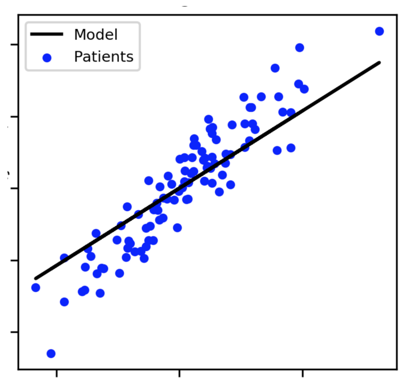
# Семинар 6: Логистическая регрессия

## И не задача вовсе, а подводка

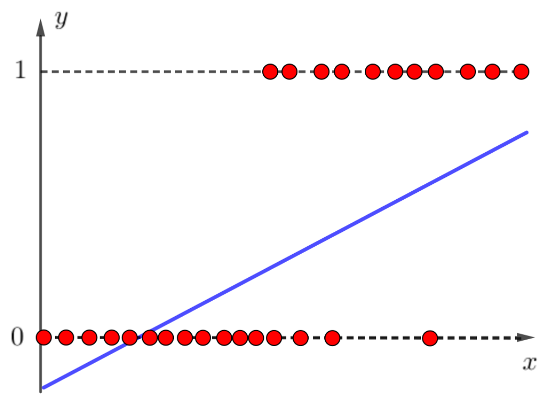
Давайте построим логистическую регрессию также как мы сделали это на семинаре. До этого мы с вами уже имели дело с обычной регрессией:

$$y = \beta_0 + \beta_1 \cdot x.$$

Когда мы оценивали её, мы рисовали через облако точек линию:



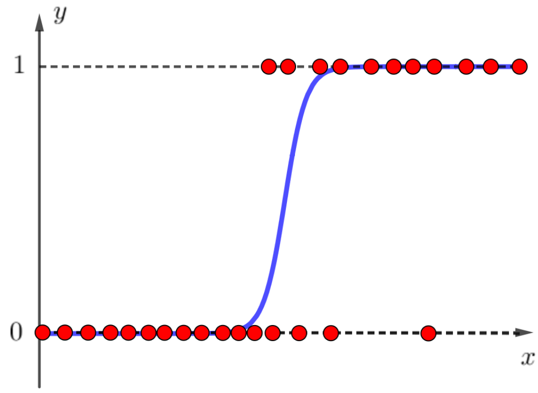

Переменная, которую мы прогнозировали, $y$, принимала любые значения и всё было хорошо. Теперь мы решаем задачу классификации. Наша переменная принимает значения либо 0 либо 1. Если мы опять будем строить обычную регрессию, мы попадём в глупую ситуацию:




Наша голубая линия регрессии снова пройдёт через облако точек. Когда мы будем пытаться на её основе построить прогноз, мы будем получать абсолютно любые значения. Это могут быть и $-7$, и 2.1, и 1, и даже 0.33.

В принципе, мы можем интерпретировать эти числа как уверенность нашей модели. Например, если получилось 55, значит модель уверена в том, что класс первый. А если получилось $-33$, модель уверена, что класс нулевой. Ну а если 0.5, то модель колеблется.

Правда эту степень уверенности хорошо было бы пронормировать. Обычно это делают на отрезок от нуля до единицы. Для этого вместо линии рисуют вот такую $S$-образную кривую:




Тогда значения принимаются на отрезке от $0$ до $1$ и мы можем их интерпретировать как вероятность первого класса, $P(y=1)$. Какую функцию можно взять для такой кривой? Из теории вероятностей вы знаете, что все функции распределения ведут себя $S$-образно. В машинном обучении обычно берут логичтическую функцию распределения, потому что она очень простая:

$$P(y=1) = \frac{1}{1+e^{-z}}.$$

Такую функцию иногда называют сигмоидой. Так и получается логичтическая регрессия. На первом шаге мы считаем "уверенность"модели:

$$z = \beta_0 + \beta_1 \cdot z,$$

а на втором шаге превращаем её в вероятность с помощью сигмоиды. Из-за того, что мы используем логистическое распределение, такую регрессию называют логистической.

