Линейные модели. Логистическая регрессия.

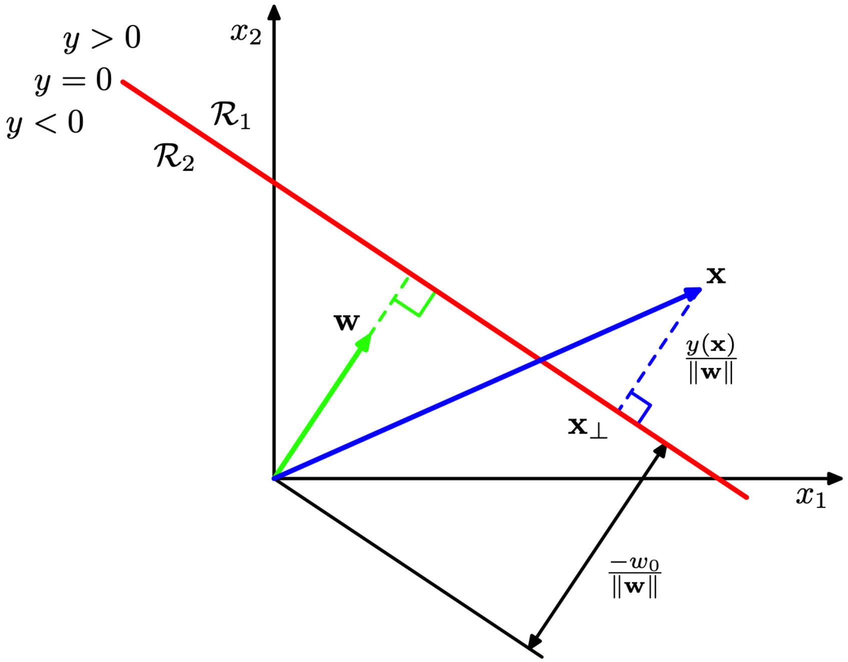
Сбертех, МФТИ

Модели классификации

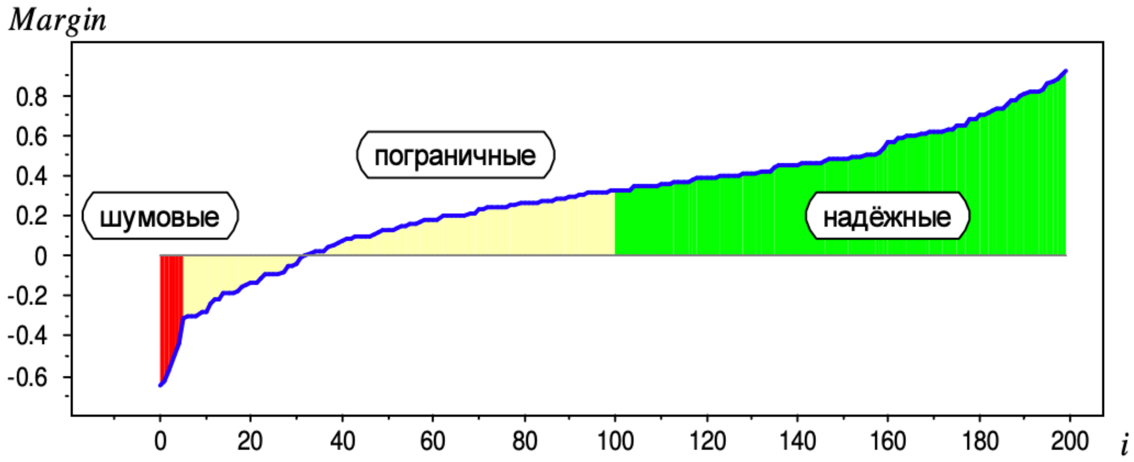
Линейные модели классификации можно представить как разделяющую гиперплоскость размерностью $(D-1)$ в пространстве D

$$y(x) = \theta_0 + \theta_1 \mathrm{x}_1 + \theta_2 \mathrm{x}_2 = \theta^T x$$

- θ_0 – сдвиг плоскости относительно начала координат
- $\theta_1, \dots, \theta_n$ - направляющий вектор плоскости
- Расстояние от точки до разделяющей гиперплоскости(обозначается как margin) - $\rho = \frac{\theta^T x}{||\theta||}$



