



Логистическая регрессия



Логистическая регрессия

- Мы посмотрели линейную регрессию и её различные варианты для предсказания значений непрерывной переменной.
- А как нам предсказывать категориальные переменные?



Логистическая регрессия

- Логистическая регрессия
 - Пусть слово регрессия в названии Вас не смущает!
 - Это алгоритм классификации для предсказания значений категориальных переменных.



Логистическая регрессия

- Обзор раздела
 - Переход от линейной регрессии к логистической
 - Математическая теория этого алгоритма
 - Простая реализация логистической регрессии для задачи классификации



Логистическая регрессия

- Обзор раздела
 - Интерпретация результатов
 - Odds Ratio и коэффициенты
 - Метрики классификации
 - Accuracy, Precision, Recall
 - ROC-графики



Логистическая регрессия

- Обзор раздела
 - Мультиклассовая классификация
 - Проект по логистической регрессии
 - Решения проекта по логистической регрессии



Логистическая регрессия

- Логистическая регрессия позволяет предсказывать значения категориальной переменной на основе исторических данных
- Целевая категориальная переменная может принимать два или более дискретных значений (классов)



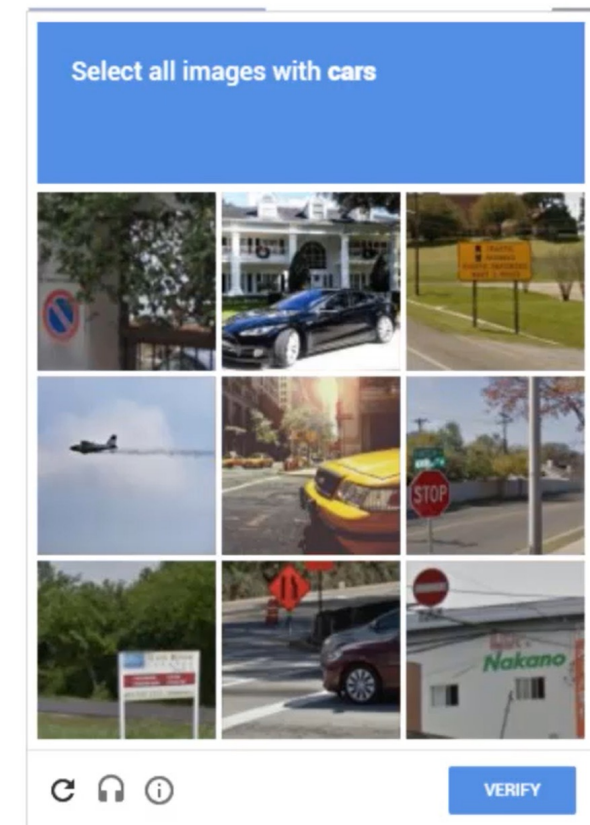
Логистическая регрессия

- Пример – указать для картинки один из классов:
 - Класс 0: изображение автомобиля
 - Класс 1: изображение улицы
 - Класс 2: изображение моста



Логистическая регрессия

- Пример – указать для картинки один из классов:
 - Класс 0: изображение автомобиля
 - Класс 1: изображение улицы
 - Класс 2: изображение моста
- Вы помогаете Google классифицировать фотографии





Логистическая регрессия

- Непрерывные переменные можно перевести в дискретные с помощью диапазонов:
 - Класс 0: цена дома \$0-100к
 - Класс 1: цена дома \$100-200к
 - Класс 2: цена дома >\$200к



Логистическая регрессия

- Зачастую алгоритмы на выходе выдают вероятность принадлежности тому или иному классу:
 - Класс 0: вероятность 10%
 - Класс 1: вероятность 85%
 - Класс 2: вероятность 5%



Логистическая регрессия

- Зачастую алгоритмы на выходе выдают вероятность принадлежности тому или иному классу:
 - Класс 0: вероятность 10% - изображение автомобиля
 - Класс 1: вероятность 85% - изображение улицы
 - Класс 2: вероятность 5% - изображение моста
 - Модель на выходе говорит “изображение улицы”



Логистическая регрессия

- Для дискретных переменных мы не сможем посчитать метрики отклонения $y - \hat{y}$.
 - Нету “изображение моста минус изображение улицы”
- Нам нужно будет разработать другой набор метрик для оценки моделей.



Логистическая регрессия

Часть 1: логистическая функция



Логистическая регрессия

- Логистическая регрессия получается преобразованием линейной регрессии в модель классификации с помощью логистической функции:

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



Логистическая регрессия

- Начнём с истории возникновения логистической функции (её ещё называют сигмоид-функцией).

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



Логистическая регрессия

- Замечание – пока мы говорим только о самой функции, а не о модели логистической регрессии!

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



Логистическая регрессия

- 1830-1850: под руководством Адольфа Кетле, Пьер Франсуа Ферхюльст разработал функцию регрессии:



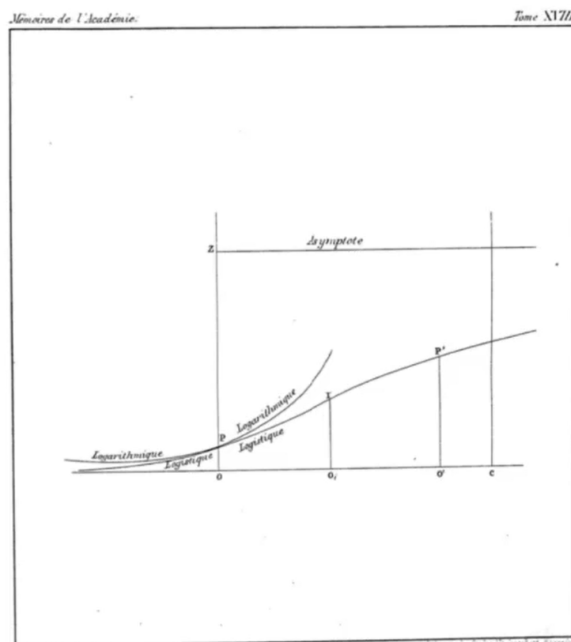
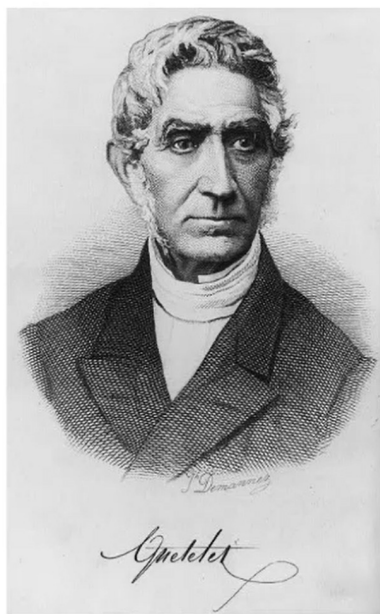
$$\sigma(x) = \frac{1}{1 + e^{-x}}$$





Логистическая регрессия

- 1830-1850: моделировали рост населения планеты



Mémoire sur la population par M. P. Verhulst.



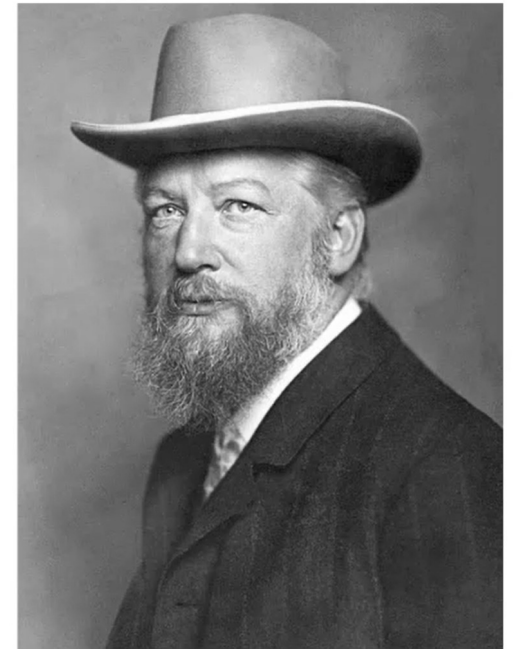
P. F. VERHULST.



Логистическая регрессия

- 1883: Вильгельм Фридрих Оствальд независимо разработал эту функцию в химии для автокатализа

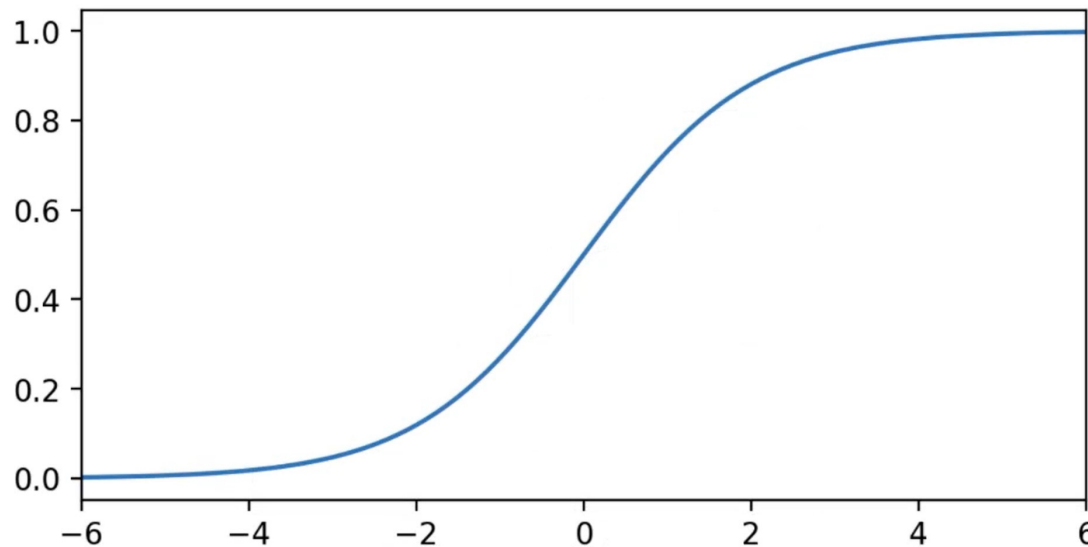
$$\sigma(x) = \frac{1}{1 + e^{-x}}$$





Логистическая регрессия

- Зачем понадобилась логистическая функция?
Почему она лучше логарифмической?

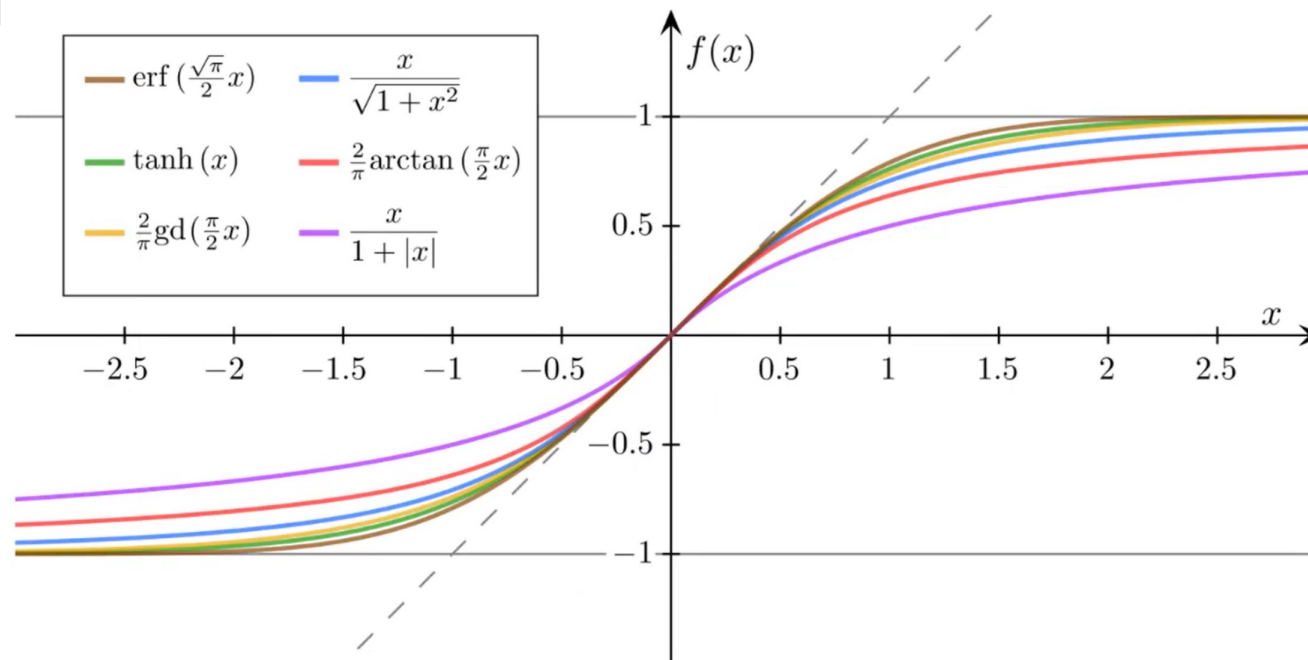


$$\sigma(x) = \frac{1}{1 + e^{-x}}$$



Логистическая регрессия

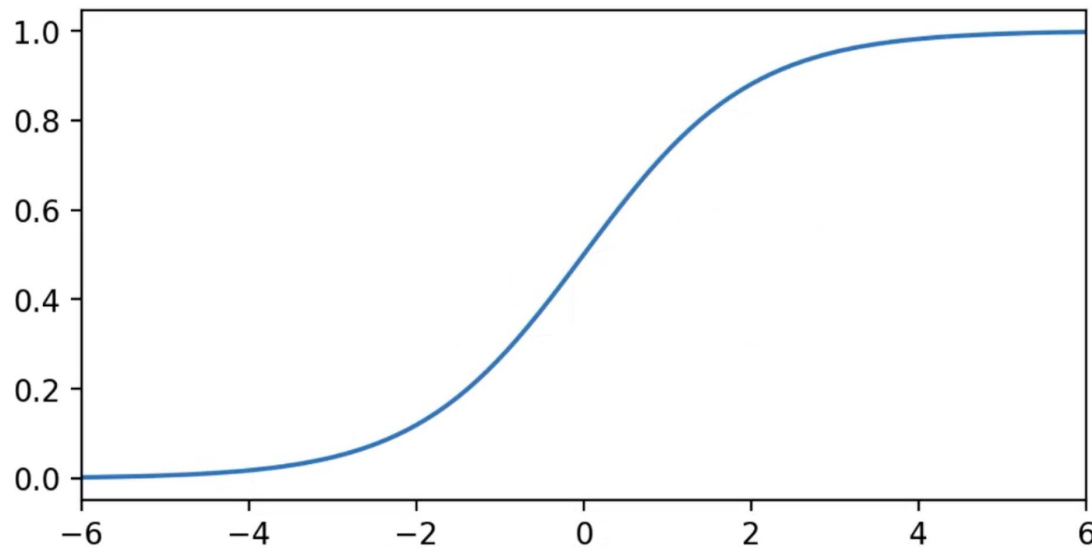
- Замечание: есть целое семейство логистических функций





Логистическая регрессия

- Обратите внимание, что кривая выходит на плато

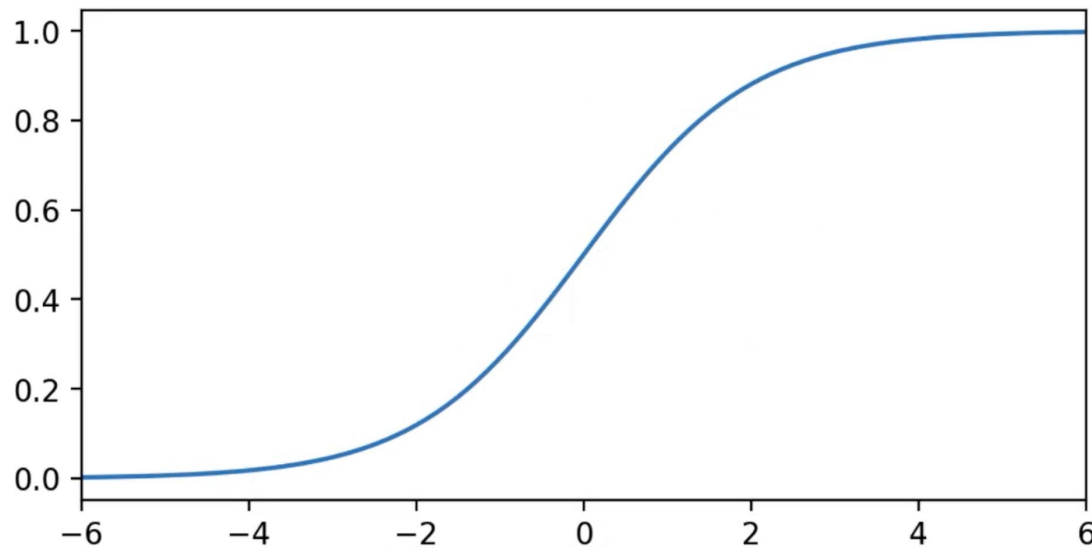


$$\sigma(x) = \frac{1}{1 + e^{-x}}$$



Логистическая регрессия

- Обратите внимание, что кривая выходит на плато
- Для любого x функция вернёт значение от 0 до 1

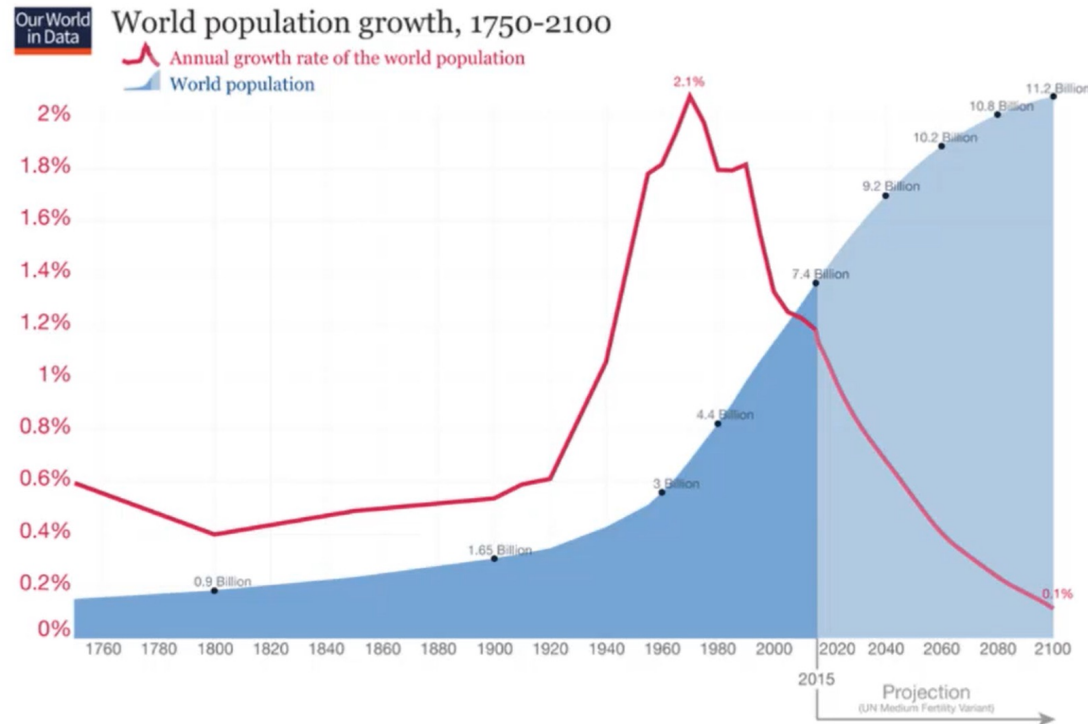


$$\sigma(x) = \frac{1}{1 + e^{-x}}$$



Логистическая регрессия

- Во многих задачах существует “естественный лимит”

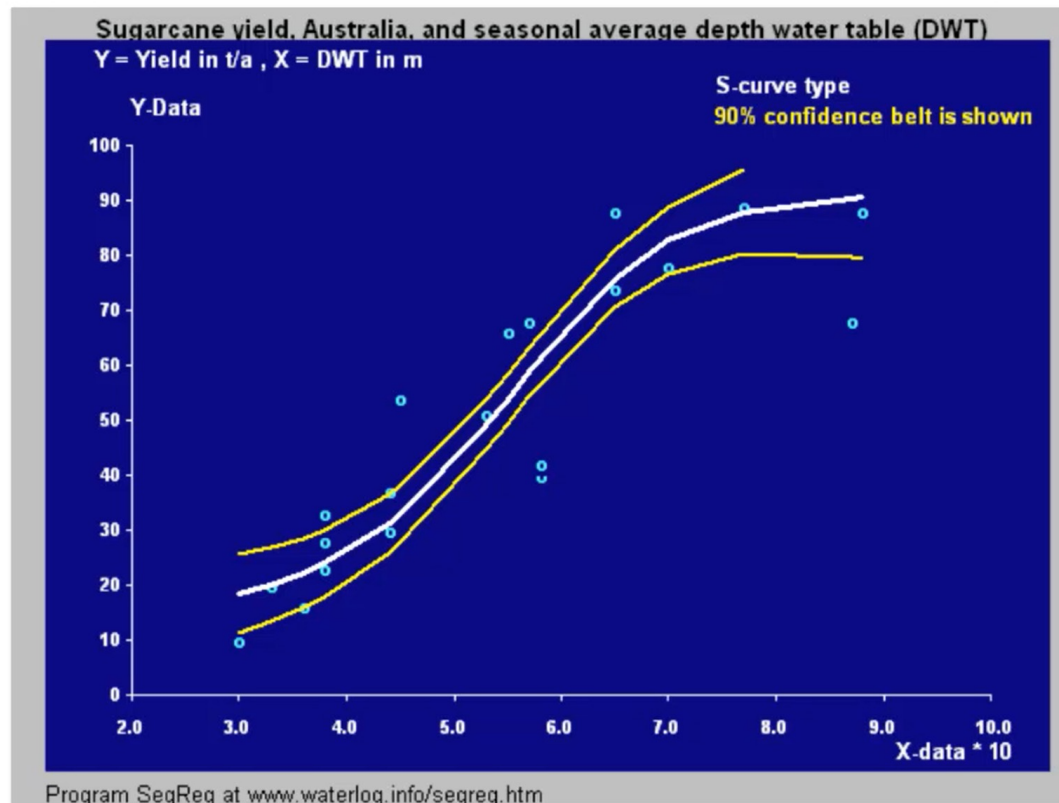


Data sources: Up to 2015 OurWorldInData series based on UN and HYDE. Projections for 2015 to 2100: UN Population Division (2015) – Medium Variant. The data visualization is taken from OurWorldInData.org. There you find the raw data and more visualizations on this topic. Licensed under CC-BY-SA by the author Max Roser.