Модель построена. Дело осталось за малым. Выбрать функцию потерь. В случае регрессии мы использовали с вами $MSE$. В случае логистической регрессии мы также можем попробовать использовать его же, но нам бы хотелось придумать что-то новое. Новая функция потерь должна подходить для нашей задачи по смыслу.

Наши $y$ могут принимать значения $1$ и $0$. Если $y=1$, мы хотим, чтобы модель спрогнозировала $\hat{p} = P(y=1)$ побольше. Если $y=0$, мы хотим, чтобы модель спрогнозировала $\hat{p}$ поменьше, то есть $1-\hat{p}$ побольше.

Тогда мы можем выписать такую штуку:

$$-1 \cdot (y \cdot \hat{p} + (1-y)(1-\hat{p})).$$

Нам надо найти её минимум по $\beta$. Тогда модель будет работать хорошо. Если $y=1$, мы будем получать большое $\hat{p}$, так ка второе слагаемое в нашей формуле будет зануляться. Если $y = 0$, то будет зануляться первое слагаемое, и мы будем пытаться получить большое $1 - \hat{p}$.

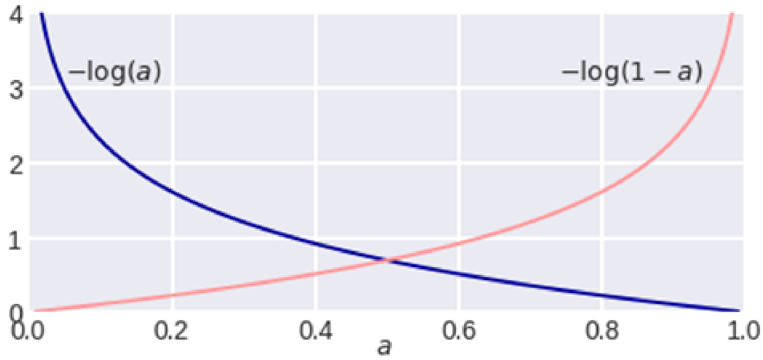
Функция потерь почти готова. Остался последний штрих. Давайте заставим нашу функцию штрафовать нас при сильных ошибках сильнее, как это было в случае MSE для регрессии. Для этого нужно взять от $\hat{p}$ логарифм и получить:

$$-1 \cdot (y \cdot \ln \hat{p} + (1 - y) \ln(1 - \hat{p})).$$

Это будет наша итоговая функция потерь. Она называется logloss и обычно используется для обучения логистической регрессии.

Остаётся только одни вопрос: почему логарифм вносит более большой штраф для более сильных ошибок. Давайте нарисуем $\ln \hat{p}$ и $\ln(1 - \hat{p})$ на картинке (картинка из интернета, из блога Дьяконова[1] , поэтому на ней $a$ это $p$.)

$$-\begin{cases} \log a_i, & y_i = 1, \\ \log(1 - a_i), & y_i = 0. \end{cases}$$




Когда $y = 1$ мы пытаемся сделать поменьше $-\ln \hat{p}$. Если он оказался очень большим, надо делать его меньше существенно сильнее, если он уже итак маленький. Посмотрите на синюю кривую. Это наш логарифм. Поначалу он убывает очень резко, а после медленнее. Это позволяет штрафовать за сильные ошибки сильнее.

И практически никакой математики. Одна сплошная интуиция.