Здесь $w == \theta$



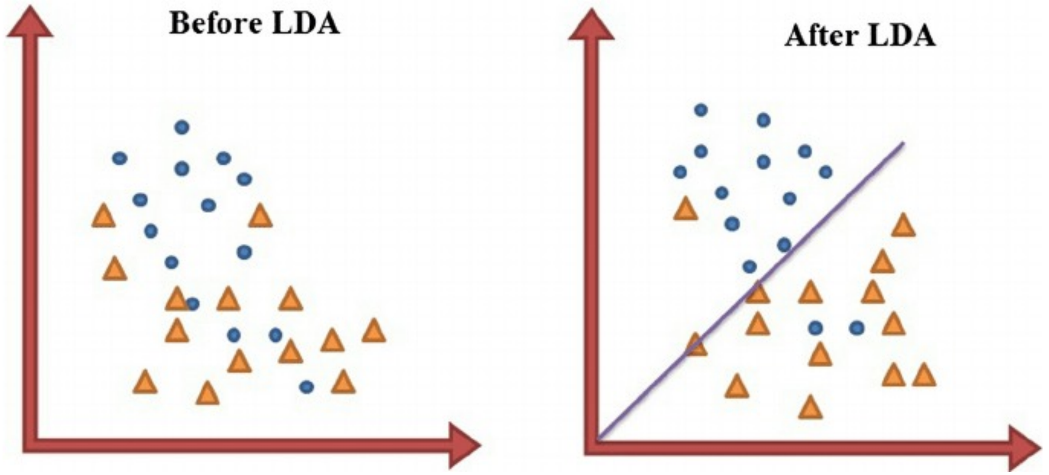
С помощью margin мы можем ранжировать объекты

Модели классификации

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

$$M_i = y_i \theta^T x$$

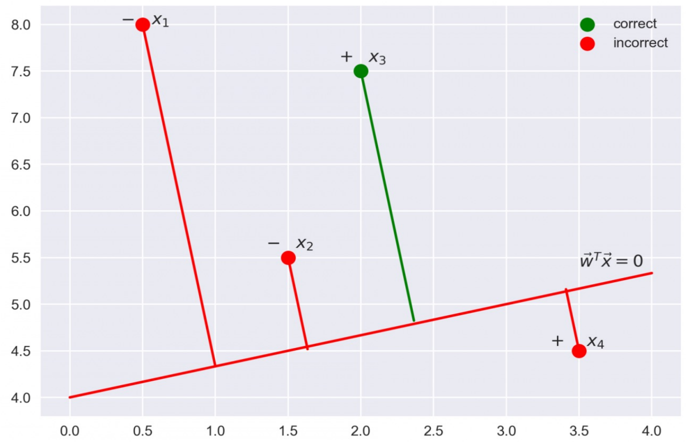
- **Если M_i > 0 - значит предсказание верное**
  1) $\theta^T x$ положительное, y_i = +1
  2) $\theta^T x$ отрицательное, y_i = -1
- **Если M_i < 0 – предсказание ошибочное**
  1) говорим что $\theta^T x$ > 0, а на самом деле лейбл меньше
  2) Говорим что $\theta^T x$ < 0, а на самом деле лейбл больше
- 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 – $\{-1\}$

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

- Если $\mathrm{f}(x_i) > 0$, значит предсказываемый объект над разделяющей прямой
  - Если $\mathrm{y_i} f(x_i) > 0$ – знак класса такой же как и предсказание, значит решение верное(x_3)
- Если $\mathrm{f}(x_i) < 0$, значит предсказываемый объект под разделяющей прямой
  - Если $\mathrm{y_i} f(x_i) < 0$ – 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

1) Модель классификации

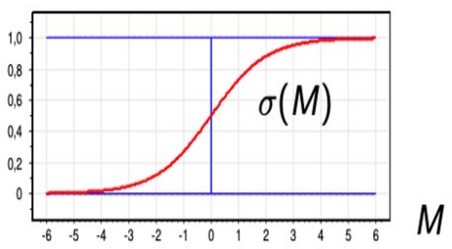
$$a_\theta = sign(\theta^T x)$$

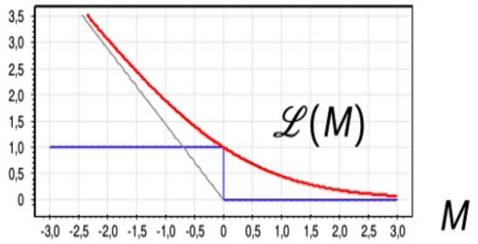
2) Ошибка минимальна при минимальном количестве неверно определённых объектов

$$\sum_{i=1}^{N}[\theta^T x_i y_i < 0] \leq \frac{1}{n}\sum_{i=1}^{N} L_\theta(y_i, x_i)$$

3) Наша задача максимизировать количество правильно предсказанных положительных объектов.