Логистическая регрессия

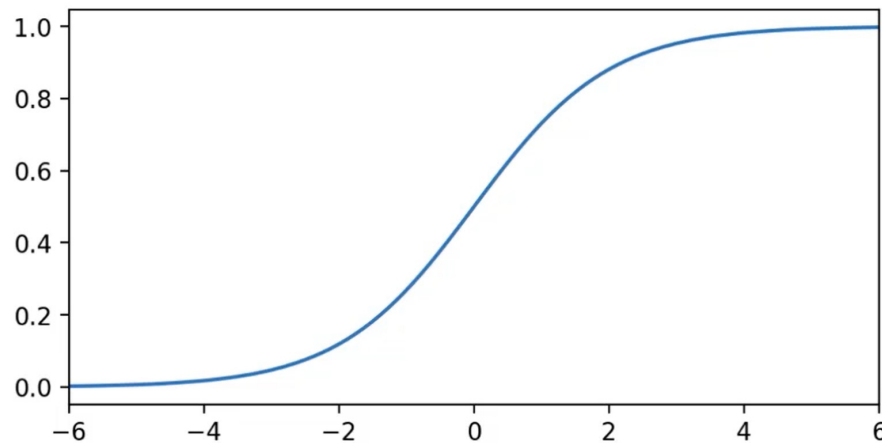
- Во многих задачах существует “естественный лимит”





Логистическая регрессия

- 1940: Джозеф Берксон применил логистическую функцию для статистического моделирования



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





Логистическая регрессия

- Интересный факт: Берксон был известным противником идеи о том, что курение вызывает рак

NOW...Scientific Evidence on Effects of Smoking!

A MEDICAL SPECIALIST in making regular bi-monthly examinations of a group of people from various walks of life, 43 percent of this group have smoked Chesterfield for an average of over ten years.

After ten months, the medical specialist reports that he observed...

no adverse effects on the nose, throat and sinuses of the group from smoking Chesterfield.

MUCH Milder CHESTERFIELD IS BEST FOR YOU

APRIL, 1952

He's one of the busiest men in town. While his door may say *Office Hours 2 to 6*, he's actually on call 24 hours a day.

The doctor is a scientist, a diplomat, and a friendly sympathetic human being all in one, no matter how long and hard his schedule.

According to a recent *Nationwide survey*:

MORE DOCTORS SMOKE CAMELS THAN ANY OTHER CIGARETTE

DOCTORS in every branch of medicine—113,000 in all—were queried in this nationwide study of cigarette preferences. Their halting research organizations made the survey. The gist of the query was: "What cigarette do you smoke, Doctor?"

The brand named most was Camels!

The rich, full flavor and cool richness of Camels' mouth blend of choice tobaccos seem to have the same appeal to the smoking tastes of doctors as to millions of other smokers. If you are a Camel smoker, this preference among doctors will hardly surprise you. If you're not—well, try Camels now.

Your "T-Zone" Will Tell You...

It's your throat... your throat... your throat... that's your growing ground for any cigarette. See if Camels don't tell you "T-Zone" is a "T-Zone."

CAMELS *Crestler Tobacco*

leaves another 'very doubtful'

and cell alteration than the lungs of nonsmokers. The peculiar lack of change also suggested how smoking might contribute to a variety of lung diseases—and how the lung damage might, in turn, lead to circulatory trouble and a harder-working heart—though this has never been directly demonstrated in any laboratory experiment.

It was these "mazy lines of converging evidence," in the committee's phrase, that led the committee to its most serious charges against cigarettes:

- **Lung Cancer:** "Cigaret smoking is causally related to lung cancer in men; the magnitude of the effect of cigarette smoking for cut-weight is still under study. The data for women, though less extensive, points in the same direction.

The risk of developing lung cancer increases with duration of smoking and the number of cigarettes smoked per day, and is diminished by discontinuing smoking."
- **Other Lung Diseases:** "Cigaret smoking is the most important of the causes of chronic bronchitis in the United States and increases the risk of dying from chronic bronchitis," as well as from pulmonary emphysema, destruction of tiny air sacs that perform the lungs' basic breathing function.
- **Heart Disease:** "Male cigarette smokers have a higher death rate from coronary artery disease than nonsmoking males, but it is not clear that the association has causal significance."
- **The report emphasized the much lesser risk of smoking cigars and pipes. But it did point out that pipe smoking seems to increase**

the risk of getting cancer of the lip.

Most surprising was the verdict on nicotine. Long considered the most harmful ingredient in tobacco because it is believed to be the stimulant that gives the smoker his kick, nicotine was found to be relatively harmless.

But the report and its conclusions are already being challenged by scientists as unconvincing as any on the surgeon general's committee. Though their numbers, these critics have persistently refused to accept the evidence against cigarettes as conclusive. An outspoken dissenter is Dr. Joseph Berkson of the Mayo Clinic, one of the nation's leading medical statisticians. Berkson has repeatedly challenged Hammond's studies, partly on the grounds that the population samples were not truly representative of the whole population. He and others have also pointed out that the interesting statistics have never been backed up by conclusive laboratory proof—e.g., lung cancer has never been induced in any experimental animal by the inhalation of cigarette smoke alone. Berkson's objections were carefully considered by the committee and answered to its own satisfaction. Berkson, however, still believes it "very doubtful" that smoking causes cancer of the lung.

A no-nonsense expression is Dr. Joseph Berkson of the Mayo Clinic. He is one of a small but conspicuous minority who insist that the case for lung cancer smoking with lung cancer remains essentially unproven.



Логистическая регрессия

- Сейчас мы уже точно знаем, что курение вредно!
- Но вернёмся к логистической регрессии..
- Далее мы посмотрим на задачи классификации, почему их нельзя решить с помощью линейной регрессии, и как нам поможет логистическая регрессия!



Логистическая регрессия

Часть 2:

от линейной регрессии к логистической



Логистическая регрессия

- Посмотрим, как можно перевести модель линейной регрессии для задачи регрессии в логистическую модель для задачи классификации
- Представьте набор данных с одним признаком “доход в прошлом году” и одной целевой переменной “невыполнение кредита”



Логистическая регрессия

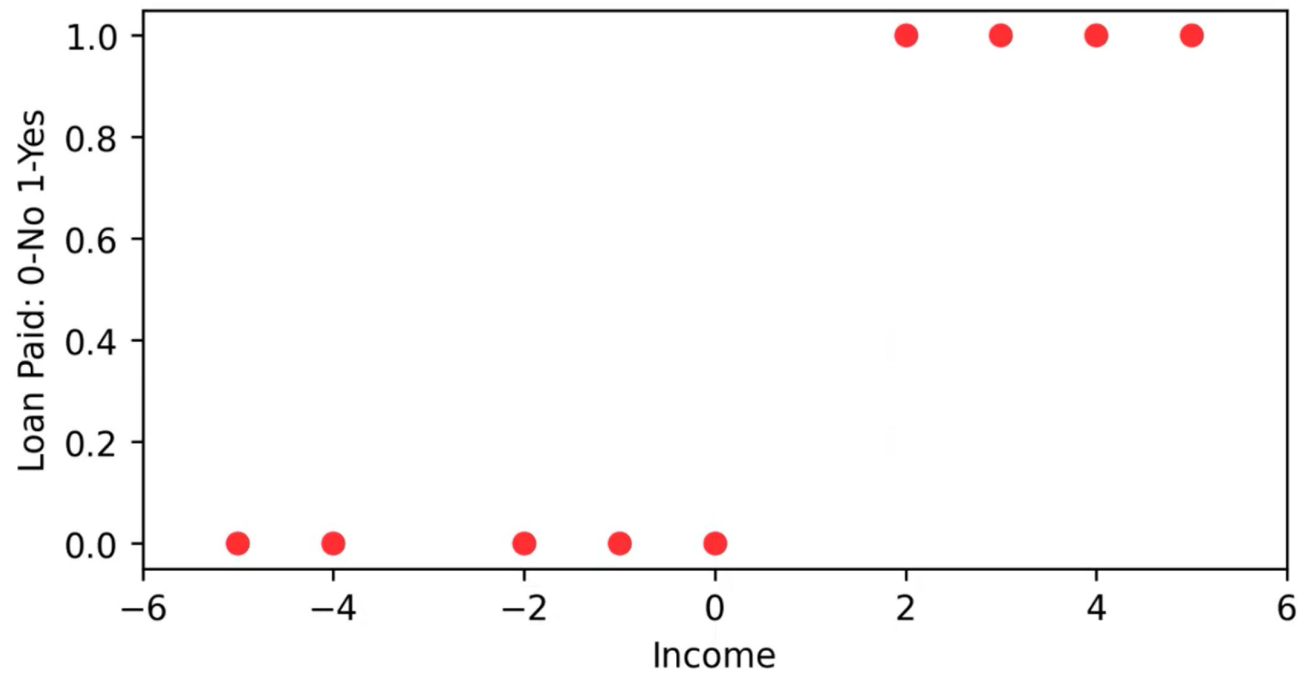
- Данные:

Income	Loan Paid
-5	0
-4	0
-2	0
-1	0
0	0
2	1
3	1
4	1
5	1



Логистическая регрессия

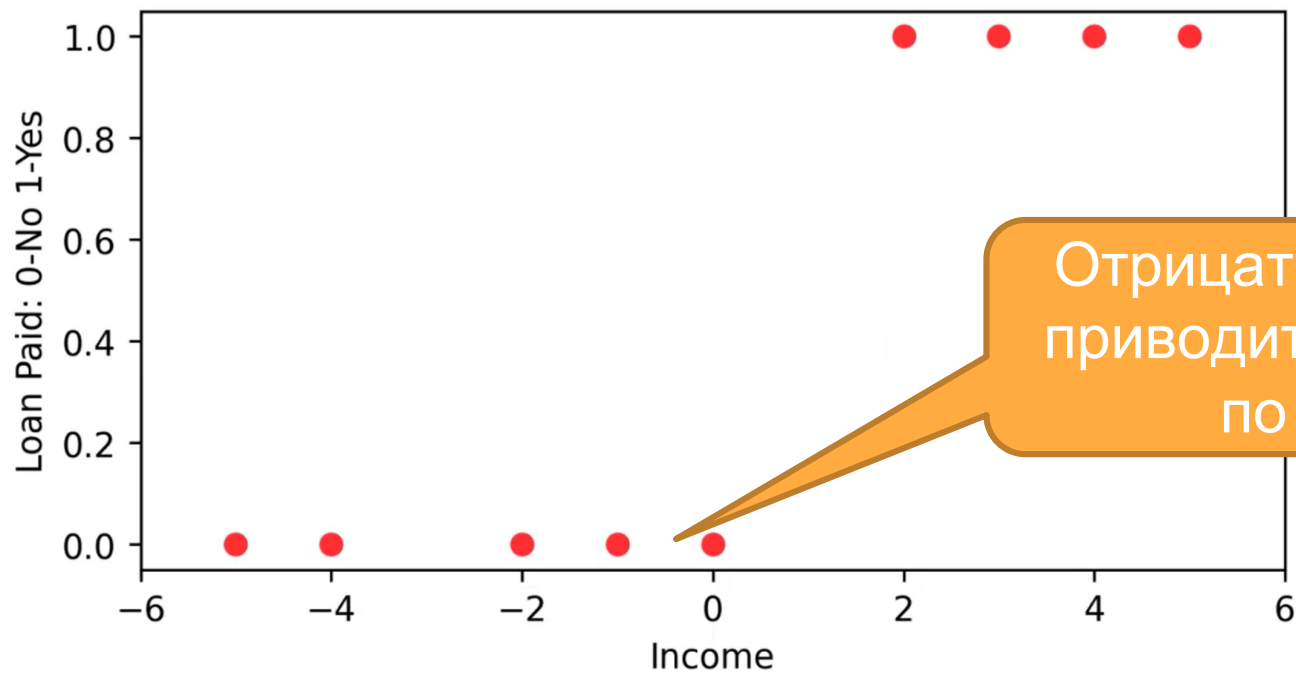
- Начнём с нанесения данных на график:





Логистическая регрессия

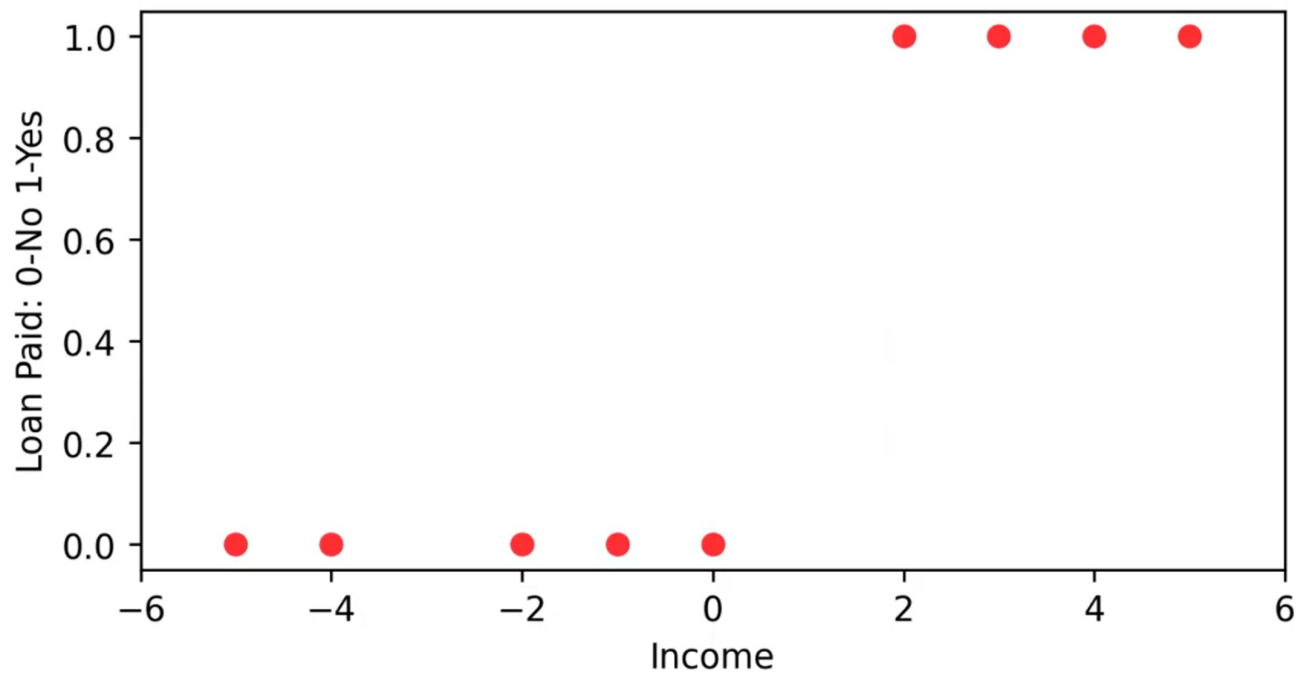
- Начнём с нанесения данных на график:





Логистическая регрессия

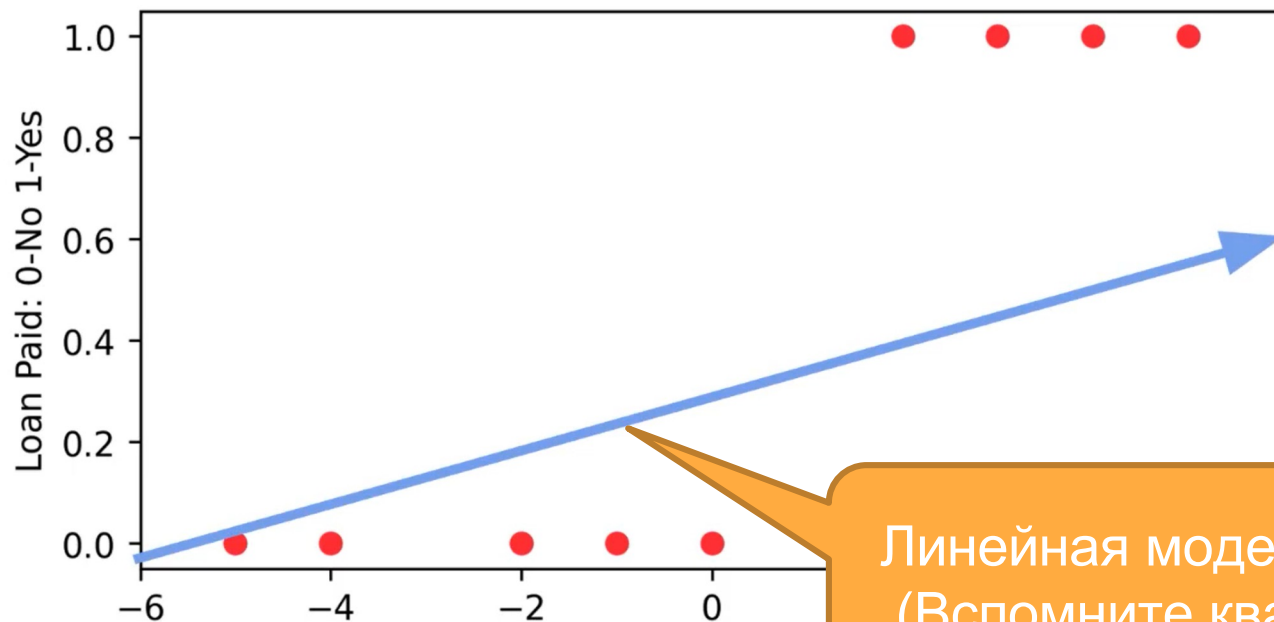
- Представьте, что мы хотим предсказать выполнение обязательств по кредиту на основе дохода:





Логистическая регрессия

- Представьте, что мы хотим предсказать выполнение обязательств по кредиту на основе дохода:

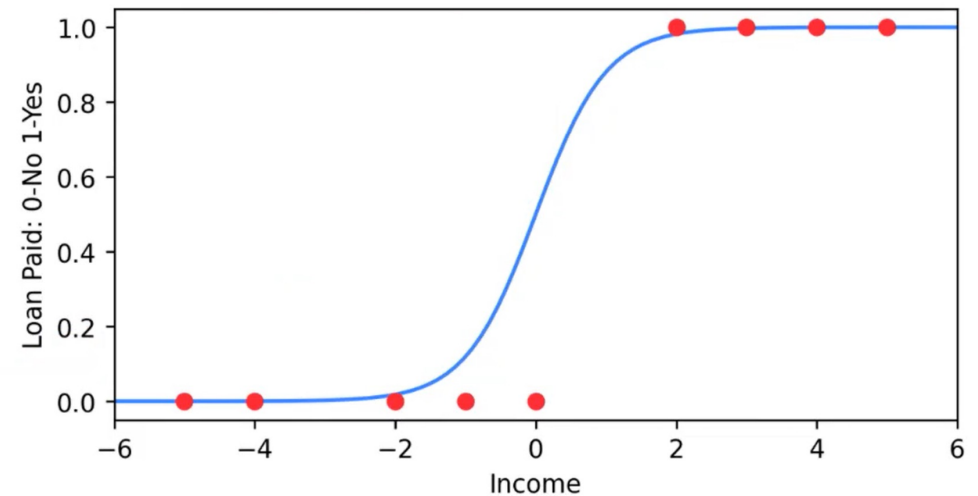
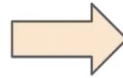
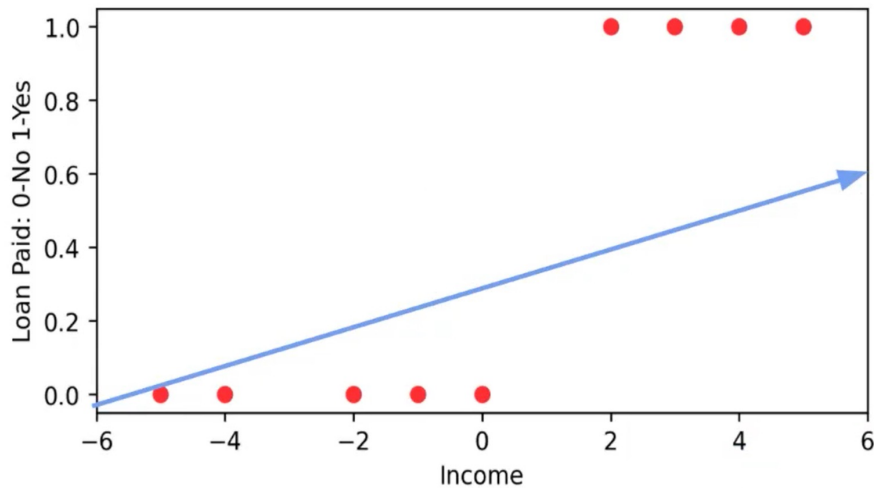