## Задача 1

Винни Пух (ВП) — исследователь и пасечник. Пчёлы ВП откладывают мёд. ВП пробует его и понимает, правильной оказалась пчела или нет. Спустя многие годы работы ВП накопил довольно большую выборку из правильных и неправильных пчёл и смог на основе неё оценить модель:

[1] https://dyakonov.org/2018/03/12

$$z = 1 + 0.5 \cdot x,$$

где $x$ — это количество мёда, которое снесла пчела. Предположим, что ВП сталкивается с новой пчелой. Он знает, что она снесла $x = 6$ килограмм мёда. Какова вероятность того, что эта пчела правильная? Предположим, что эта пчела оказалась неправильной. Какой logloss совершает ВП? Какой logloss будет, если эта пчела оказалась правильной?

**Решение**:

Найдём "уверенность"ВП в правильности пчелы:

$$z = 1 + 0.5 \cdot 6 = 4.$$

Превратим её в вероятность с помощью сигмоиды:

$$\hat{P}(y = 1) = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-4}} = 0.98.$$

Получаем, что вероятность того, что пчела правильная $0.98$. Пусть эта пчела в реальности оказалась правильной. Найдём для неё logloss:

$$-1 \cdot (1 \cdot \ln 0.98 + (1 - 1) \cdot \ln(1 - 0.98)) = 0.021.$$

За ошибку отвечает первое слагаемое. Вероятность того, что пчела была правильной оказалась довольно высокой. Пчела и правда оказалась правильной. Однако вероятность была чуть меньше единицы. Мы не были абсолютно уверены, а значит немного ошибались. logloss формализует это "немного"в виде числа.

Теперь найдём logloss в случае, если пчела в реальности оказалась неправильной. Тут за ошибку отвечает второе слагаемое. Мы, сказав что вероятность её правильности $0.98$ очень сильно ошиблись. Посмотрим каким станет logloss:

$$-1 \cdot (0 \cdot \ln 0.98 + (1 - 0) \cdot \ln(1 - 0.98)) = 4.1.$$

Ошибка довольно сильно выросла.

## Задача 2

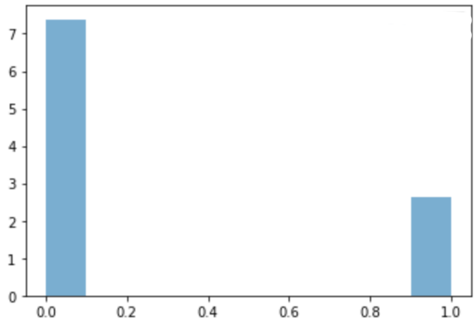
У ВП в тестовой выборке есть две пчелы. Одна правильная, одна неправильная. Он хочет проверить на этой выборке свою модель. Она предсказала, что первая правильная с вероятностью $0.6$, а вторая с вероятностью $0.4$. Найдите средний logloss на этой выборке.

**Решение**:

$$\text{logloss} = -\frac{1}{2} \cdot ((1 \cdot \ln(0.6) + (1-1) \cdot \ln(1-0.6)) + (0 \cdot \ln(0.4) + (1-0) \cdot \ln(1-0.4))) = -0.5 \cdot (\ln 0.6 + \ln 0.6) = -\ln 0.6 \approx 0.51$$

## Задача 3

ВП построил для своей выборки картинку, чтобы посмотреть насколько его выборка сбалансированна. Получилось вот так:




Обычно ВП минимизировал такой logloss:

$$-\frac{1}{n} \sum_{i=1}^{n} (y_i \cdot \ln \hat{p}_i + (1-y_i) \cdot \ln(1-\hat{p}_i))$$

В этот раз ВП решил минимизировать немного модернизированную функцию:

$$-\frac{1}{n} \sum_{i=1}^{n} ({\color{red}1\cdot} y_i \cdot \ln \hat{p}_i + {\color{red}3\cdot}(1-y_i) \cdot \ln(1-\hat{p}_i))$$

Как думаете, зачем ВП сделал это?

**Решение**: Думаю, что вы справитесь.

## Задача 4 (для настоящих дата саунистов)

**Задачка со звёздочкой. Для тех, кто интересуется. Её не будет в самостоялке! Несмотря на это она довольно простая**. На семинаре мы сказали, что логистическую регрессию тоже можно учить с помощью градиентного спуска. Давайте попробуем сделать один шаг этой процедуры.