Модель $p(y_i|x_i, \theta)$ ⇄ Модель $a_\theta(x_i)$ и функция ошибки L





Функция ошибки не дифференцируема, попробуем подобрать функцию правдоподобия для поиска оптимальных параметров θ

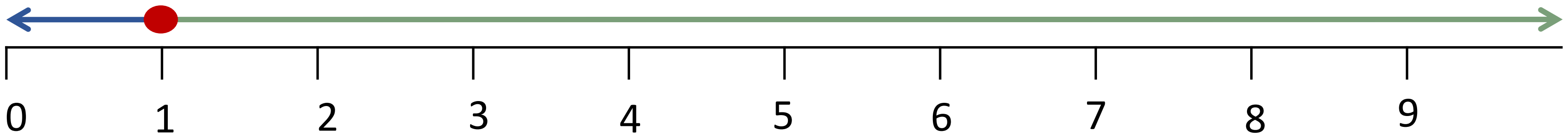
Log Odds

Будем моделировать отношение шансов с помощью $\boldsymbol{\theta^T x}$

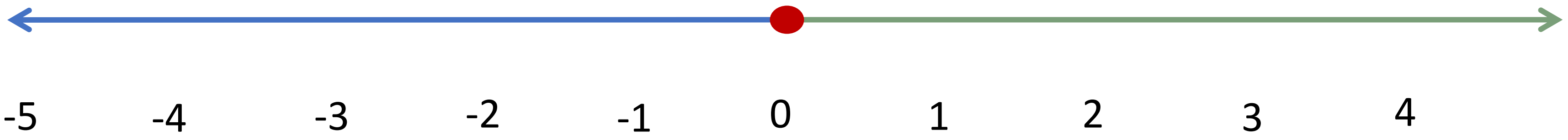
Пускай соотношение победы команды X над командой Y составляет 1:4 в игре x. Обозначим вероятность выигрыша $X - p_X(x) \in [0; 1]$, а вероятность выигрыша $Y - p_Y(x) \in [0; 1]$. Отношение шансов победы X и Y и вероятность происшествия события X обладает одинаковой информацией. Выразим с помощью вероятности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

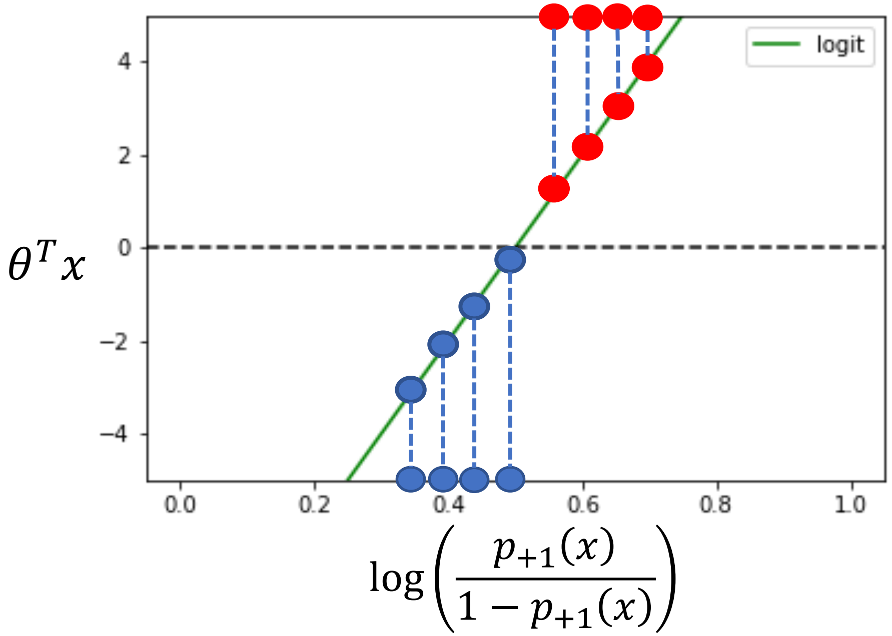
Задача классификации – научить модель определять отношения вероятности положительного класса к отрицательному с помощью функции распределения значений.

1) Запишем логарифм отношения вероятностей выбрать класс +1 - $Log(odds) = \log(\frac{p_{+1}(x)}{p_{-1}(x)}) = \log(\frac{p_{+1}(x)}{1 - p_{+1}(x)})$