Линейная модель не сработает!
(Вспомните квартет Энскомба)



Логистическая регрессия

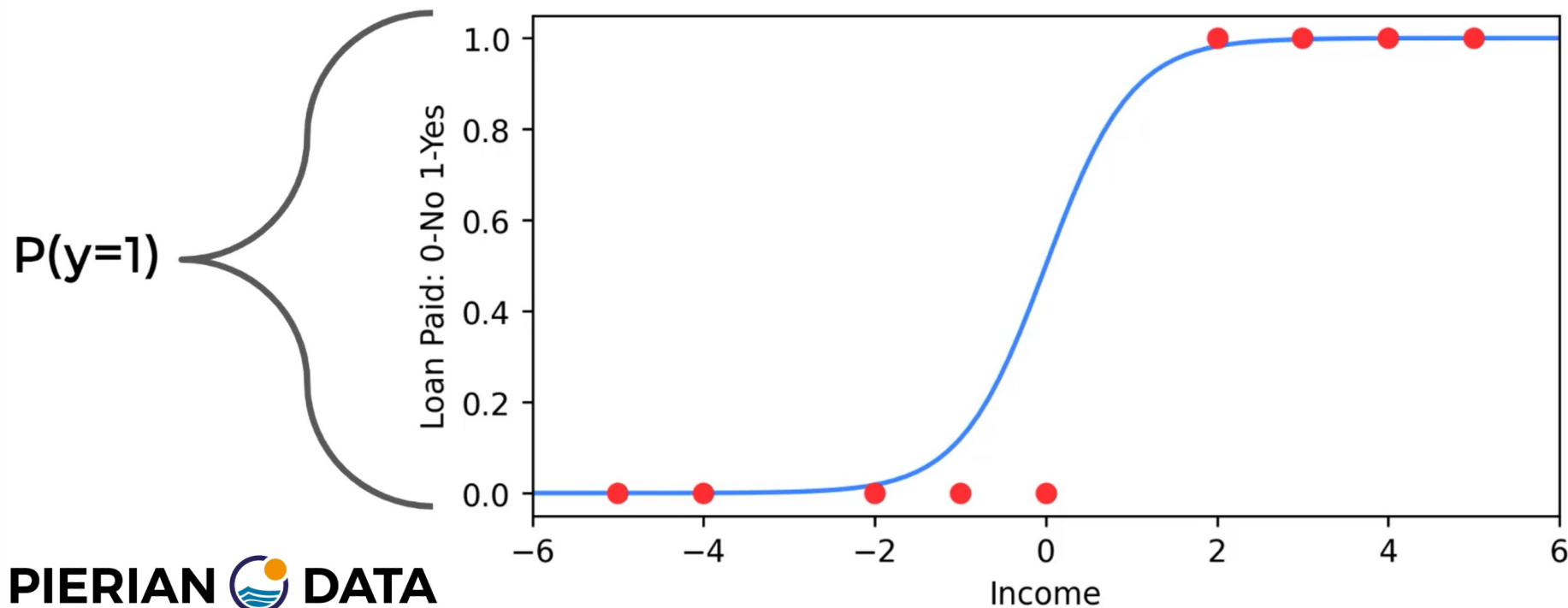
- Мы можем применить логистическую функцию:





Логистическая регрессия

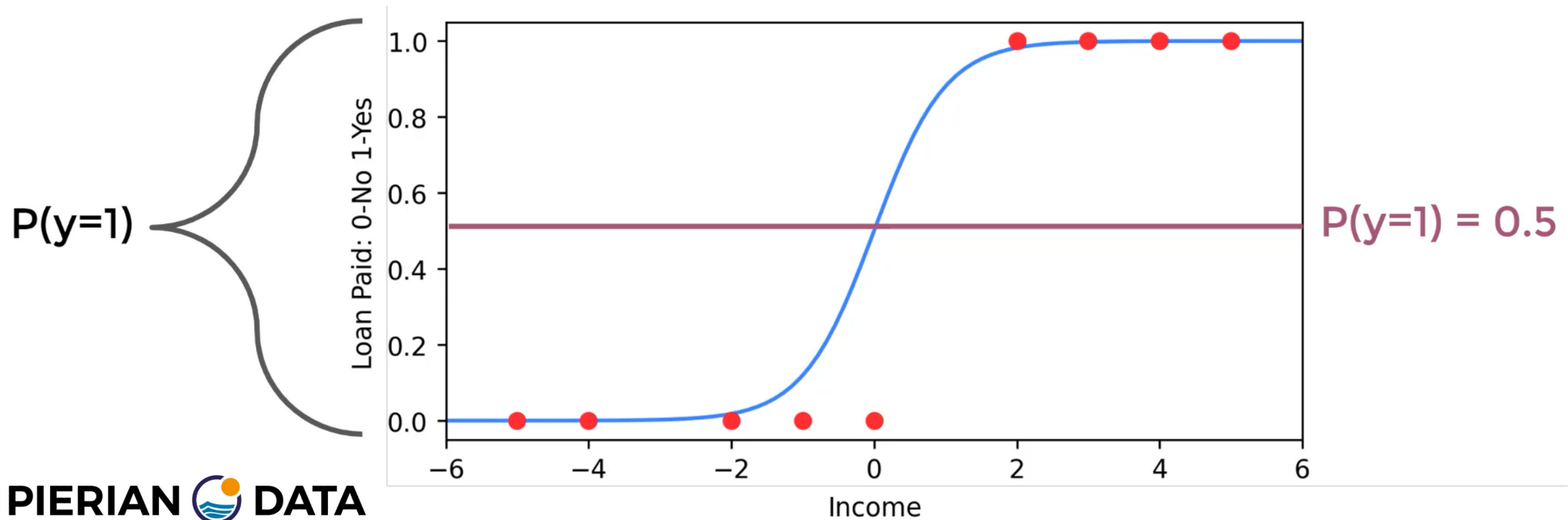
- Ось y можно интерпретировать как вероятность принадлежности к тому или иному классу:





Логистическая регрессия

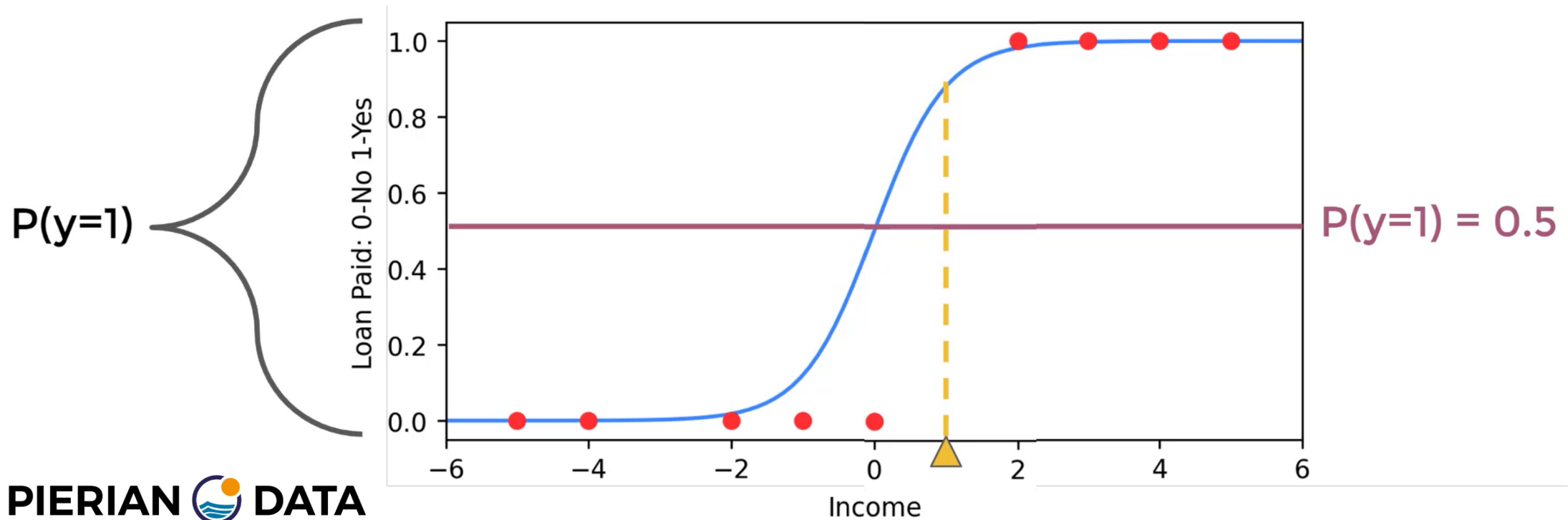
- $P(y=1) \geq 0.5$ как точка отсечки для классификации:





Логистическая регрессия

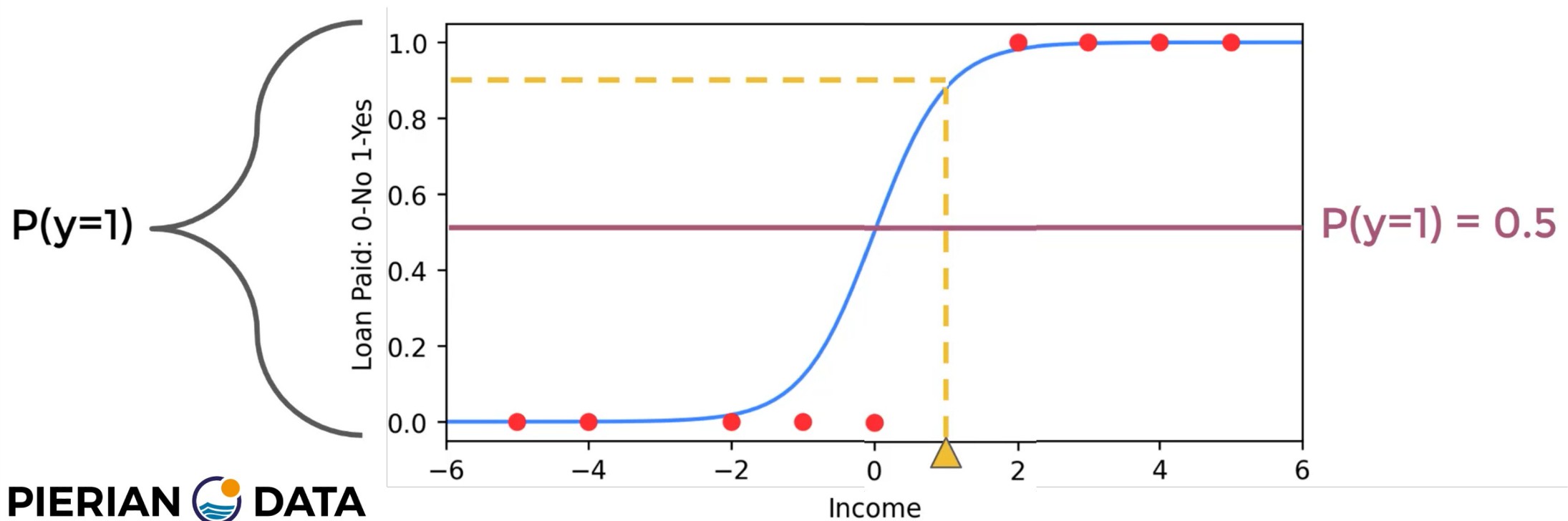
- Пример: income = 1.





Логистическая регрессия

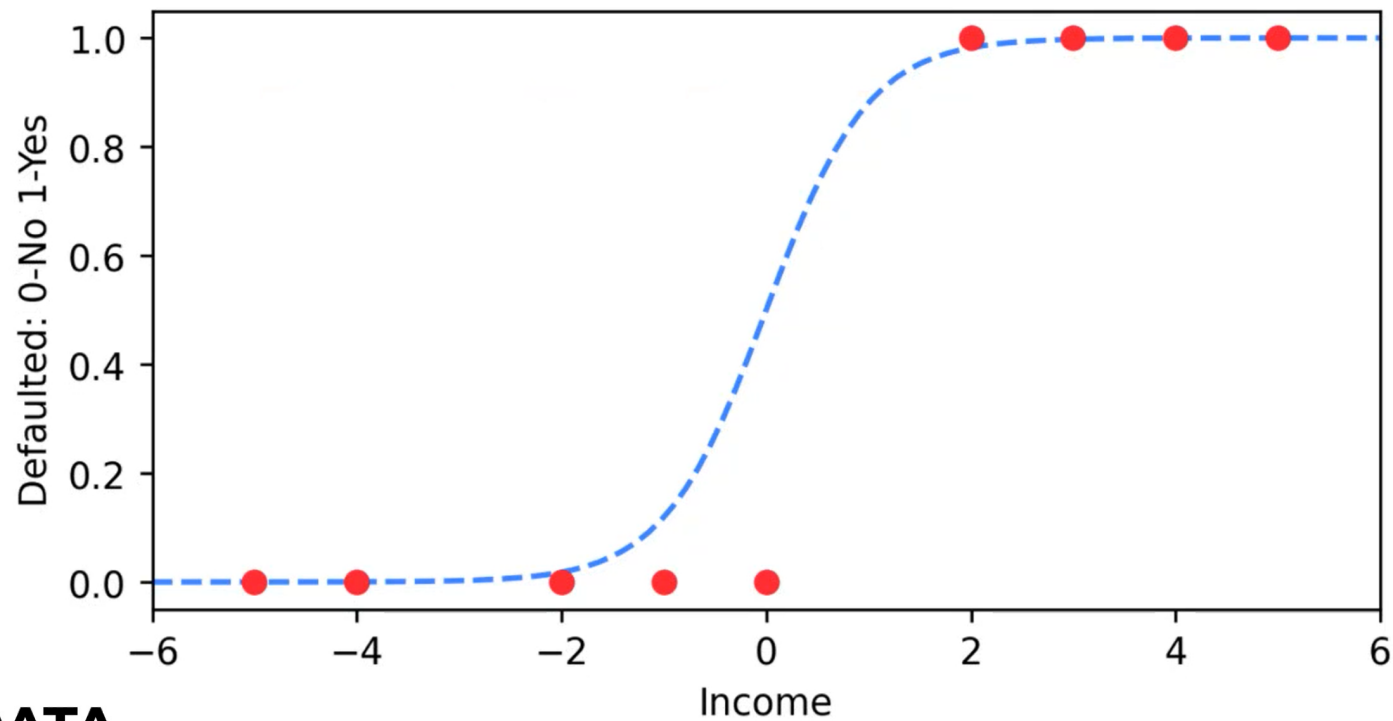
- Пример: $\text{income} = 1$.
- Вероятность возврата кредита 90%.





Логистическая регрессия

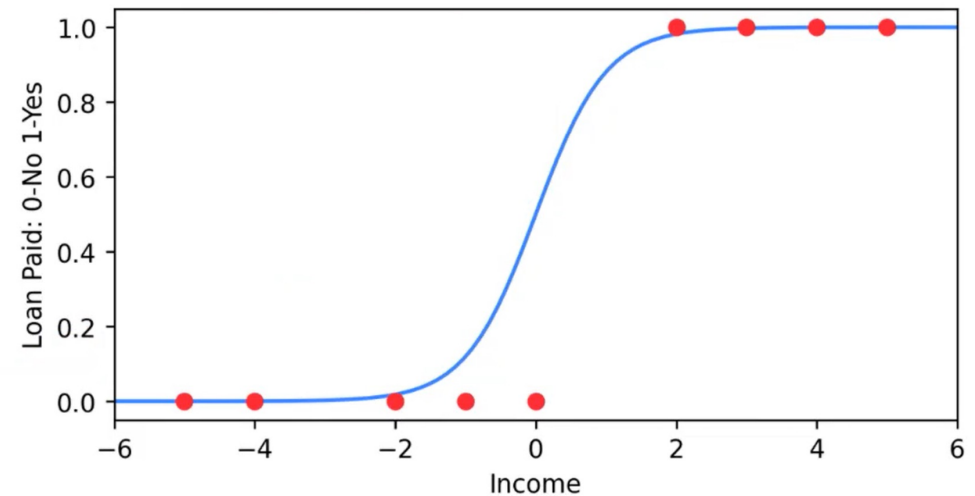
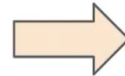
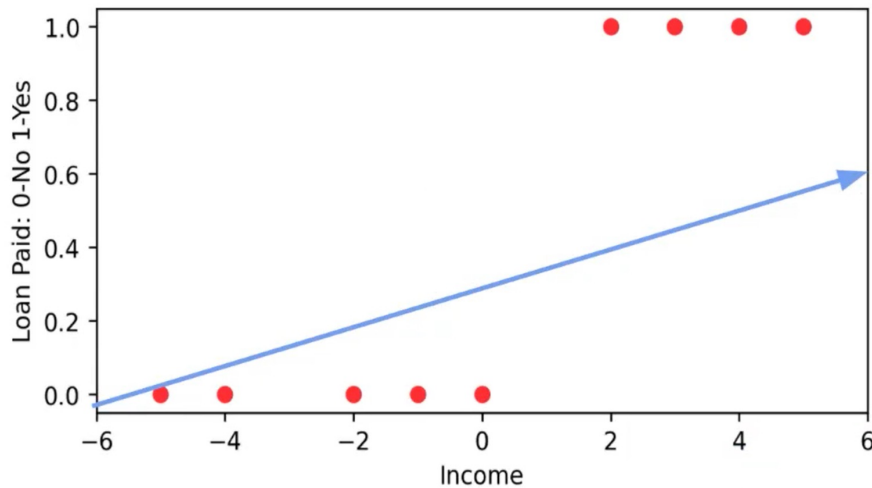
- Но как нарисовать такую линию?





Логистическая регрессия

- К счастью, математическое преобразование простое – мы рассмотрим его в следующей лекции





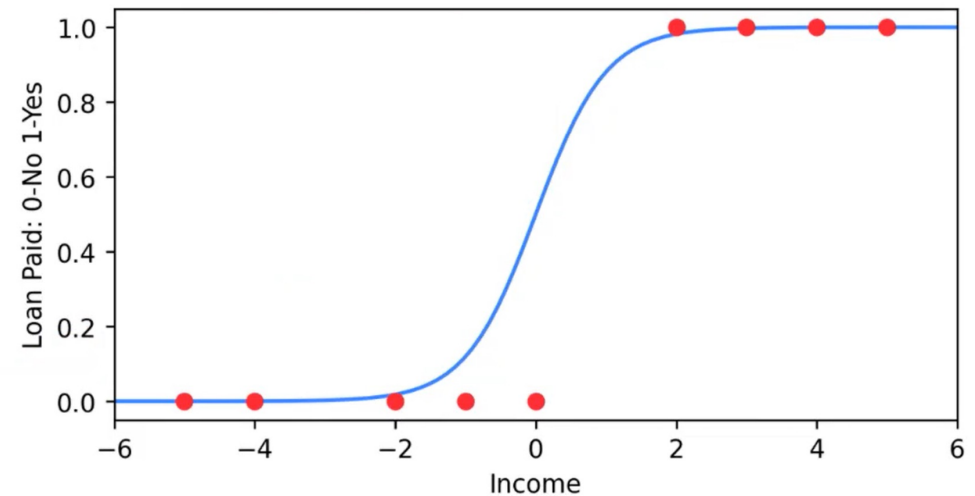
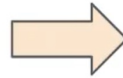
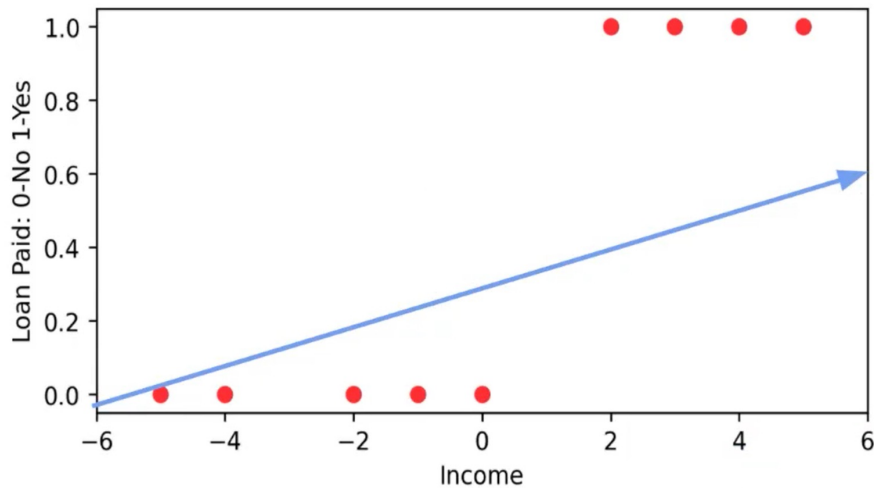
Логистическая регрессия

Часть 3: математика перехода
от линейной регрессии к логистической



Логистическая регрессия

- Посмотрим путь перехода от линейной регрессии к логистической





Логистическая регрессия

- Мы знаем уравнение линейной регрессии

$$\hat{y} = \beta_0 x_0 + \dots + \beta_n x_n$$

$$\hat{y} = \sum_{i=0}^n \beta_i x_i$$



Логистическая регрессия

- И мы знаем логистическую функцию, которая для любого значения возвращает число от 0 до 1

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



Логистическая регрессия

- Достаточно просто подставить уравнение линейной регрессии в логистическую функцию, и мы получим логистическую регрессию.

$$\hat{y} = \beta_0 x_0 + \dots + \beta_n x_n$$

$$\hat{y} = \sum_{i=0}^n \beta_i x_i$$

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



Логистическая регрессия

- В терминах логистической функции:

$$\hat{y} = \sigma(\beta_0 x_0 + \dots + \beta_n x_n)$$

$$\hat{y} = \sigma\left(\sum_{i=0}^n \beta_i x_i\right)$$



Логистическая регрессия

- Более полно:

$$\hat{y} = \frac{1}{1 + e^{-\sum_{i=0}^n \beta_i x_i}}$$



Логистическая регрессия

- Более полно:

$$\hat{y} = \frac{1}{1 + e^{-\sum_{i=0}^n \beta_i x_i}}$$

Как интерпретировать
коэффициенты бетта?



