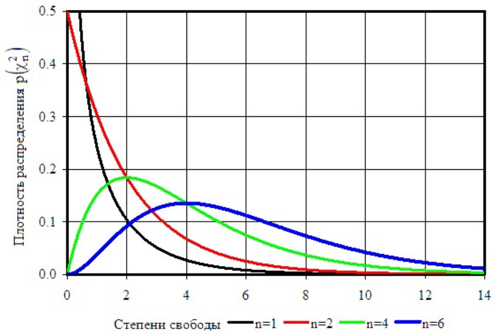
$$LL_{model} = \sum_{i=0}^{n} \log(p(y = X|w^T, x))$$

$$R^2 = \frac{LL_{mean} - LL_{model}}{LL_{model}} \in [0; 1]$$



$$p(y_i|w^T, x) = \sum_{i=0}^{n} \log(p(y = y_i|w^T, x))$$

$$2(LL_{model} - LL_{mean}) \sim \chi^2(\mathrm{df}_{\mathrm{model}} - \mathrm{df}_{\mathrm{mean}})$$

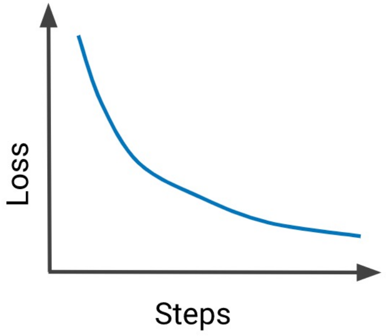


H_0: коэфициенты модели равны 0

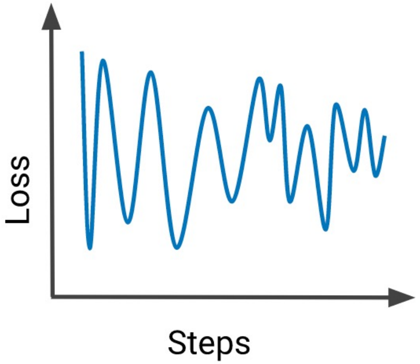
H_1: хотя бы 1 из коэфициентов модели значим

Поведение функций ошибки

Кривая обучения

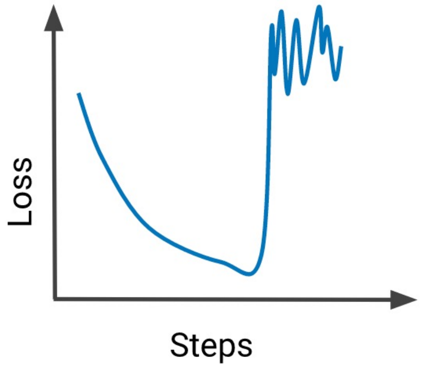


Модель не обучается



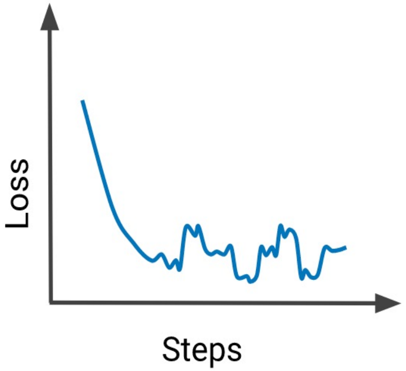
- Проверьте, что модель может правильно предсказывать очевидные примеры из train
- Сделайте LR меньше
- Проверьте работу модели на маленьком сете – добейтесь на нем наименьшего loss – продолжайте работать на большом сете
- Сделайте простую модель и проверьте, что она лучше бейслайна(например, среднее)

Взрыв кривой обучения



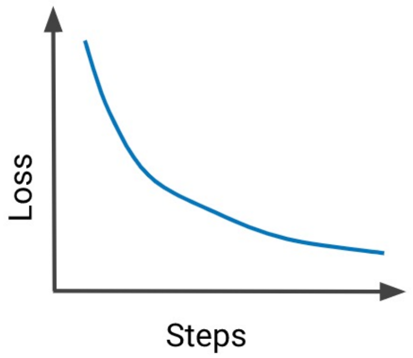
- Данные содержат NaN
- Аномальные объекты в выборке
- Деление на 0
- Логарифм 0 или отрицательного числа

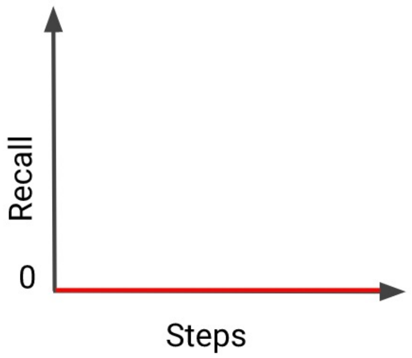
Модель застряла

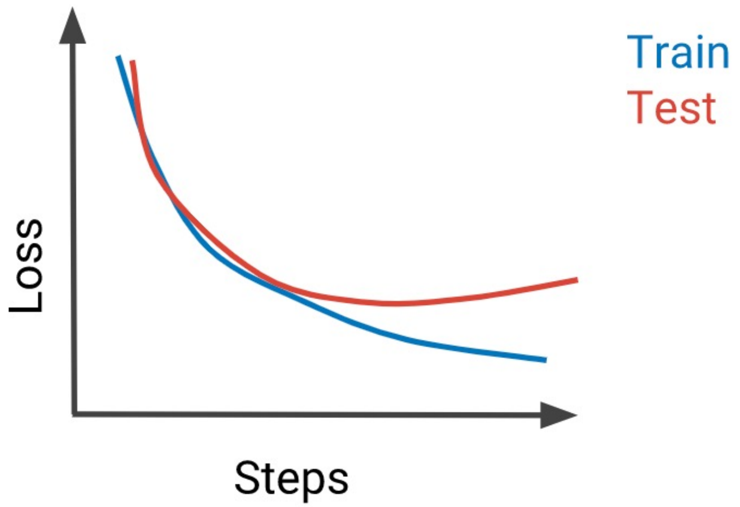


- Данные повторяются, стоит перемешать

Поведение функций ошибки







- Модель не предсказывает положительный класс, так как вероятность предсказания ниже порога(default - 0.5)
    - Часто встречается в сетах с большим дисбалансом классов
- Стоит проверить другие метрики(которые не учитывают порог)

- Эффект переобучения
    - Сделайте модель проще
    - Добавьте регуляризацию
- Проверьте, что train и test совпадают