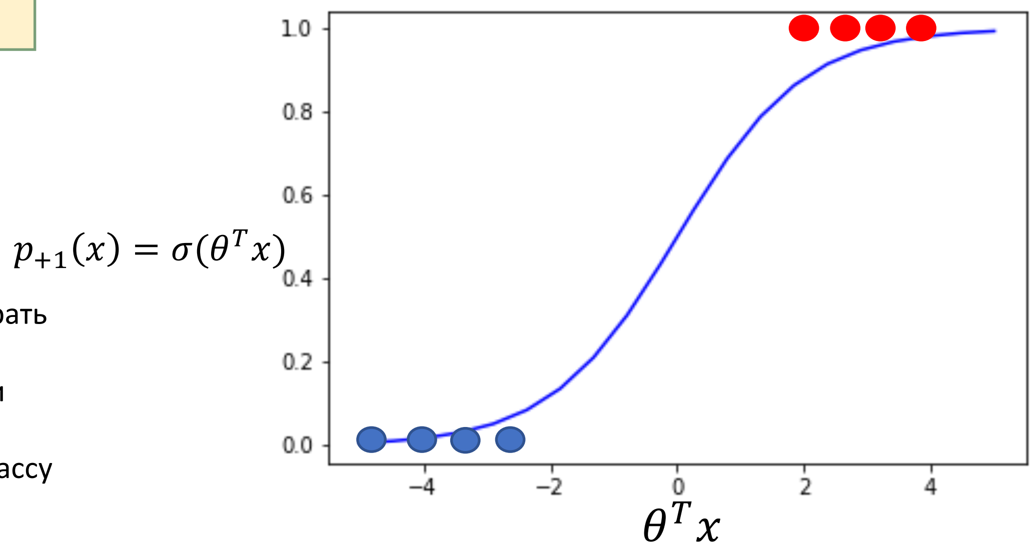
2) Отношение шансов мы хотим моделировать с помощью линейной модели $\log\left(\frac{p_{+1}(x)}{1-p_{+1}(x)}\right) = \theta^T x$

3) Выразим $p_{+1}(X)$ из пункта **2** - $p_{+1}(x) = \boxed{\frac{1}{1+e^{-\theta^T x}}} = \sigma(\theta^T x)$

Сигмоида



Значит мы можем подобрать параметры модели θ, подставить в $\sigma(\theta^T x)$, и получить вероятность отнесения объекта x к классу $+1$

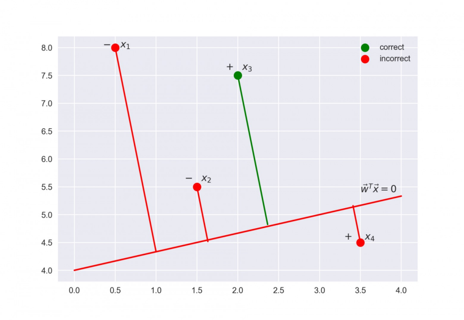


Логистическая функция потерь

$$p_{+1}(x) = P(y = 1|x, \theta^T) = \sigma(\theta^T x)$$

$$p_{-1}(x) = P(y = -1|x, \theta^T) = 1 - \sigma(\theta^T x) = \sigma(-\theta^T x)$$

$$P(y = y_c|x, \theta^T) = \sigma(\boxed{y_c\theta^T x})$$



Чем больше $y_c\theta^T x$, тем больше мы уверены в правильности ответа(margin)

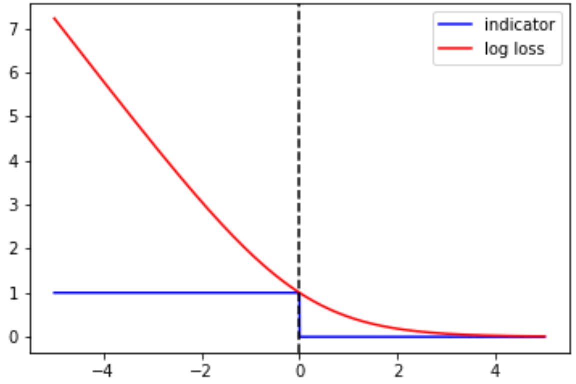
1) Запишем функцию правдоподобия выборки:

$$P(y|x, \theta) = \prod_{i=1}^{n} P(y = y_i|x_i, \theta^T)$$

2) Прологарифмируем выражение:

$$\log P(y|x, \theta^T) = \sum_{i=1}^{n} \log(P(y_i|x_i, \theta^T) = \sum_{i=1}^{n} \log \sigma(y_i\theta^T x_i)$$

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



3) Получим логистическую функцию потерь:

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

Задача логистической функции ошибки – максимизация отступа от разделяющей прямой на каждого класса

Минимизация вероятности отнесения к неверному классу

Пускай мы рассмотрим вместо классов $\{-1; +1\}$, классы $\{0; 1\}$. Используя свойство сигмоиды, получим:

$$p_1(x) = \sigma(\theta^T x)$$

$$p_0(x) = 1 - p_1(x) = \sigma(-\theta^T x)$$

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

Аналогично прошлому рассуждению, рассмотрим функцию правдоподобия:

$$P(y = y_c | x, \theta) = \sigma(\theta^T x)^{y_c} * \left(1 - \sigma(\theta^T x)\right)^{1-y_c}$$