Логистическая регрессия

- Для этого нам понадобится термин “odds” (шансы).
- Этот термин часто используется в букмекерских конторах – каковы шансы, что на скачках первой придёт та или иная лошадь.





Логистическая регрессия

- Например, шансы “10 к 1”.
- Но откуда взялся этот термин?





Логистическая регрессия

- Шансы (odds) события с вероятностью p вычисляется как вероятность возникновения этого события, делённая на вероятность не возникновения этого события.

$$\frac{p}{1 - p}$$



Логистическая регрессия

- Предоставьте событие, возникающее с вероятностью 50%. Для него $0.5/(1-0.5) = 0.5/0.5$ – это то же самое, что $1/1$ – шансы 1 к 1

$$\frac{p}{1 - p}$$



Логистическая регрессия

- Мы можем переписать формулу ниже так, чтобы показать что это эквивалентно логарифму шансов (odds) в виде линейной комбинации признаков

$$\hat{y} = \frac{1}{1 + e^{-\sum_{i=0}^n \beta_i x_i}}$$



Логистическая регрессия

- Это позволит нам решить это уравнение для коэффициентов и признаков x в терминах логарифмов шансов (log odds)

$$\hat{y} = \frac{1}{1 + e^{-\sum_{i=0}^n \beta_i x_i}}$$



Логистическая регрессия

- Решаем для логарифмов шансов (log odds)

$$\hat{y} = \frac{1}{1 + e^{-\sum_{i=0}^n \beta_i x_i}}$$



Логистическая регрессия

- Решаем для логарифмов шансов (log odds)

$$\hat{y} = \frac{1}{1 + e^{-\sum_{i=0}^n \beta_i x_i}}$$

$$\hat{y} + \hat{y} e^{-\sum_{i=0}^n \beta_i x_i} = 1$$



Логистическая регрессия

- Решаем для логарифмов шансов (log odds)

$$\hat{y} + \hat{y}e^{-\sum_{i=0}^n \beta_i x_i} = 1$$



Логистическая регрессия

- Решаем для логарифмов шансов (log odds)

$$\hat{y} + \hat{y}e^{-\sum_{i=0}^n \beta_i x_i} = 1$$

$$\hat{y}e^{-\sum_{i=0}^n \beta_i x_i} = 1 - \hat{y}$$



Логистическая регрессия

- Решаем для логарифмов шансов (log odds)

$$\hat{y} + \hat{y}e^{-\sum_{i=0}^n \beta_i x_i} = 1$$

$$\hat{y}e^{-\sum_{i=0}^n \beta_i x_i} = 1 - \hat{y}$$

$$\frac{\hat{y}}{1 - \hat{y}} = e^{\sum_{i=0}^n \beta_i x_i}$$



Логистическая регрессия

- Решаем для логарифмов шансов (log odds)

$$\frac{\hat{y}}{1 - \hat{y}} = e^{\sum_{i=0}^n \beta_i x_i}$$



Логистическая регрессия

- Решаем для логарифмов шансов (log odds)

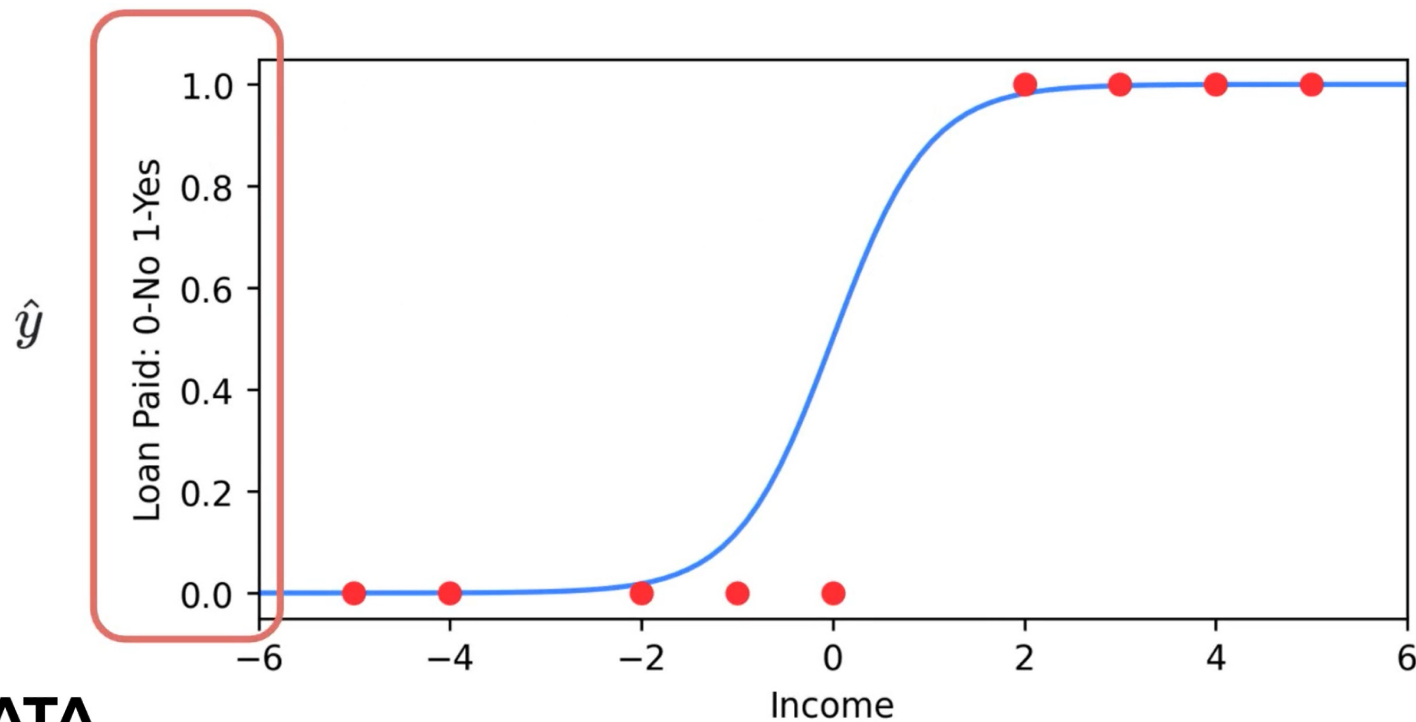
$$\frac{\hat{y}}{1 - \hat{y}} = e^{\sum_{i=0}^n \beta_i x_i}$$

$$\ln \left(\frac{\hat{y}}{1 - \hat{y}} \right) = \sum_{i=0}^n \beta_i x_i$$



Логистическая регрессия

- Как будет выглядеть график функции в терминах логарифма шансов (log odds)?

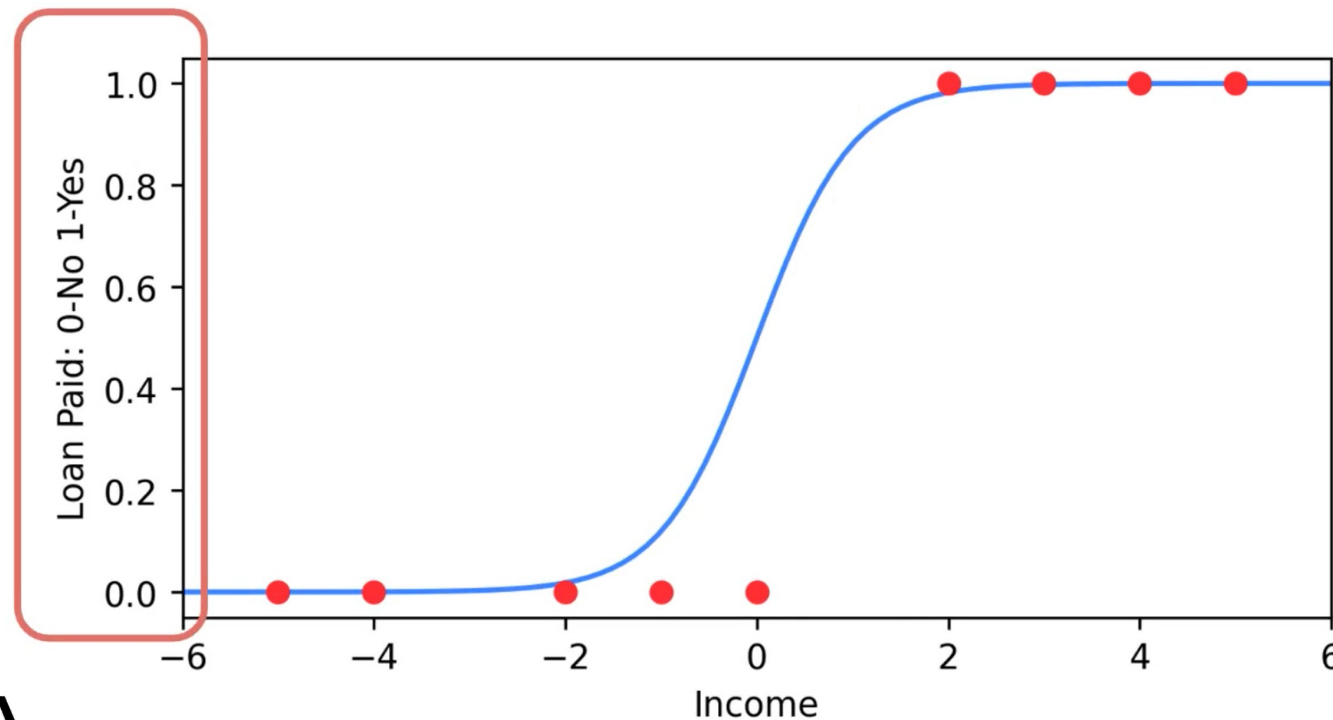




Логистическая регрессия

- Как будет выглядеть график функции в терминах логарифма шансов (log odds)?

$$\ln \left(\frac{\hat{y}}{1 - \hat{y}} \right) \leftarrow \hat{y}$$

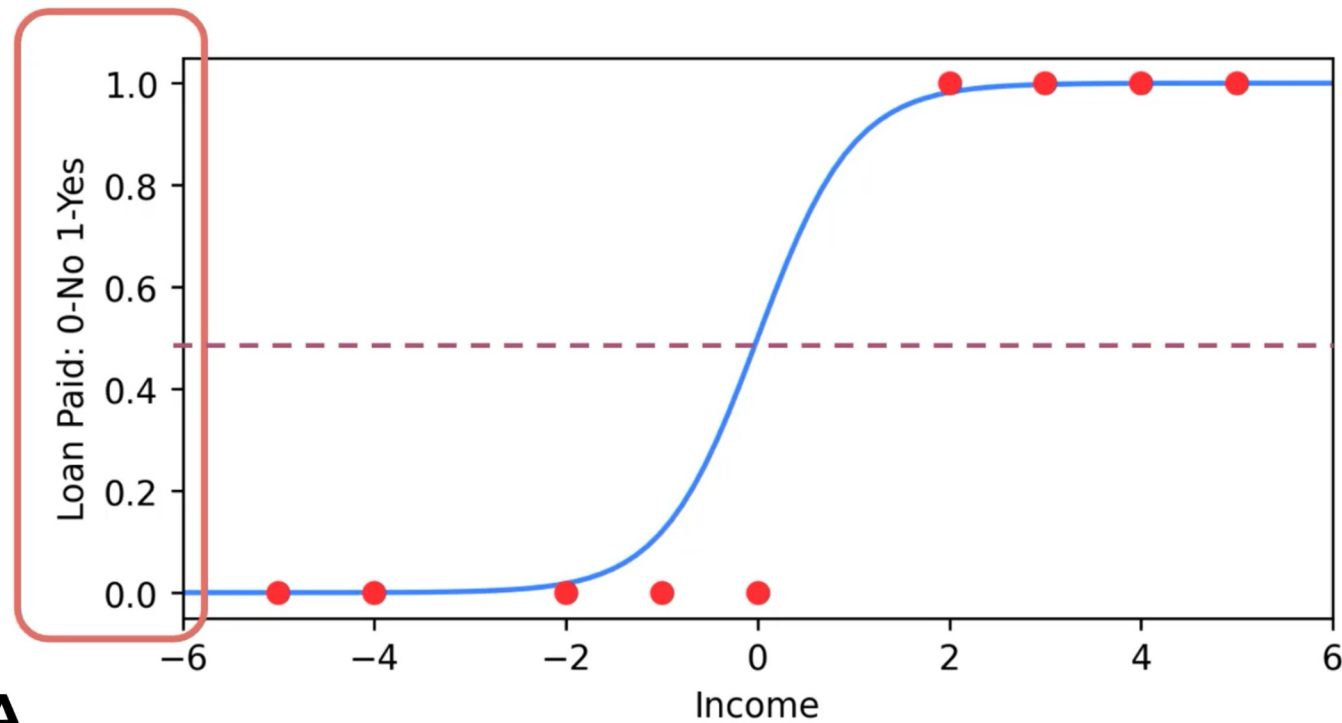




Логистическая регрессия

- Например, для $p=0.5$

$$\ln\left(\frac{0.5}{1-0.5}\right) = 0$$





Логистическая регрессия

- Например, для $p=0.5$, серединная точка в нуле.