У ВП есть логистическая регрессия и функция потерь:

$$z = \beta \cdot x$$
$$P(y=1) = \frac{1}{1+e^{-z}}$$
$$logloss = -1(y \cdot \ln \hat{p} + (1-y) \cdot \ln(1-p))$$

Оказалось, что $x = -5$, а $y = 1$. Сделайте один шаг градиентного спуска, если $\beta_0 = 1$, а скорость обучения $\gamma = 0.01$.

**Решение**:

Сначала нам надо найти $logloss'_\beta$. В принципе в этом и заключается вся сложность задачки. Давайте подстави вместо $\hat{p}$ в logloss сигмоиду.

$$logloss = -1\left(y \cdot \ln\left(\frac{1}{1+e^{-z}}\right) + (1-y) \cdot \ln\left(1 - \frac{1}{1+e^{-z}}\right)\right)$$

Теперь подставим вместо $z$ уравнение регрессии:

$$logloss = -1\left(y \cdot \ln\left(\frac{1}{1+e^{-\beta \cdot x}}\right) + (1-y) \cdot \ln\left(1 - \frac{1}{1+e^{-\beta \cdot x}}\right)\right)$$

Если говорить без прекрас, это и есть наша функция потерь. От неё нам нужно найти производную. Давайте подготовимся. Делай раз, найдём производную $logloss$ по$\hat{p}$:

$$logloss'_{\hat{p}} = -1\left(y \cdot \frac{1}{\hat{p}} - (1-y) \cdot \frac{1}{(1-p)}\right)$$

Делай два, найдём производную $\frac{1}{1+e^{-z}}$ по $z$:

$$\left(\frac{1}{1+e^{-z}}\right)'_z = -\frac{1}{(1+e^{-z})^2} \cdot e^{-z} \cdot (-1) = \frac{1}{1+e^{-z}} \cdot \frac{e^{-z}}{1+e^{-z}} = \frac{1}{1+e^{-z}} \cdot \left(1 - \frac{1}{1+e^{-z}}\right)$$

По-другому это можно записать как $\hat{p} \cdot (1-\hat{p})$. **Всё.** Давайте искать полную производную. Начнём изнутри:

$$\left(e^{-\beta \cdot x}\right)'_\beta = e^{-\beta \cdot x} \cdot (-\beta)$$

Расширяемся:

$$\left(\frac{1}{1+e^{-\beta \cdot x}}\right)'_\beta = \frac{1}{1+e^{-\beta \cdot x}} \cdot \left(1 - \frac{1}{1+e^{-\beta \cdot x}}\right) \cdot e^{-\beta \cdot x} \cdot (-\beta)$$

Расширяемся до конца:

$$\text{logloss}'_\beta = -1\left(y \cdot \frac{1}{\hat{p}} \cdot \hat{p} \cdot \big(1-\hat{p})\big) \cdot e^{-\beta \cdot x} \cdot (-\beta) - (1-y) \cdot \frac{1}{(1-\hat{p})} \cdot \hat{p} \cdot \big(1-\hat{p})\big) \cdot e^{-\beta \cdot x} \cdot (-\beta)\right) =$$

$$= y \cdot \big(1-\hat{p}\big) \cdot e^{-\beta \cdot x} \cdot \beta + (1-y) \cdot \hat{p} \cdot e^{-\beta \cdot x} \cdot \beta$$

Найдём значение производной в точке $\beta_0 = 1$ для нашего наблюдения $x = -5, y = 1$:

$$1 \cdot \left(1 - \frac{1}{1+e^{-1 \cdot -5}}\right) \cdot e^{-1 \cdot -5} \cdot 1 \approx 147.41$$

Делаем шаг градиентного спуска:

$$\beta_1 = 1 - 0.01 \cdot 147.41 \approx -0.47$$