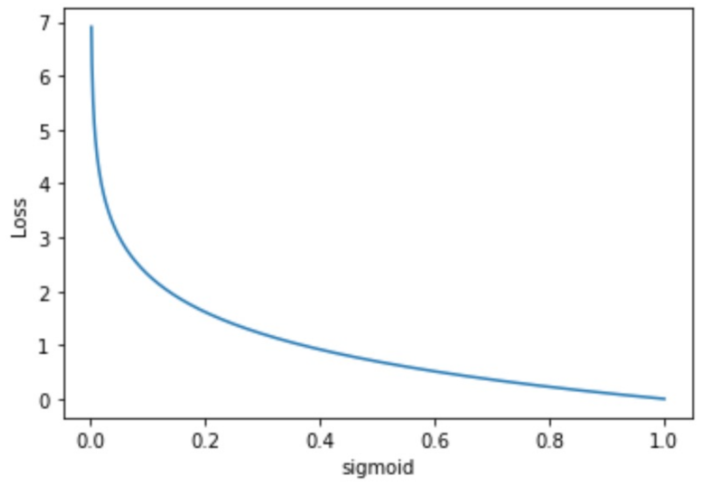
$$\log P(y|x, \theta^T) = \sum_{i=1}^{n} \log(P(y_i|x_i, \theta^T))$$

$$\sum_{i=1}^{n} \log(p(y|x, \theta^T)) = \boxed{-\sum_{i=1}^{n} y_i \, log(\sigma(\theta^T x_i)) + (1 - y_i) log(1 - (\sigma(\theta^T x_i))}$$

$$= -\sum_{i=1}^{n} p(x)_i \, log(\sigma(\theta^T x_i)) + (1 - p(x)) log(1 - (\sigma(\theta^T x_i))$$

$$= -\sum_{i=1}^{n} \sum_{k=1}^{C} y_k \, log(p(x_i)_k)$$

Бинарная кросс-энтропия



Ошибка на положительных объекта маленькая, если большая уверенность модели

Кросс - энтропия

Кросс-энтропия – перекрестная энтропия между распределением целевой переменной и предсказанным распределением

$$L(x,y) = -\sum_{i=1}^{n}\sum_{k=1}^{C} y_k\, log(p(x_i)_k)$$

$$H(x) = -\sum p(x)\log(p(x))$$

Чем больше мы уверены в правильных ответах, тем меньше ошибка

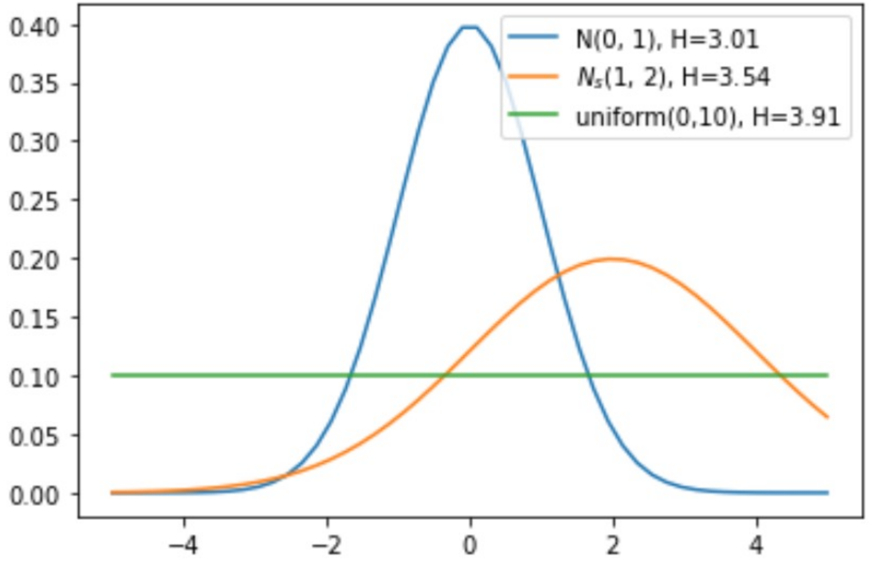
$H(x)$ – количество неопределённости в сообщении.

- *Чем больше неуверенности, тем больше хаос.*

Энтропия некоторого события x это некоторая функция h от вероятности этого события $p(x)$. Эта функция - $\boldsymbol{-log}$

$p(x,y) = p(x) * p(y)$ – события независимы
$h(x,y) = h(x) + h(y)$ – совместная информация от двух независимых событий

- Так как , $p(x) \in [0;1]$ – энтропия не отрицательна
- Если $p(x) = const$, то $H(x) = -1 * log * 1 = 0$ – значение случайной величины абсолютно предсказуемо.
- Если $p(x) = uniform$, то $H(x) = -\frac{1}{K} * \log\left(\frac{1}{K}\right) = K$ – значение случайной величины наибольшее.