$$\ln\left(\frac{0.5}{1 - 0.5}\right) = 0$$





Логистическая регрессия

- Когда p приближается к 1, log odds становится ∞

$$\lim_{p \rightarrow 1} \ln\left(\frac{p}{1-p}\right) = \infty$$

$$\ln\left(\frac{0.5}{1-0.5}\right) = 0$$





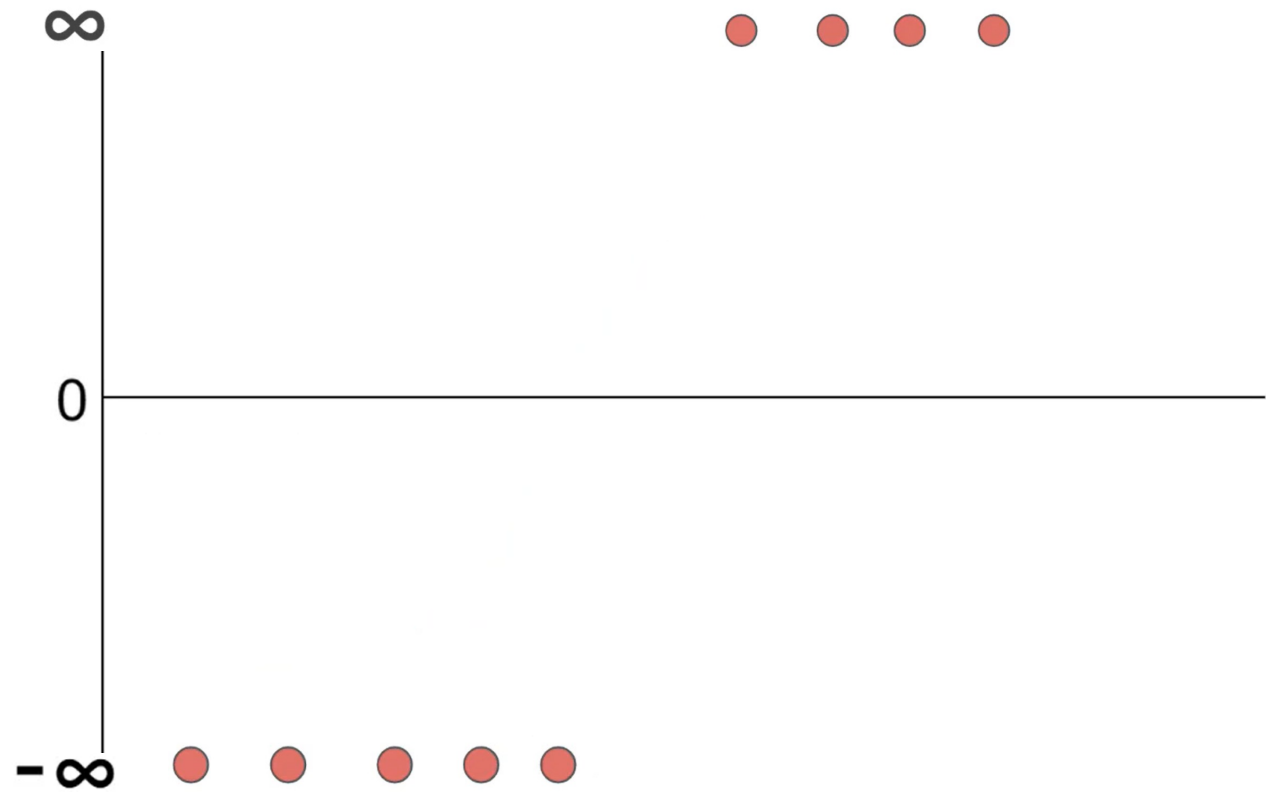
Логистическая регрессия

- Точки рядом с бесконечностями

$$\lim_{p \rightarrow 1} \ln\left(\frac{p}{1-p}\right) = \infty$$

$$\ln\left(\frac{0.5}{1-0.5}\right) = 0$$

$$\lim_{p \rightarrow 0} \ln\left(\frac{p}{1-p}\right) = -\infty$$





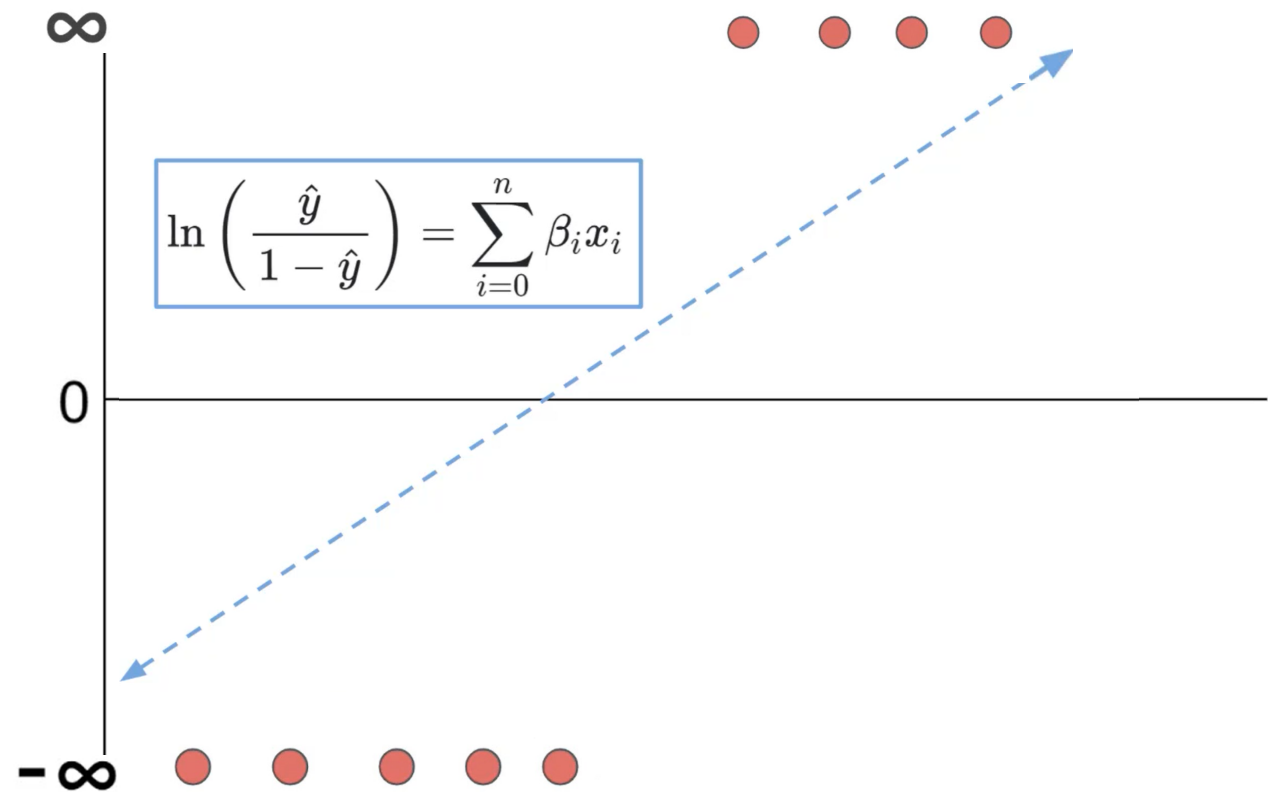
Логистическая регрессия

- Коэффициенты в терминах изменения log odds

$$\lim_{p \rightarrow 1} \ln\left(\frac{p}{1-p}\right) = \infty$$

$$\ln\left(\frac{0.5}{1-0.5}\right) = 0$$

$$\lim_{p \rightarrow 0} \ln\left(\frac{p}{1-p}\right) = -\infty$$





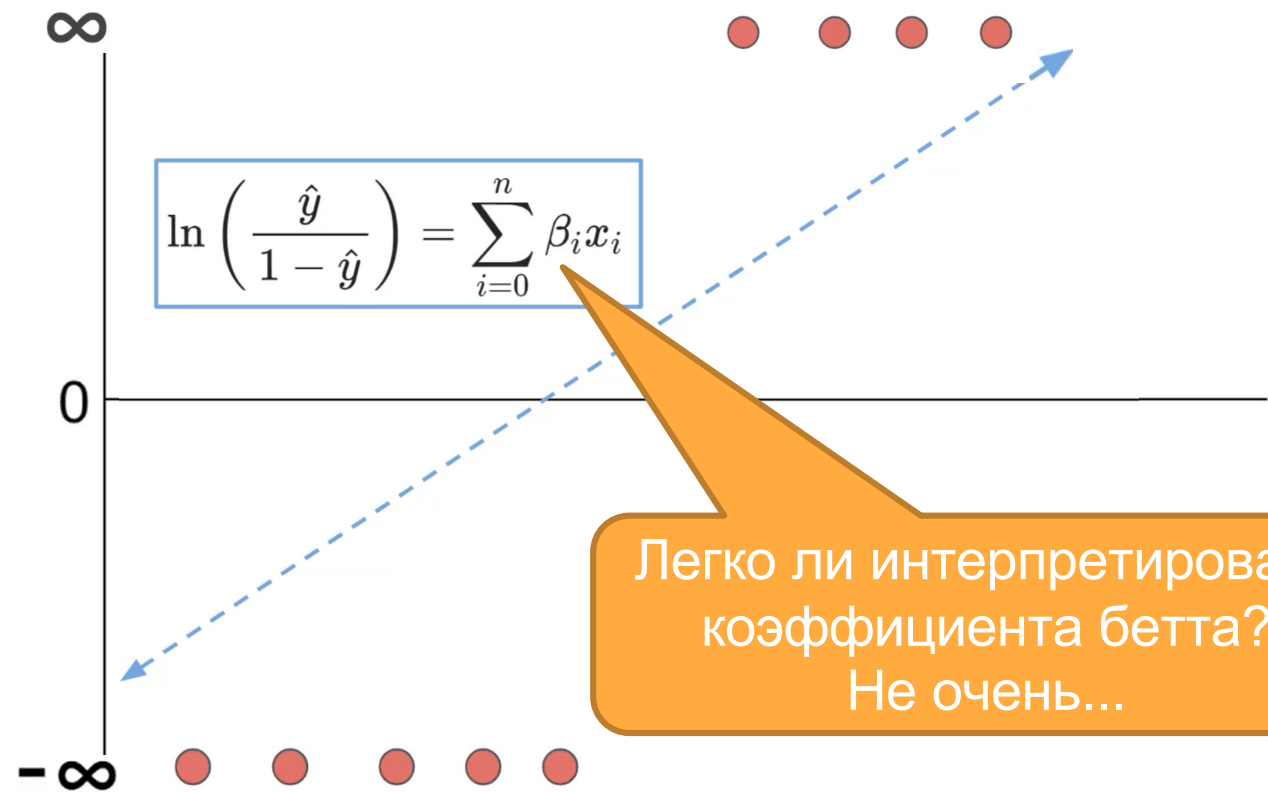
Логистическая регрессия

- Коэффициенты в терминах изменения log odds

$$\lim_{p \rightarrow 1} \ln\left(\frac{p}{1-p}\right) = \infty$$

$$\ln\left(\frac{0.5}{1-0.5}\right) = 0$$

$$\lim_{p \rightarrow 0} \ln\left(\frac{p}{1-p}\right) = -\infty$$





Логистическая регрессия

- Поскольку шкала для log odds не является линейной, коэффициенты бетта нельзя понимать как “увеличение на одну единицу”, как это было для линейной регрессии.

$$\ln \left(\frac{\hat{y}}{1 - \hat{y}} \right) = \sum_{i=0}^n \beta_i x_i$$



Логистическая регрессия

- Но тем не менее, мы можем сделать кое-какие выводы...

$$\ln \left(\frac{\hat{y}}{1 - \hat{y}} \right) = \sum_{i=0}^n \beta_i x_i$$



Логистическая регрессия

- Знак коэффициента
 - Положительные бетта означают увеличение вероятности принадлежности к классу 1 при увеличении признака x
 - Отрицательные бетта означают уменьшение вероятности принадлежности к классу 1 при увеличении признака x



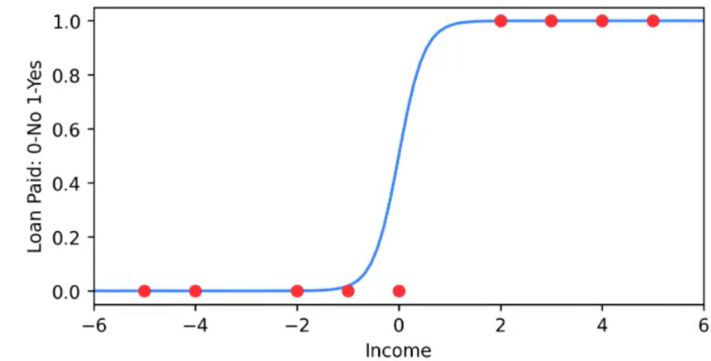
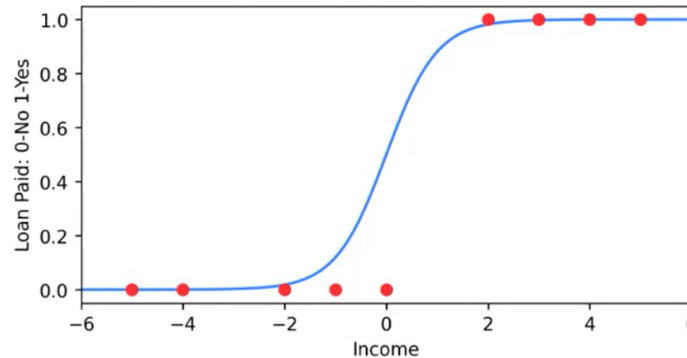
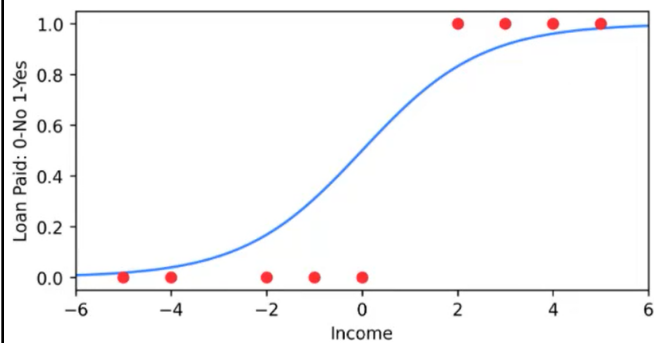
Логистическая регрессия

- Величина коэффициента
 - Сложно интерпретировать напрямую, особенно когда у нас есть и непрерывные, и дискретные признаки
 - Но мы можем сравнивать величины коэффициентов между собой, используя т.н. odds ratio.
 - Так мы можем определить, какие признаки больше влияют на результат предсказания.



Логистическая регрессия

- Последнее, что нам осталось обсудить – это как провести такую линию через точки, то есть как выполнить подгонку (fit).





Логистическая регрессия

Часть 4: подгоняем функцию (best fit)



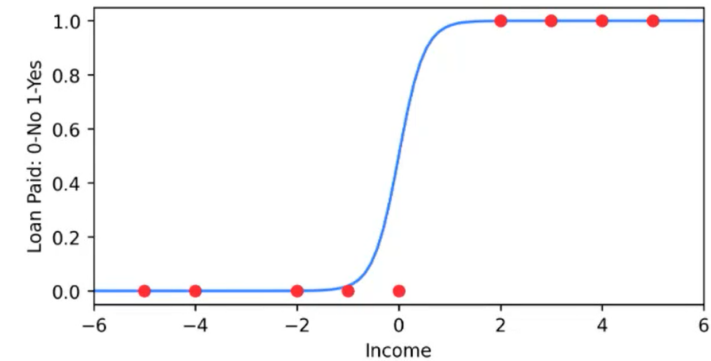
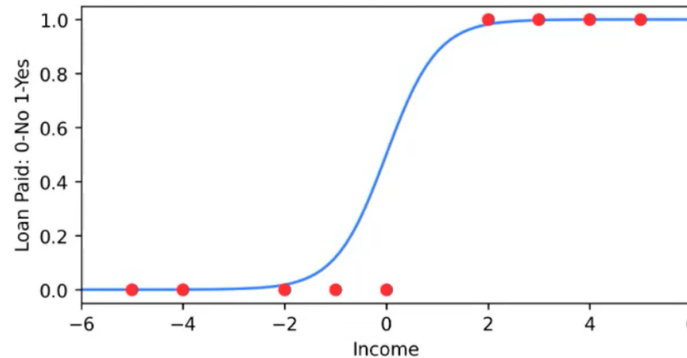
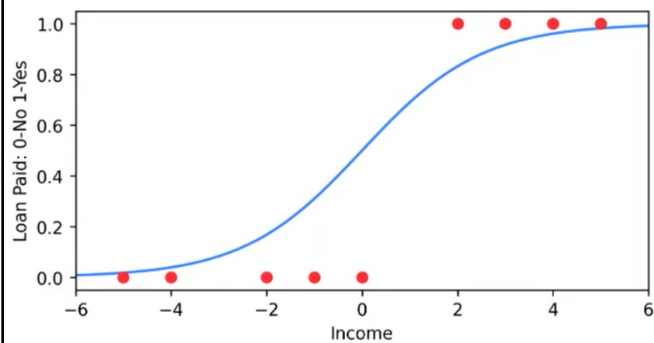
Логистическая регрессия

- Для подгонки модели логистическая регрессия использует метод максимального правдоподобия (maximum likelihood).
- В этой лекции мы посмотрим, как это работает. Также мы посмотрим на функцию стоимости и градиентный спуск, которые для нас применяет компьютер.



Логистическая регрессия

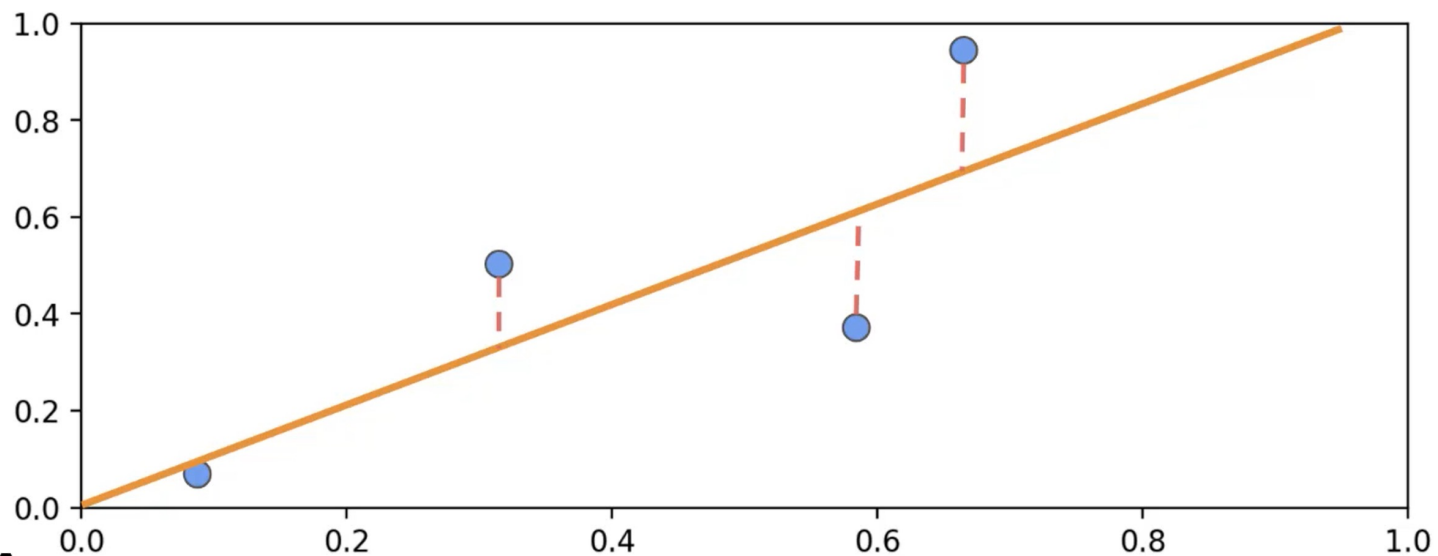
- На этом слайде мы видим три графика, для разных значений бетта.
- Какой из этих графиков лучший?





Логистическая регрессия

- Вспомните – для линейной регрессии мы минимизировали сумму квадратов остатков (RSS - residual sum of squares).

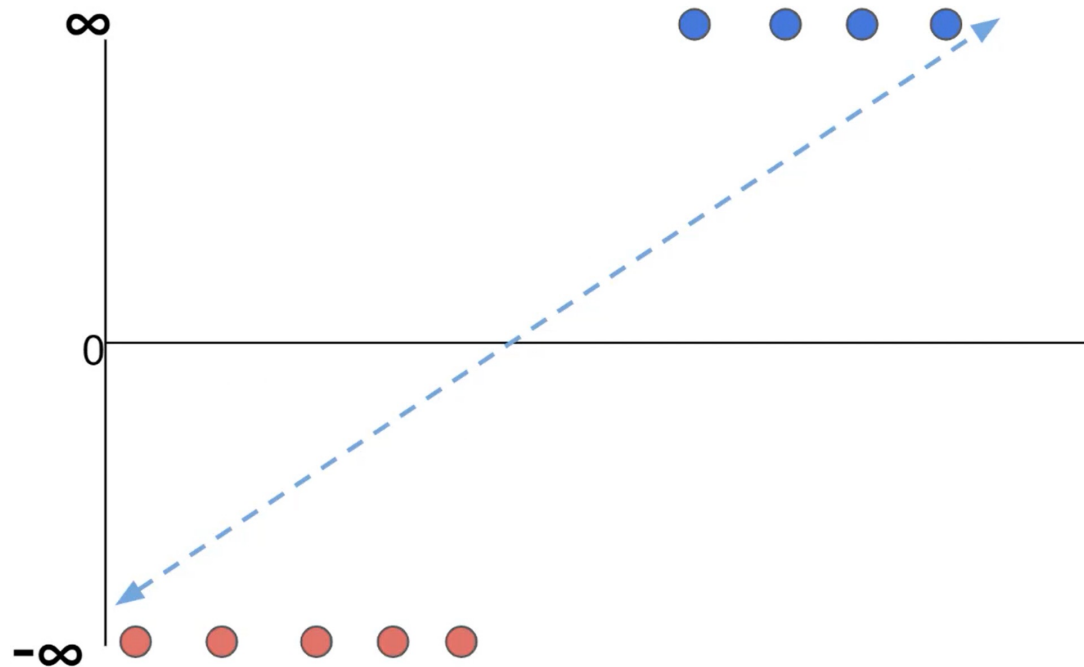




Логистическая регрессия

- К сожалению, в случае log odds значения близятся к бесконечности, и RSS здесь не подходит

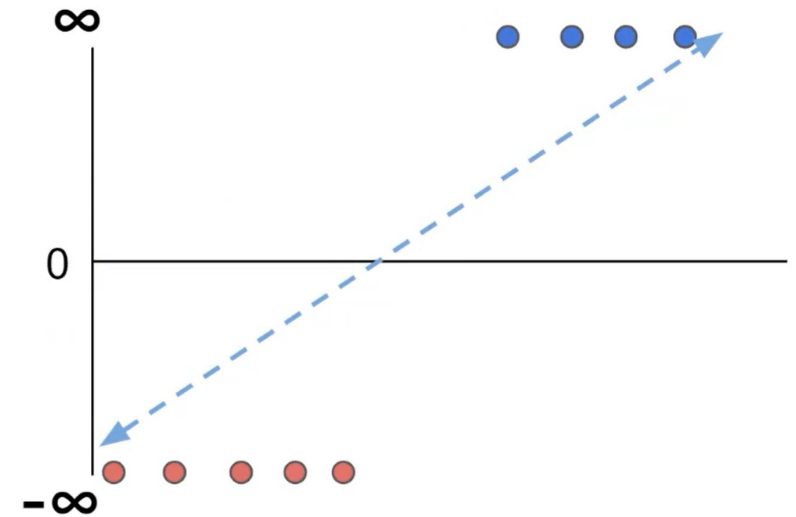
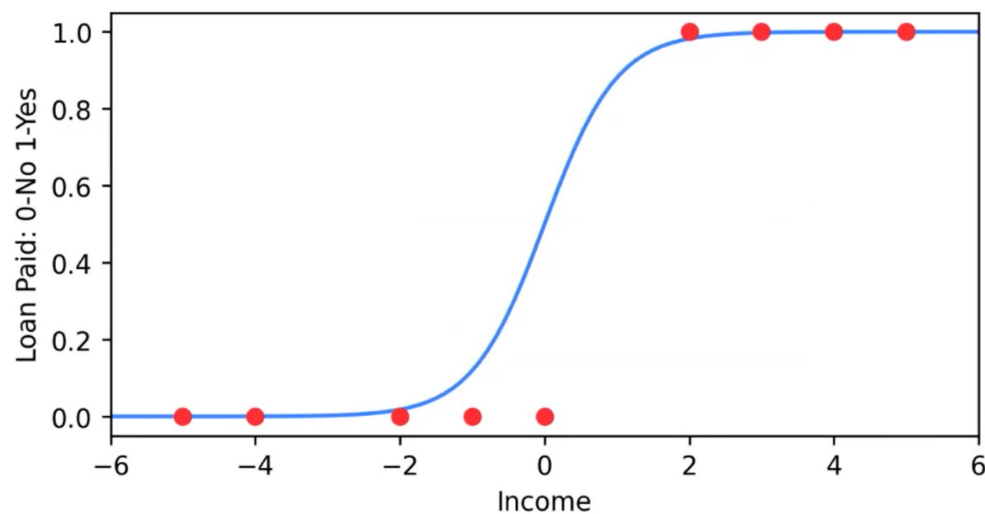
$$\ln \left(\frac{\hat{y}}{1 - \hat{y}} \right) = \sum_{i=0}^n \beta_i x_i$$





Логистическая регрессия

- Первый шаг метода максимального правдоподобия – перейти от log odds обратно к вероятностям