Энтропия распределений

Обучение логистической регрессии

Рассмотрим функцию ошибки

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

Воспользуемся правилом дифференцирования сложной функции:

$$\boldsymbol{\frac{dL}{d\theta} = \frac{dL}{d\sigma} * \frac{d\sigma}{d\theta}}$$

Сначала дифференцируем сигмоиду по θ:

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

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

А функцию ошибки по сигмоиде:

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

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

Перемножим производные:

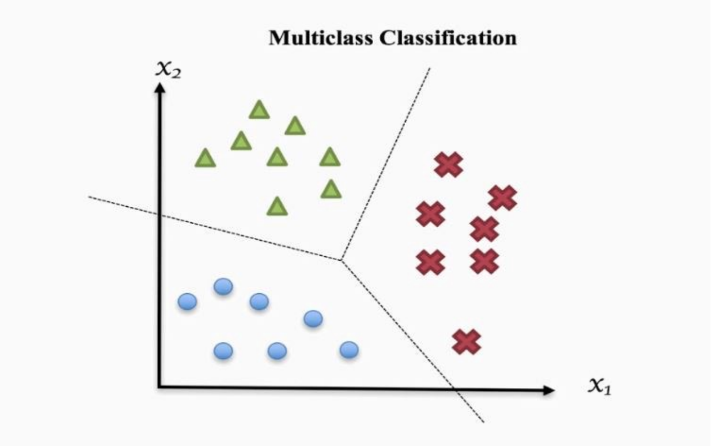
$$\frac{dL}{d\theta} = \frac{dL}{d\sigma} * \frac{d\sigma}{d\theta} = \frac{\sigma - y}{\sigma(1-\sigma)}\left(x\sigma(1-\sigma)\right) = \boldsymbol{x(\sigma - y)}$$

Подбирать параметры логистической регрессии будем через градиентный спуск:

$$\theta_{i+1} = \theta_i - \eta * \frac{dL}{d\theta}$$

One vs All & One vs One классификация

Допустим у нас C классов в задаче. Как мы можем решить задачу много-классовой классификации?



One vs All

1. Строим C датасетов, где представляем задачу, как бинарную.
    - Для каждого класса в выборке
        - $+1$ объектов текущего класса
        - -1 для остальных объектов
2. Обучаем C классификаторов
3. Для каждого объекта выбираем класс за который проголосовали большее число классификаторов.

One vs One

1. Строим $\mathrm{C} * \frac{(C-1)}{2}$ датасетов, где представляем задачу как бинарную.
    1. Для каждого класса в выборке, среди C-1 остальных классов
        - $+1$ для объектов выбранного класса
        - -1 другой класс
2. Обучаем $\mathrm{C} * \frac{(C-1)}{2}$ классификаторов
3. Для каждого объекта выбираем класс за который проголосовали большее число классификаторов.

Много-классовая логистическая регрессия

При двух классах, мы оцениваем вероятность отнесения к положительному объекта к положительному классу. Моделирование происходит с помощью сигмоиды:

$$p(C_1|x) = \frac{1}{1+e^{-x}} = \sigma \qquad x = \log(\frac{\sigma}{1 - \sigma})$$

$p(C_1|x)$ – апостериорная вероятность, можно посчитать по формуле Байеса (**но мы используем отношение шансов для расчета**) :

$$p(C_1|x) = \frac{p(x|C_1)p(C_1)}{\sum_{k=1}^{2} p(x|C_k)p(C_k)}$$

В случае С классов, мы уже получаем распределение на классы, а не одну вероятность:

$$p(C_k|x) = \boxed{\frac{p(x|C_k)p(C_k)}{\sum_k p(x|C_k)p(C_k)}}$$

Общий случай логистической регрессии

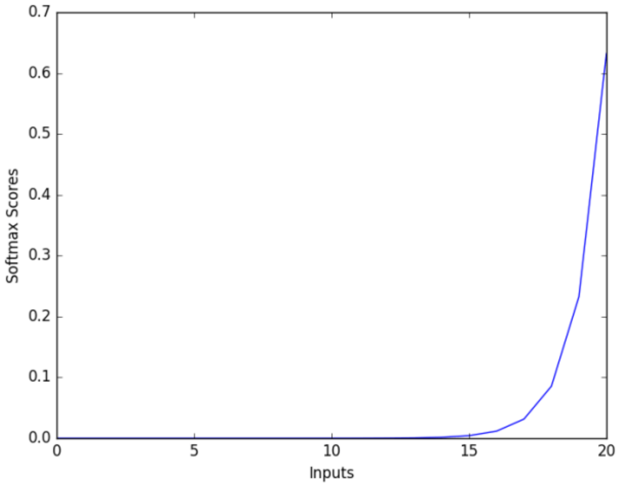
По аналогии с сигмоидой:

- функция должна быть дифференцируема,
- высокие значит логита должны соответствовать высоким значениям вероятности.
- Область значений функции $[0; 1]$

Для этого подходит **softmax** функция:

$$p(C_k|x) = \boxed{\frac{e^{-x}}{\sum_k e^{-x}}}$$

Выводится через one-vs-all логистических регрессий



Много-классовая логистическая регрессия

1. Зададим представление целевой переменной y_i, как one-hot вектор – вектор из 0 на всех индексах, кроме индекса соответствующему корректному классу. Например, y_i = 2, тогда one-hot представление $y_\mathrm{i} = [0, 0, 1]$.
2. Расчет предсказания модели:

$$A = X\theta$$

- X – матрица объекты признаки, размером $n * f$
- θ – матрица параметров, размером $\mathrm{f} * c$

На выходе матрица $n * c$ – **логиты**

- Далее применим операцию $softmax$ каждой строке матрицы A – получим распределение по каждому классу для каждого объекта *(суммируется в 1).*
- С помощью операции $argmax$ по каждой строке получим самый вероятный класс для каждого объекта выборки.
- Обучаем градиентным спуском

Функция ошибки

$$L_\theta(x, y) = -\frac{1}{N}\sum_{i=0}^{N} \log\left(\frac{e^{-x_i\theta y_i^T}}{\sum_{c=0}^{C} e^{-x_i\theta y_c^T}}\right)$$

$$= \frac{1}{N}\sum_{i=0}^{N} x_i\theta y_i^T + \sum_{i=1}^{N} \log\left(\sum_{c=0}^{C} e^{-x_i\theta y_c^T}\right))$$

Градиент функции ошибки

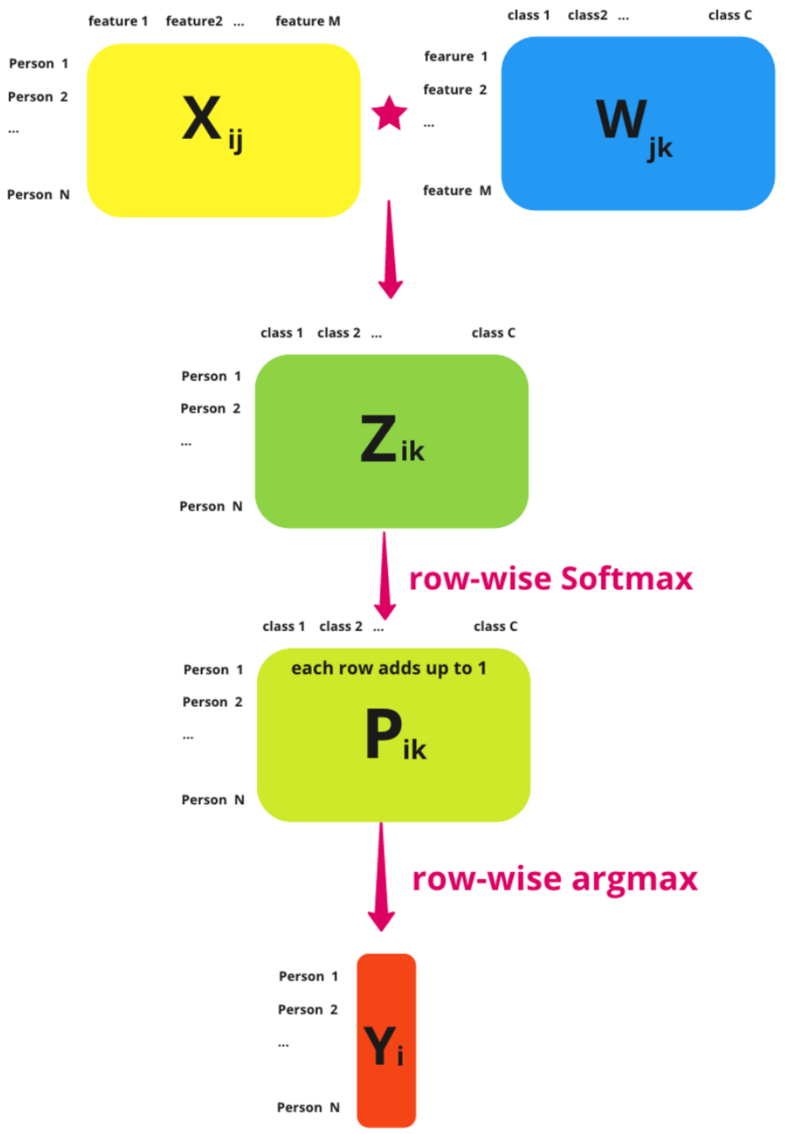
$$\nabla L_\theta(x, y) = \frac{1}{N}(X^T(Y_{oh} \; - P))$$