Логистическая регрессия

- Первый шаг метода максимального правдоподобия – перейти от log odds обратно к вероятностям

$$\ln\left(\frac{p}{1-p}\right) = \ln(odds)$$



Логистическая регрессия

- Первый шаг метода максимального правдоподобия – перейти от log odds обратно к вероятностям

$$\ln\left(\frac{p}{1-p}\right) = \ln(odds)$$

$$\frac{p}{1-p} = e^{\ln(odds)}$$



Логистическая регрессия

- Первый шаг метода максимального правдоподобия – перейти от $\log \text{odds}$ обратно к вероятностям

$$\ln\left(\frac{p}{1-p}\right) = \ln(\text{odds})$$

$$\frac{p}{1-p} = e^{\ln(\text{odds})}$$

$$p = (1-p)e^{\ln(\text{odds})}$$



Логистическая регрессия

- Первый шаг метода максимального правдоподобия – перейти от log odds обратно к вероятностям

$$p = (1 - p)e^{\ln(odds)}$$



Логистическая регрессия

- Первый шаг метода максимального правдоподобия – перейти от log odds обратно к вероятностям

$$p = (1 - p)e^{\ln(odds)}$$

$$p = e^{\ln(odds)} - pe^{\ln(odds)}$$



Логистическая регрессия

- Первый шаг метода максимального правдоподобия – перейти от log odds обратно к вероятностям

$$p = (1 - p)e^{\ln(odds)}$$

$$p = e^{\ln(odds)} - pe^{\ln(odds)}$$

$$p + pe^{\ln(odds)} = e^{\ln(odds)}$$



Логистическая регрессия

- Первый шаг метода максимального правдоподобия – перейти от log odds обратно к вероятностям

$$p = (1 - p)e^{\ln(odds)}$$

$$p = e^{\ln(odds)} - pe^{\ln(odds)}$$

$$p + pe^{\ln(odds)} = e^{\ln(odds)}$$

$$p(1 + e^{\ln(odds)}) = e^{\ln(odds)}$$



Логистическая регрессия

- Первый шаг метода максимального правдоподобия – перейти от log odds обратно к вероятностям

$$p = e^{\ln(odds)} - pe^{\ln(odds)}$$

$$p + pe^{\ln(odds)} = e^{\ln(odds)}$$

$$p(1 + e^{\ln(odds)}) = e^{\ln(odds)}$$

$$p = \frac{e^{\ln(odds)}}{1 + e^{\ln(odds)}}$$



Логистическая регрессия

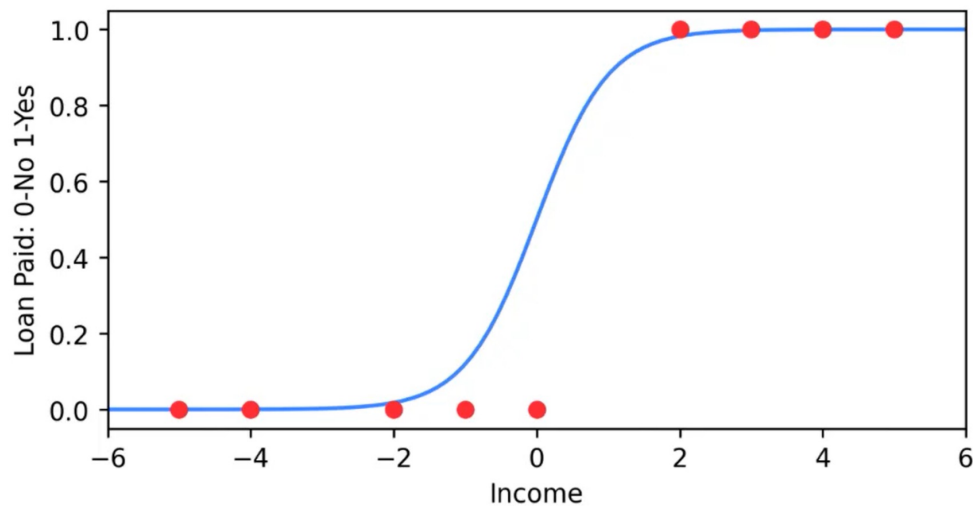
- Первый шаг метода максимального правдоподобия – перейти от log odds обратно к вероятностям

$$p = \frac{e^{\ln(odds)}}{1 + e^{\ln(odds)}}$$

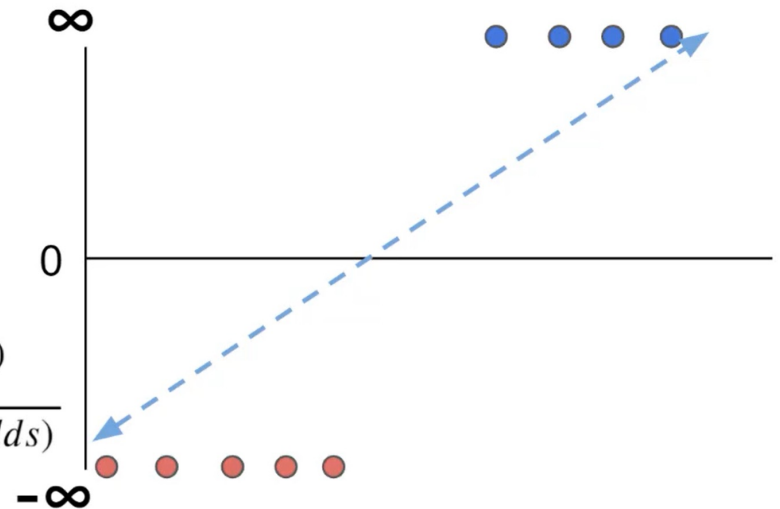


Логистическая регрессия

- Теперь мы можем перейти от log odds к вероятностям



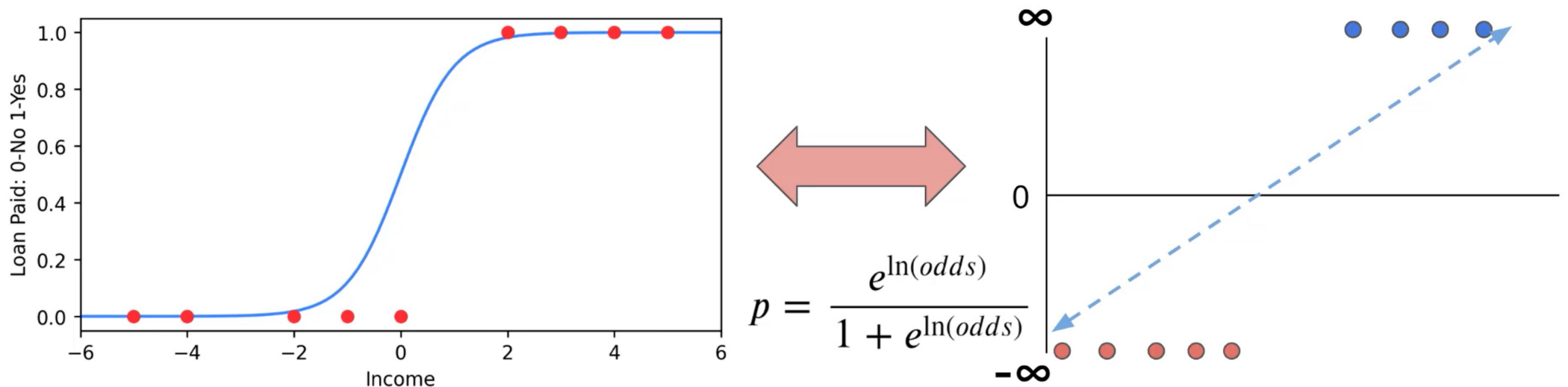
$$p = \frac{e^{\ln(odds)}}{1 + e^{\ln(odds)}}$$





Логистическая регрессия

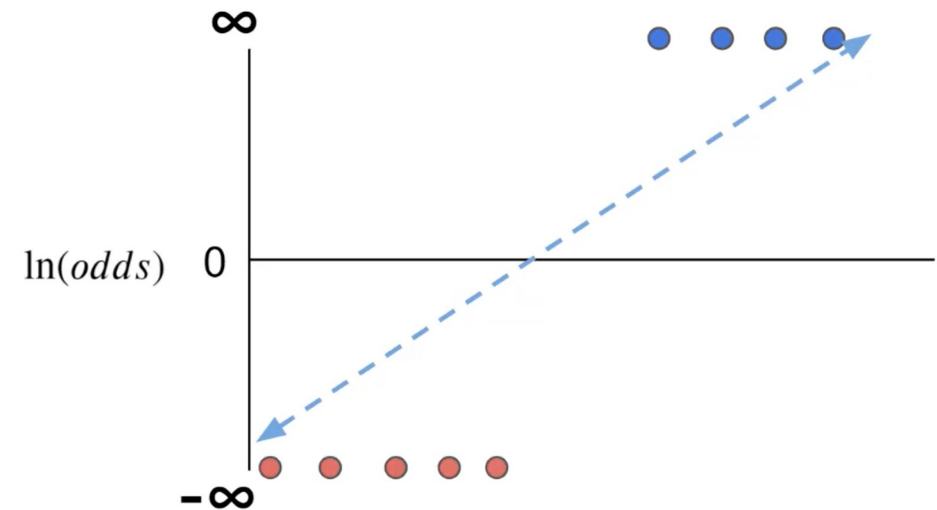
- Далее посмотрим на принцип работы метода максимального правдоподобия





Логистическая регрессия

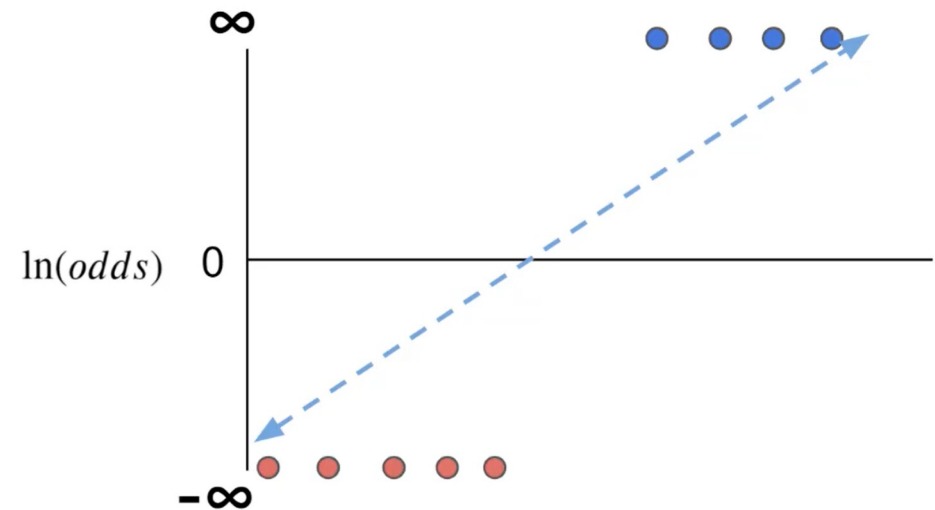
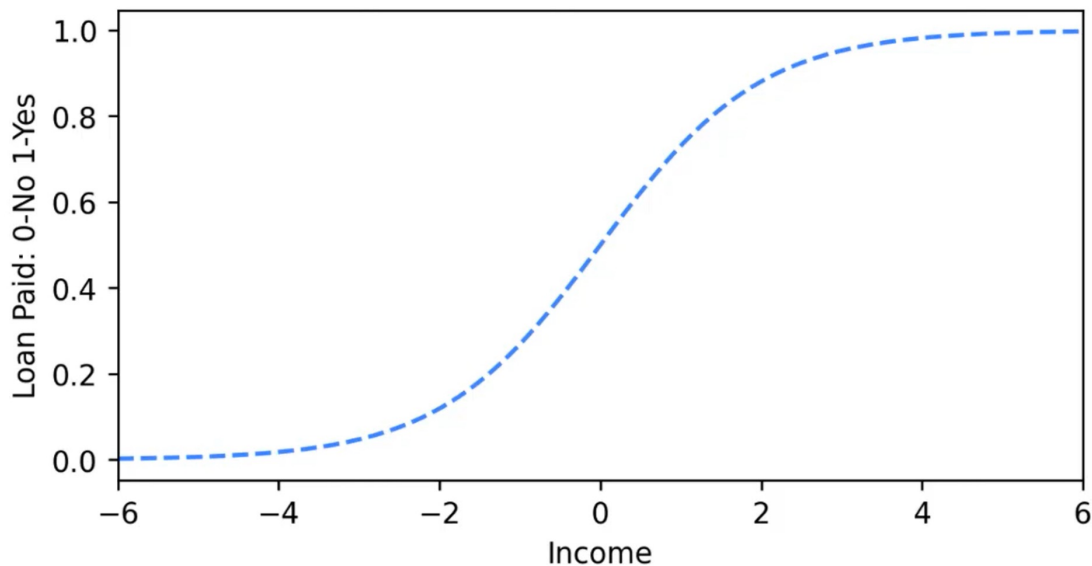
- Мы проводим линию на графике $\log \text{odds}$ и проецируем точки на эту линию





Логистическая регрессия

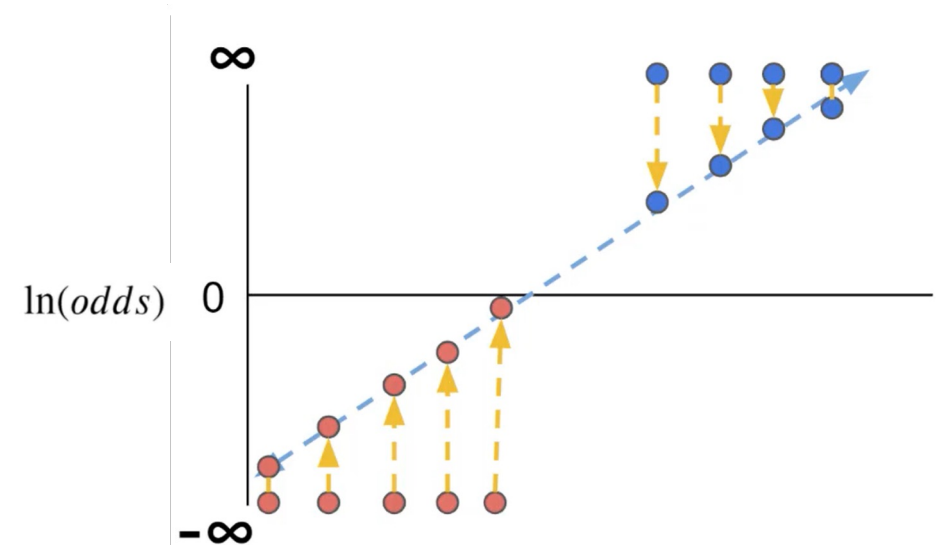
- Этой линии должна соответствовать некоторая кривая на шкале вероятностей





Логистическая регрессия

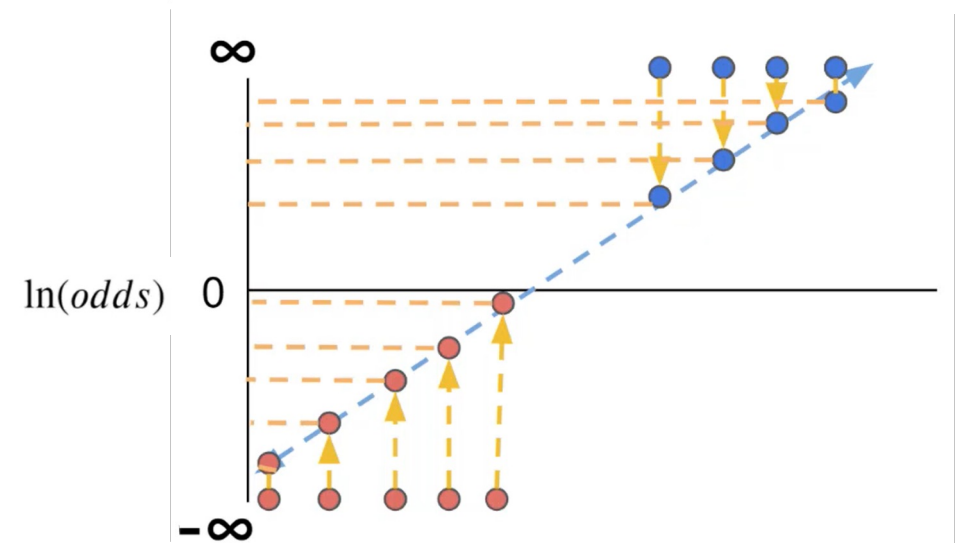
- Проецируем точки на линию





Логистическая регрессия

- Проецируем точки на линию
- Вычисляем соответствующие log odds

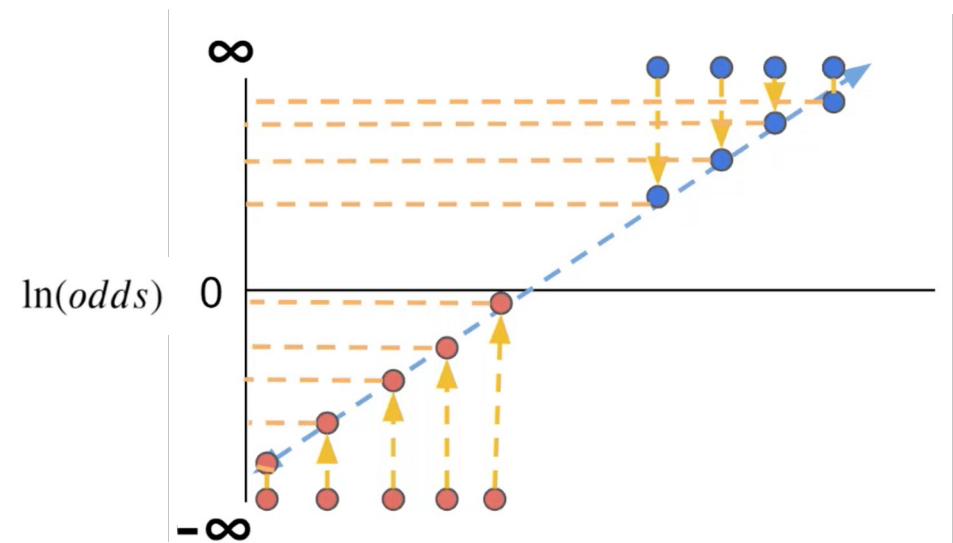




Логистическая регрессия

- Наносим эти значения как вероятности на график

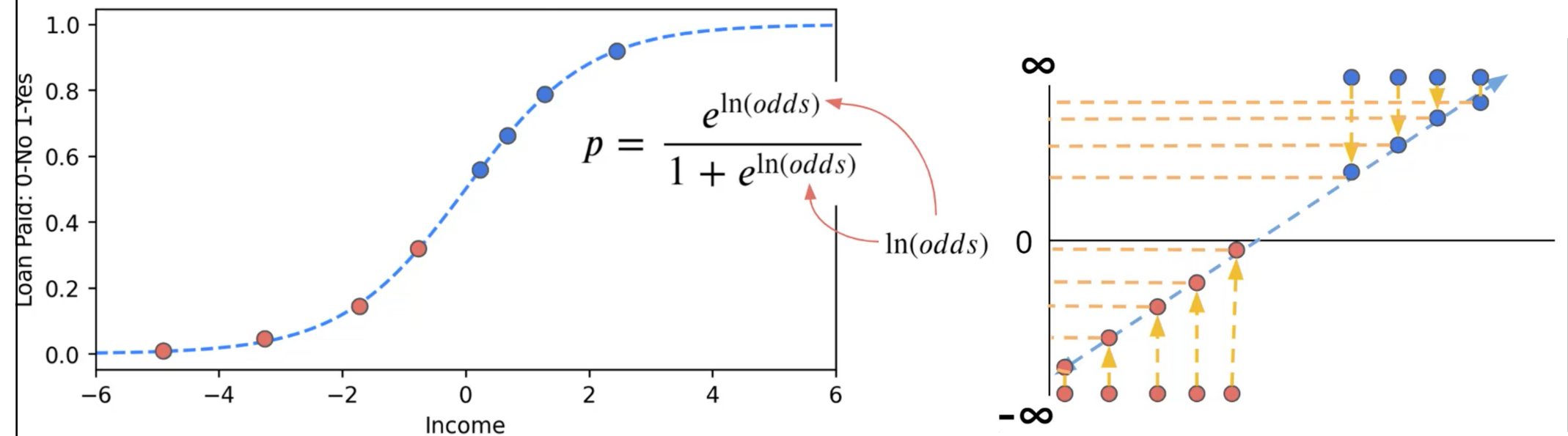
$$p = \frac{e^{\ln(odds)}}{1 + e^{\ln(odds)}}$$





Логистическая регрессия

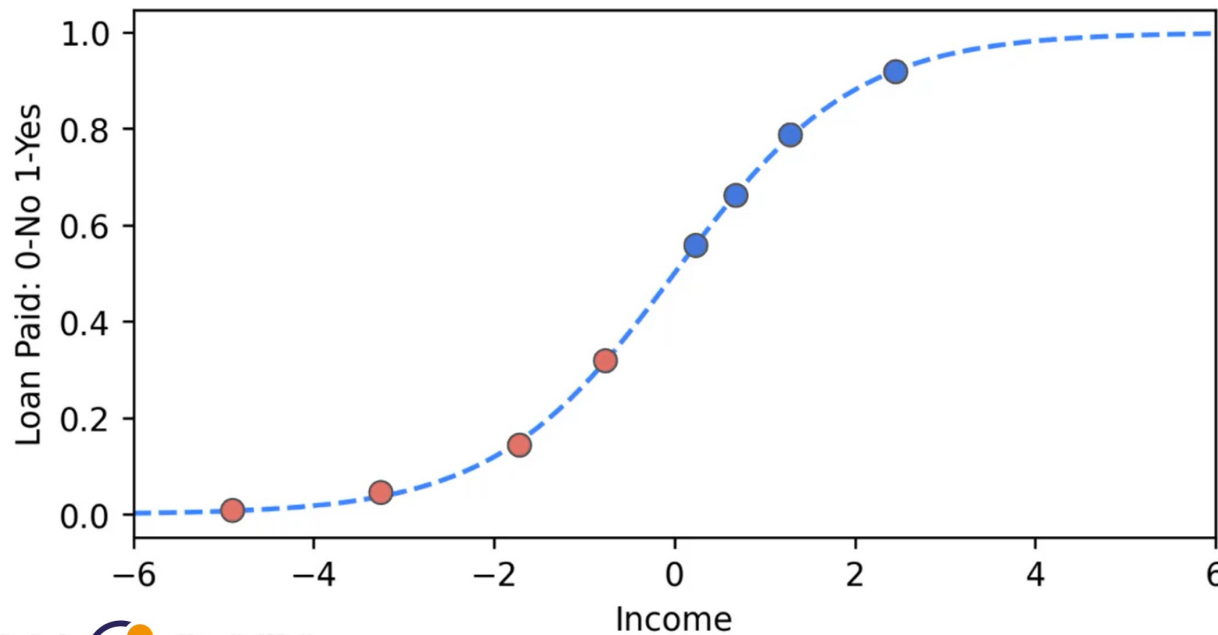
- Наносим эти значения как вероятности на график





Логистическая регрессия

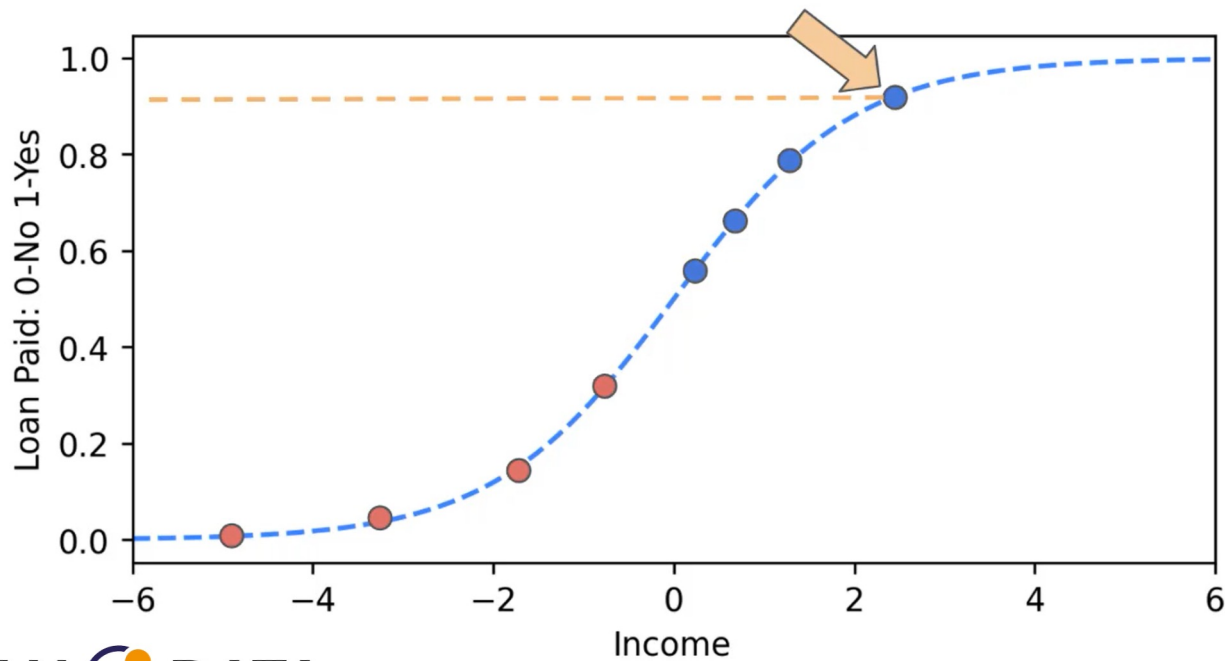
- Далее измеряем правдоподобие этих вероятностей – произведение вероятностей точек класса 1





Логистическая регрессия

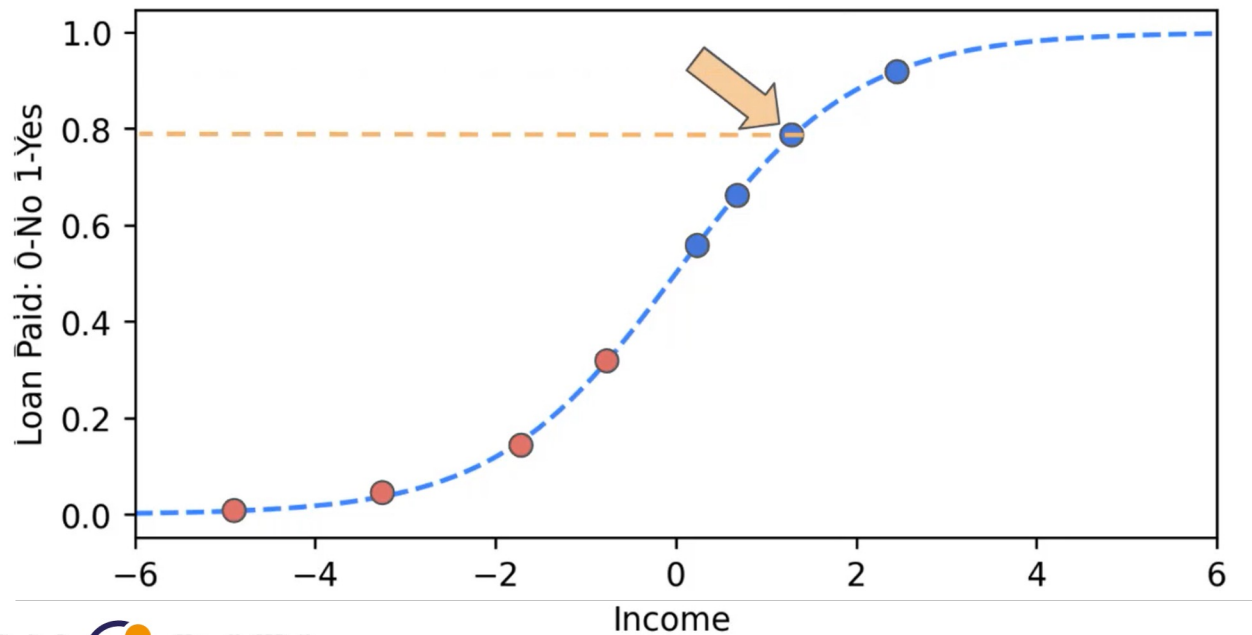
- Правдоподобие = 0.9 * ...





Логистическая регрессия

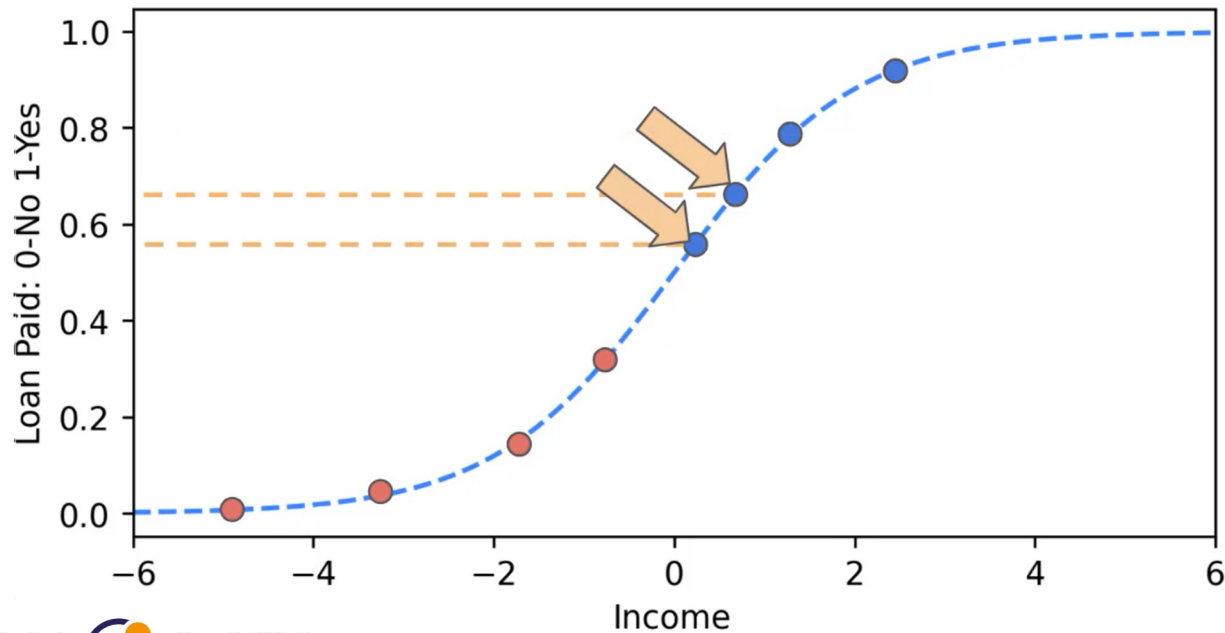
- Правдоподобие = $0.9 * 0.8 * \dots$





Логистическая регрессия

- Правдоподобие = $0.9 * 0.8 * 0.65 * 0.55 \dots$

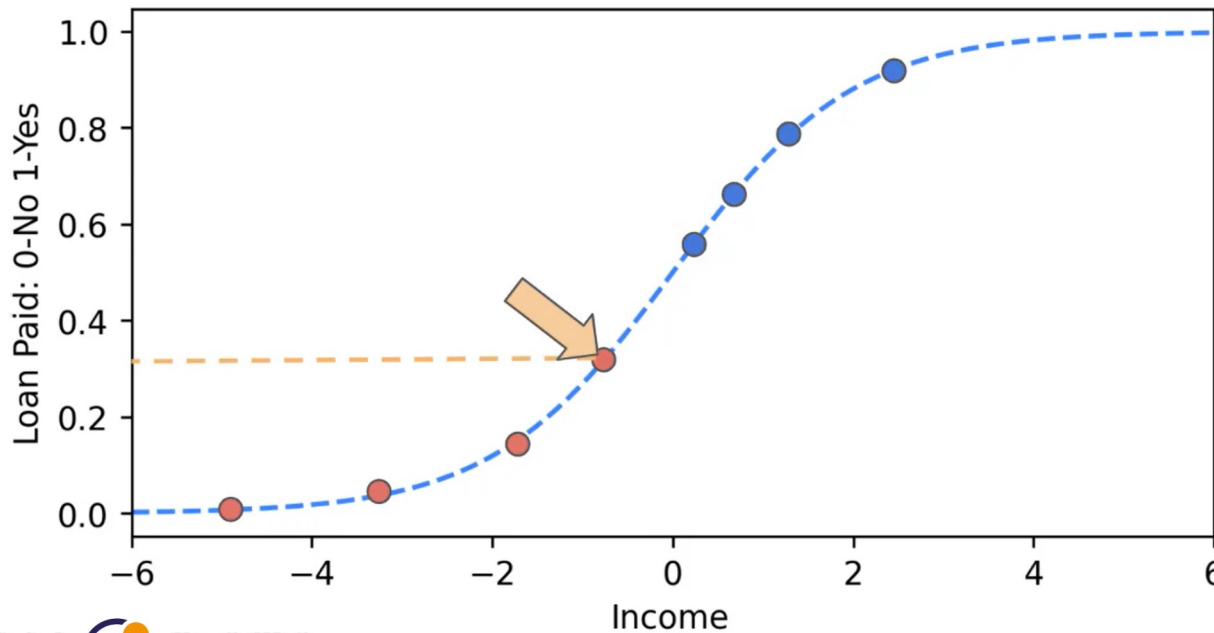




Логистическая регрессия

(1-p)

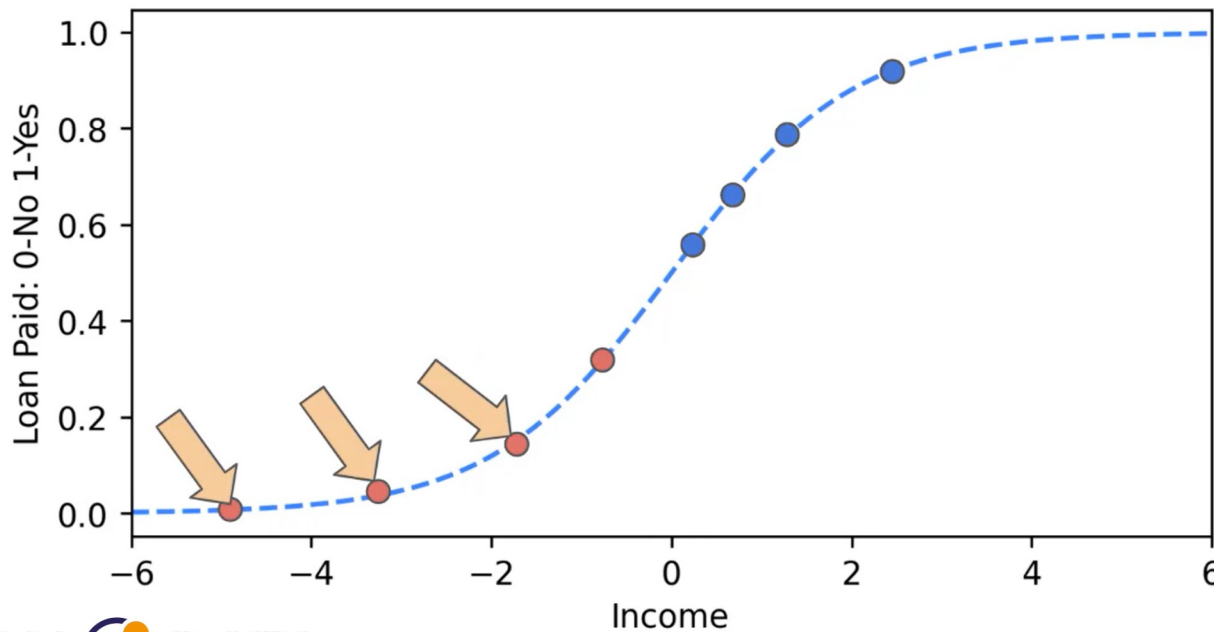
- Правдоподобие = $0.9 * 0.8 * 0.65 * 0.55 * (1-0.3) * \dots$





Логистическая регрессия

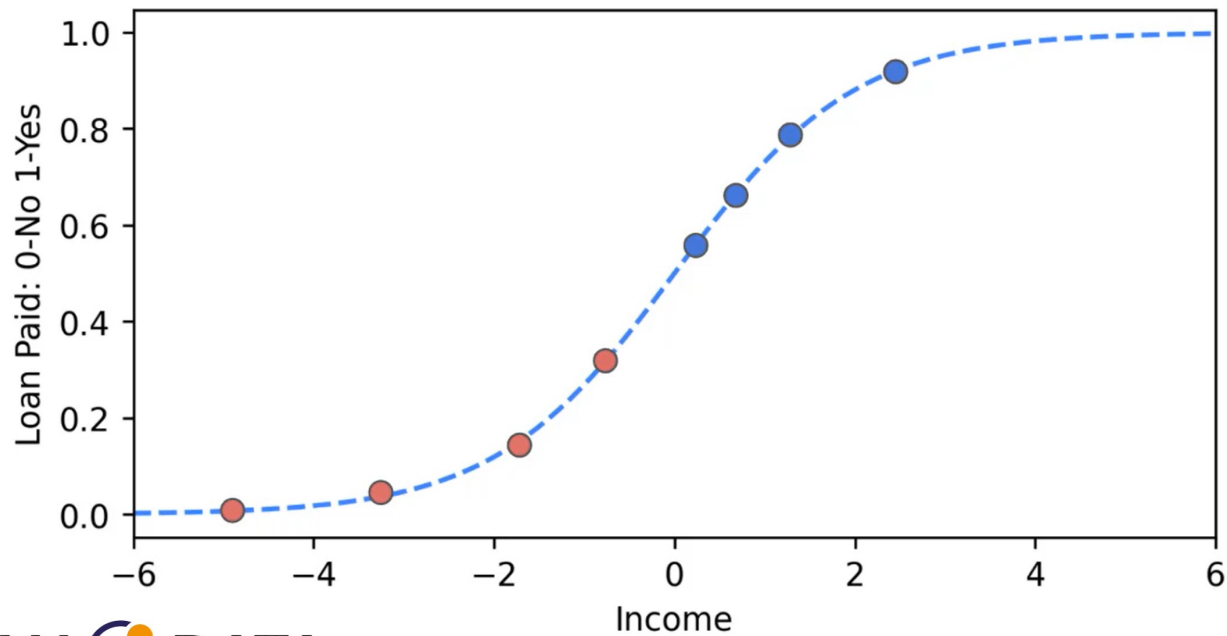
- Правдоподобие = $0.9 * 0.8 * 0.65 * 0.55 * (1-0.3) * (1-0.2) * (1-0.08) * (1-0.02)$





Логистическая регрессия

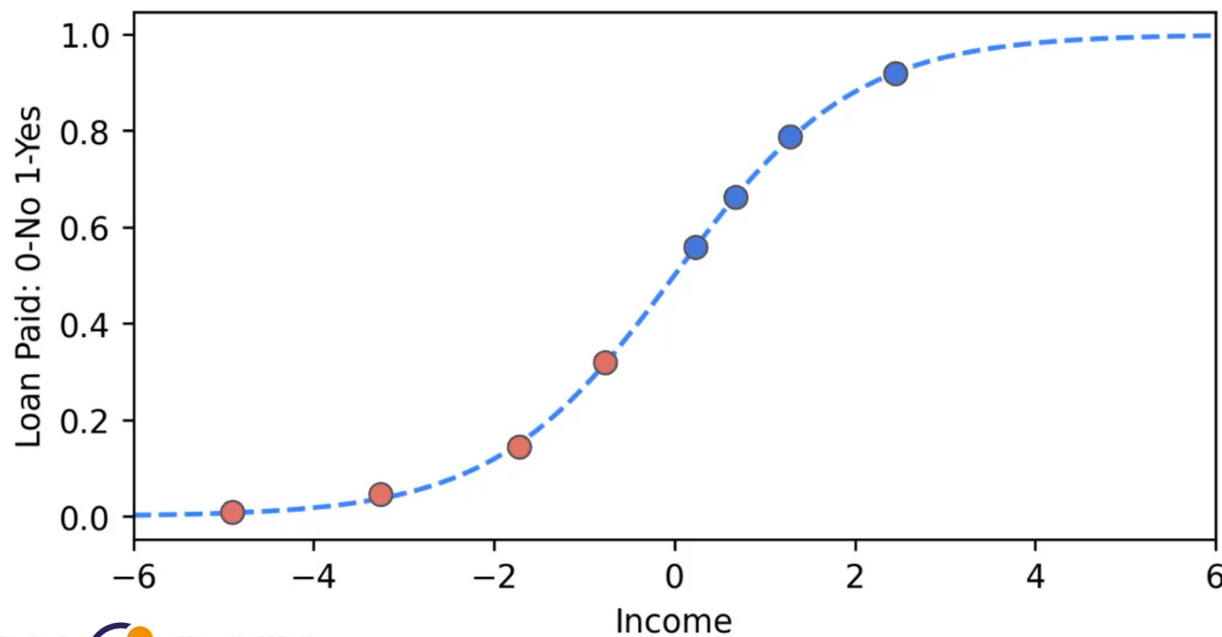
- Правдоподобие = 0.129





Логистическая регрессия

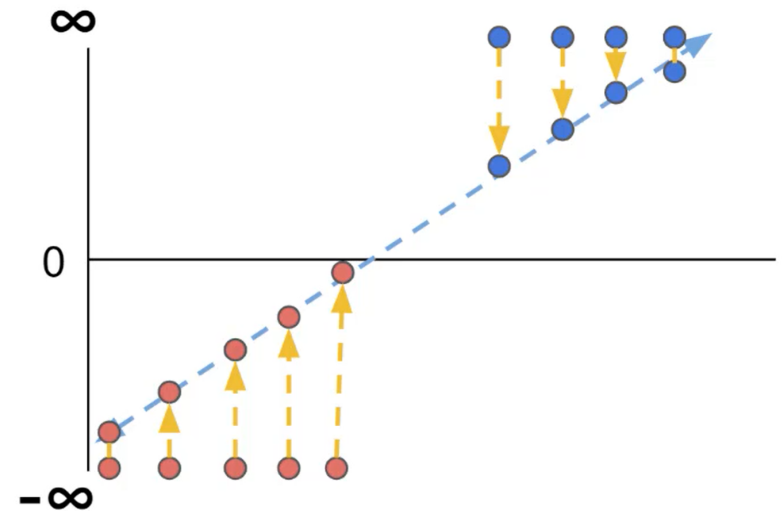
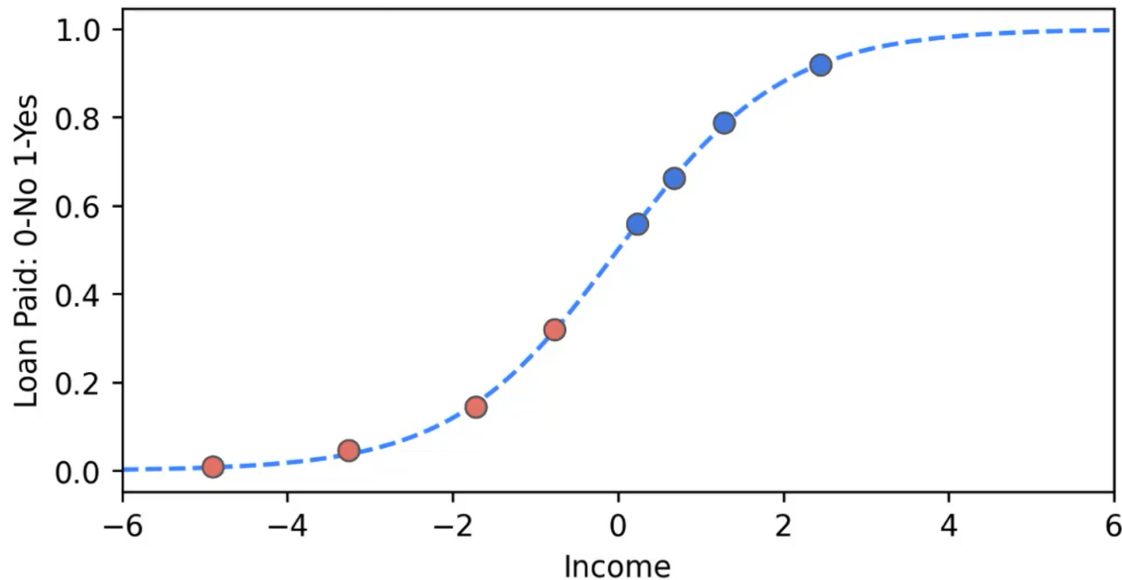
- Правдоподобие = 0.129; В терминах логарифмов мы максимизируем: $\ln(0.9) + \ln(0.8) + \dots$