Y_{oh} - матрица one-hot векторов классов

P – матрица распределения по классам на каждом объекте

https://sophiamyang.github.io/DS/optimization/multiclass-logistic/multiclass-logistic.html

Алгоритм предсказания



Интерпретация логистической регрессии

https://www.statsmodels.org/dev/generated/statsmodels.discrete.discrete_model.Logit.html#statsmodels.discrete.discrete_model.Logit

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                 result   No. Observations:                   20
Model:                          Logit   Df Residuals:                       17
Method:                           MLE   Df Model:                            2
Date:                Mon, 22 Aug 2022   Pseudo R-squ.:                  0.1894
Time:                        09:53:35   Log-Likelihood:                -11.156
converged:                       True   LL-Null:                       -13.763
Covariance Type:            nonrobust   LLR p-value:                   0.07375
==============================================================================
                coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -2.1569      1.416     -1.523      0.128      -4.932       0.618
method[T.B]   0.0875      1.051      0.083      0.934      -1.973       2.148
hours         0.4909      0.245      2.002      0.045       0.010       0.972
==============================================================================
```

Мы обучаем модель с помощью ММП, поэтому нам важно знать достигли ли мы сходимости ошибки.

Псевдо R^2

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

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

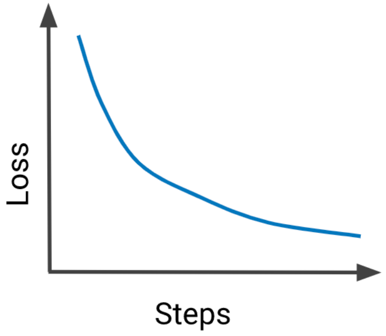
$$R^2 = \frac{LL_{mean} - LL_{model}}{LL_{model}} \in [0; 1]$$

$$z = \frac{coef}{\sigma}$$

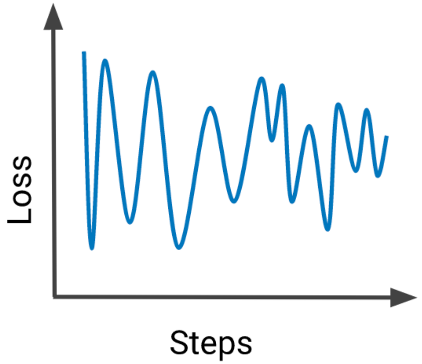
Оценка неопределённости коэффициента.
Чем больше z, тем меньше неопределённости – с ростом коэффициента падает σ.

Поведение функций ошибки

Кривая обучения

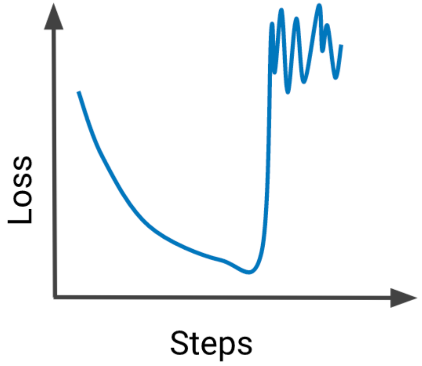


Модель не обучается



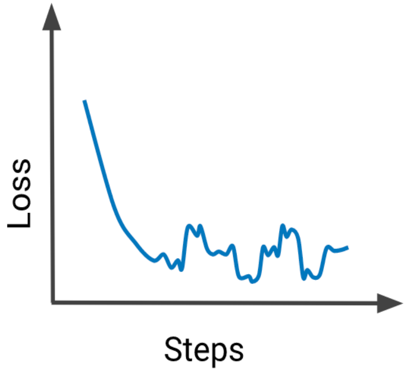
- Проверьте, что модель может правильно предсказывать очевидные примеры из train
- Сделайте LR меньше
- Проверьте работу модели на маленьком сете – добейтесь на нем наименьшего loss – продолжайте работать на большом сете
- Сделайте простую модель и проверьте, что она лучше бейслайна(например, среднее)

Взрыв кривой обучения



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

Модель застряла



- Данные повторяются, стоит перемешать

Применимость в индустрии

Логистическая регрессия удобна тем, что дает возможность оценить вероятность события и оценить возможные потери.

$$R(x,s) = \sum_{i=1}^{N} L_{xs} * p_{-1}(x)$$

- L_{xs} - потери размера s на объекте x
- $p_{-1}(x)$ – вероятность отрицательного события на объекте x

Value at Risk - это величина убытков, которая с вероятностью, равной уровню доверия (например, 99 %), не будет превышена. Следовательно, в 1 % случаев убыток составит величину, большую чем VaR.

- Можем провести эксперимент на клиентах банка и определить размер резервируемого капитала в случае невыплаты кредитов.
- Можем оценить логистические затраты на возврат товаров на маркетплейсах.
- Оценить риски при покупке строящейся недвижимости.