Логистическая регрессия

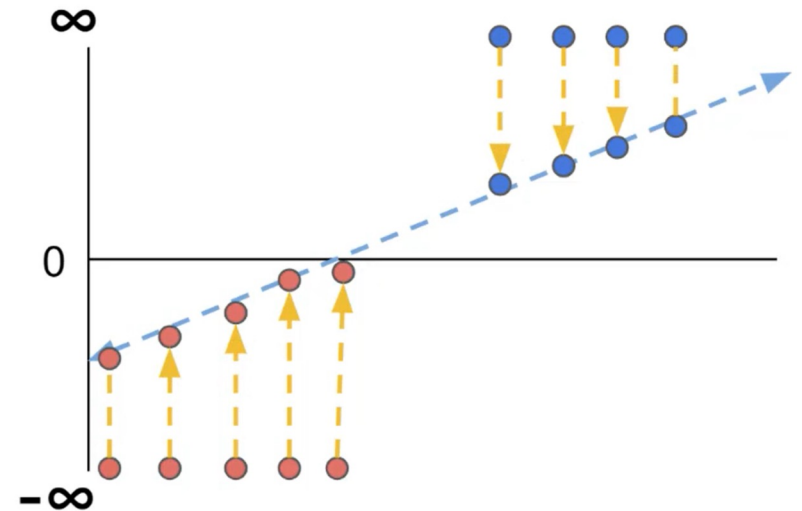
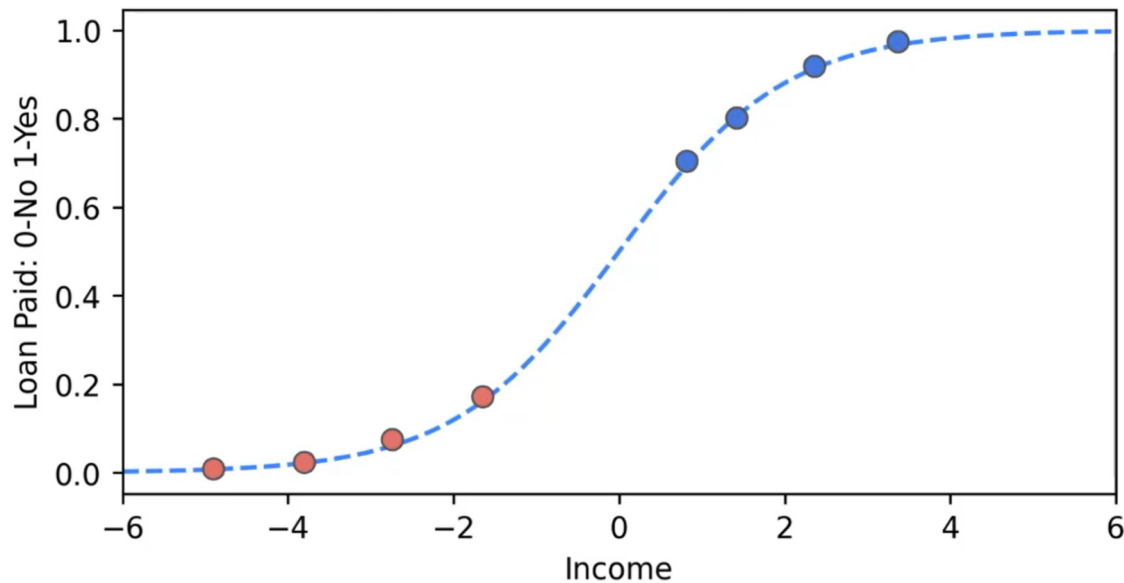
- Мы максимизируем: $\ln(0.9) + \ln(0.8) + \dots$
- Некоторые коэффициенты дадут такой максимум





Логистическая регрессия

- Ищем оптимальные коэффициенты в терминах log odds, которые максимизируют правдоподобие





Логистическая регрессия

- Хотя здесь мы ищем максимум правдоподобия, нам нужно будет что-то минимизировать, потому что градиентный спуск ищет минимумы.



Логистическая регрессия

- В терминах функции стоимости (cost function) мы минимизируем следующее выражение (log loss):

$$J(\mathbf{x}) = -\frac{1}{m} \sum_{j=1}^m y^j \log(\hat{y}^j) + (1 - y^j) \log(1 - \hat{y}^j)$$

$$J(\mathbf{x}) = -\frac{1}{m} \sum_{j=1}^m \left(y^j \log \left(\frac{1}{1 + e^{-\sum_{i=0}^n \beta_i x_i^j}} \right) + (1 - y^j) \log \left(1 - \frac{1}{1 + e^{-\sum_{i=0}^n \beta_i x_i^j}} \right) \right)$$



Логистическая регрессия

- В терминах функции стоимости (cost function) мы минимизируем следующее выражение (log loss):

$$J(\mathbf{x}) = -\frac{1}{m} \sum_{j=1}^m y^j \log(\hat{y}^j) + (1 - y^j) \log(1 - \hat{y}^j)$$

$$J(\mathbf{x}) = -\frac{1}{m} \sum_{j=1}^m \left(y^j \log \left(\frac{1}{1 + e^{-\sum_{i=0}^n \beta_i x_i^j}} \right) + (1 - y^j) \log \left(1 - \frac{1}{1 + e^{-\sum_{i=0}^n \beta_i x_i^j}} \right) \right)$$

Как и для линейной регрессии, для решения применяем градиентный спуск



Логистическая регрессия

- Не стоит волноваться по поводу градиентного спуска – на практике мы не будем реализовывать его самостоятельно
- Главное, что нам нужно здесь – это понять связь между log odds и вероятностями



Логистическая регрессия

- Далее мы перейдём к применению логистической регрессии в Python!



Логистическая регрессия в Scikit-Learn

Часть 1: исследование данных



Логистическая регрессия в Scikit-Learn

Часть 2: создание и обучение модели



Метрики оценки модели классификации

Часть 1: матрица ошибок



Метрики оценки модели классификации

- Наверняка Вы слышали про такие метрики, как “ложно-положительные” (false positive) и “ложно-отрицательные” (false negative).
А также такие метрики, как “accuracy”.
Что же означают эти метрики?



Метрики оценки модели классификации

- Представьте, что модель должна предсказывать наличие/отсутствие заболевания на основе некоторых биологических признаков.
- Это можно рассмотреть как задачу логистической регрессии, в которой:
 - 0 – заболевания нет (тест отрицательный)
 - 1 – заболевание есть (тест положительный)



Метрики оценки модели классификации

- Однако вряд ли модель будет работать идеально во всех случаях. Поэтому возможны 4 варианта:
 - Заболевший человек – тест положительный
 - Здоровый человек – тест отрицательный



Метрики оценки модели классификации

- Однако вряд ли модель будет работать идеально во всех случаях. Поэтому возможны 4 варианта:
 - Заболевший человек – тест положительный
 - Здоровый человек – тест отрицательный
 - Заболевший человек – тест отрицательный
 - Здоровый человек – тест положительный



Метрики оценки модели классификации

- На основе этих 4-х случаев мы можем рассчитать некоторые метрики.
- Для начала расположим эти 4 случая в виде матрицы



Метрики оценки модели классификации

- Матрица ошибок (confusion matrix)

		ACTUAL	
		INFECTED	HEALTHY
	PREDICTED		



Метрики оценки модели классификации

- Матрица ошибок (confusion matrix)

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED		
	HEALTHY		



Метрики оценки модели классификации

- Матрица ошибок (confusion matrix)

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	TRUE POSITIVE	
	HEALTHY		



Метрики оценки модели классификации

- Матрица ошибок (confusion matrix)

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	TRUE POSITIVE	
	HEALTHY		TRUE NEGATIVE



Метрики оценки модели классификации

- Матрица ошибок (confusion matrix)

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	TRUE POSITIVE	FALSE POSITIVE
	HEALTHY		TRUE NEGATIVE



Метрики оценки модели классификации

- Матрица ошибок (confusion matrix)

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	TRUE POSITIVE	FALSE POSITIVE
	HEALTHY	FALSE NEGATIVE	TRUE NEGATIVE



Метрики оценки модели классификации

- Представим группу из 100 человек.

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED		
	HEALTHY		



Метрики оценки модели классификации

- Представим группу из 100 человек.
- Из них 5 заболевших и 95 здоровых.

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED		
	HEALTHY		



Метрики оценки модели классификации

- Мы провели тесты и получили такие результаты:

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93



Метрики оценки модели классификации

- Ассурасу – доля правильно классифицированных объектов.

$$\text{Acc} = (TP+TN)/\text{Total}$$

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93



Метрики оценки модели классификации

- Ассурасу – доля правильно классифицированных объектов.

$$\text{Acc} = (TP+TN)/\text{Total}$$

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

$$(4+93)/100 = 97\% \text{ Accuracy}$$



Метрики оценки модели классификации

- Ассурасу – доля правильно классифицированных объектов.

$$\text{Acc} = (TP+TN)/\text{Total}$$

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

Это хорошее значение?

$$(4+93)/100 = 97\% \text{ Accuracy}$$



Метрики оценки модели классификации

- Парадокс ассигасы.... представьте, что наша модель всегда говорит, что пациент здоров.

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	0	0
	HEALTHY	5	95



Метрики оценки модели классификации

- Парадокс ассурасу.... представьте, что наша модель всегда говорит, что пациент здоров.

ACTUAL

$$\text{Acc} = (\text{TP} + \text{TN}) / \text{Total}$$

PREDICTED	ACTUAL	
	INFECTED	HEALTHY
INFECTED	0	0
HEALTHY	5	95

$$(0 + 95) / 100 = 95\% \text{ Accuracy}$$



Метрики оценки модели классификации

- Вы можете подумать: ну, это здесь такие данные! В реальной жизни, если модель всегда возвращает один и тот же результат, то эффективность модели должна получиться низкой.



Метрики оценки модели классификации

- В этом и состоит парадокс accuracy!
 - Если классы несбалансированы, то любой классификатор столкнётся с проблемой парадокса accuracy
 - Несбалансированные классы всегда приводят к искажённым (завышенным) результатам



Метрики оценки модели классификации

- В этом и состоит парадокс accuracy!
 - Если классы несбалансированы, то любой классификатор столкнётся с проблемой парадокса accuracy
 - Несбалансированные классы всегда приводят к искажённым (завышенным) результатам
 - Несбалансированные классы встречаются очень часто!
 - Заболевает небольшой процент населения
 - Фрод кредитных карт для небольшого процента карт



Метрики оценки модели классификации

- Если один из двух классов имеет небольшой процент $n\%$, то модель, предсказывающая “основной” класс, будет иметь ассигасу $(1-n)$
- В нашем предыдущем примере было только 5% заболевших.
- Поэтому ассигасу модели “все здоровы” была 95%.



Метрики оценки модели классификации

- Это значит, что полагаться только на метрику accuracy (доля правильно классифицированных объектов) нам нельзя.
- Ещё понадобятся другие метрики:
 - Precision (точность)
 - Recall (полнота)
 - F1-score (F1-оценка или F1-мера)
- Рассмотрим эти метрики в следующей лекции



Метрики оценки модели классификации

Часть 2: Precision и Recall



Метрики оценки модели классификации

- Мы посмотрели метрику accuracy (доля правильно классифицированных объектов) и связанный с ней парадокс.
- Рассмотрим теперь следующие метрики – они позволят нам лучше понять эффективность модели:
 - Recall (полнота; aka sensitivity - чувствительность)
 - Precision (точность)
 - F1-score (F1-оценка или F1-мера)



Метрики оценки модели классификации

- Начнём с метрики Recall (полнота): когда фактический результат actual положителен, как часто модель права?

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

$(TP) / \text{Total Actual Positives}$



Метрики оценки модели классификации

- Начнём с метрики Recall (полнота): когда фактический результат actual положителен, как часто модель права?

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

$$\text{Recall} = (4)/5 = 0.8$$

(TP)/Total Actual
Positives



Метрики оценки модели классификации

- Чему будет равен Recall, если модель будет всегда отвечать “пациент здоров”?

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	0	0
	HEALTHY	5	95

(TP)/Total Actual
Positives



Метрики оценки модели классификации

- Чему будет равен Recall, если модель будет всегда отвечать “пациент здоров”?

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	0	0
	HEALTHY	5	95

(TP)/Total Actual
Positives

$$\text{Recall} = (0)/5 !$$

Recall = 0
Повод
насторожиться!



Метрики оценки модели классификации

- Далее рассмотрим Precision (точность): когда предсказание положительно, как часто модель права?

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

(TP)/Total Predicted
Positives



Метрики оценки модели классификации

- Далее рассмотрим Precision (точность): когда предсказание положительно, как часто модель права?

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

(TP)/Total Predicted Positives

$$\text{Precision} = (\text{TP})/6 = (4)/6 = 0.66$$



Метрики оценки модели классификации

- Чему будет равен Recall, если модель будет всегда отвечать “пациент здоров”?

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	0	0
	HEALTHY	5	95

(TP)/Total Predicted
Positives



Метрики оценки модели классификации

- Чему будет равен Recall, если модель будет всегда отвечать “пациент здоров”?

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	0	0
	HEALTHY	5	95

(TP)/Total Predicted
Positives

Precision = 0/0



Метрики оценки модели классификации

- Метрики Precision и Recall помогают оценить работу модели именно для положительных случаев.
- В зависимости от модели, очень часто между этими двумя метриками нужно искать некоторый баланс. Мы рассмотрим это позже, когда будем говорить о ROC-кривых.



Метрики оценки модели классификации

- Поскольку метрики Precision и Recall связаны между собой общим числителем (TP), мы можем также посчитать F1-score по следующей формуле гармонического среднего:

$$F = \frac{2 \times \textit{precision} \times \textit{recall}}{\textit{precision} + \textit{recall}}$$



Метрики оценки модели классификации

- Поскольку метрики Precision и Recall связаны между собой общим числителем (TP), мы можем также посчитать F1-score по следующей формуле гармонического среднего:

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

Равно нулю, если хотя бы одно из слагаемых нулевое.



Метрики оценки модели классификации

- И последнее замечание по матрице ошибок – существует множество других метрик:

		True condition			
<u>Total population</u>		Condition positive	Condition negative	$\text{Prevalence} = \frac{\Sigma \text{Condition positive}}{\Sigma \text{Total population}}$	$\text{Accuracy (ACC)} = \frac{\Sigma \text{True positive} + \Sigma \text{True negative}}{\Sigma \text{Total population}}$
Predicted condition	Predicted condition positive	True positive	False positive, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{True positive}}{\Sigma \text{Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{False positive}}{\Sigma \text{Predicted condition positive}}$
	Predicted condition negative	False negative, Type II error	True negative	False omission rate (FOR) = $\frac{\Sigma \text{False negative}}{\Sigma \text{Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{True negative}}{\Sigma \text{Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection, $\text{Power} = \frac{\Sigma \text{True positive}}{\Sigma \text{Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{False positive}}{\Sigma \text{Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR-}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{False negative}}{\Sigma \text{Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{True negative}}{\Sigma \text{Condition negative}}$	Negative likelihood ratio (LR-) = $\frac{\text{FNR}}{\text{TNR}}$	
				$F_1 \text{ score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$	



Метрики оценки модели классификации

- Далее мы рассмотрим способы визуализировать связи между метриками (такими как precision, recall) с помощью графиков.



Метрики оценки модели классификации

Часть 3: ROC-кривые



Метрики оценки модели классификации

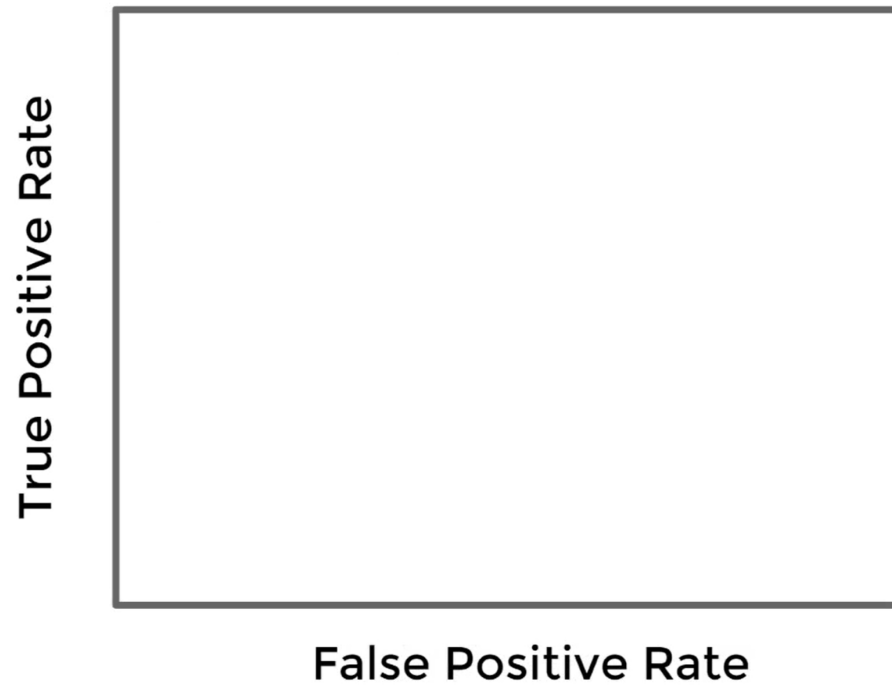
- Во время второй мировой войны получили развитие радары для обнаружения самолётов.
- Технология была новой, и нужна была методика определения эффективности радаров.





Метрики оценки модели классификации

- Для оценки радаров была придумана ROC-кривая (ROC = Receiver Operator Characteristic).





Метрики оценки модели классификации

- Для оценки радаров была придумана ROC-кривая (ROC = Receiver Operator Characteristic).





Метрики оценки модели классификации

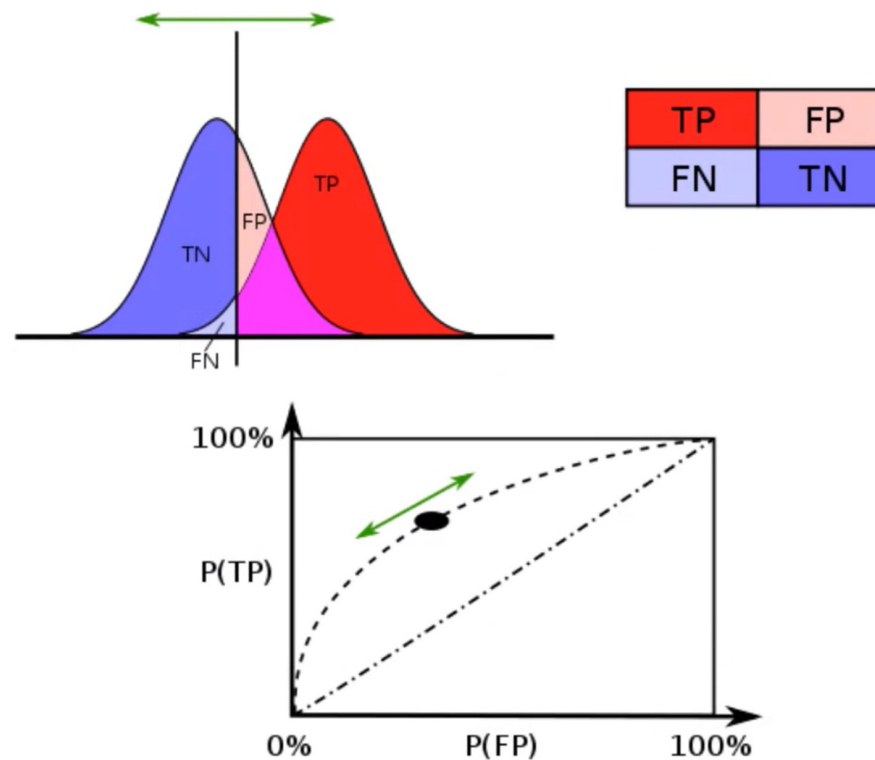
- Для оценки радаров была придумана ROC-кривая (ROC = Receiver Operator Characteristic).





Метрики оценки модели классификации

- Между истинно-положительными и ложно-положительными точками может быть компромисс

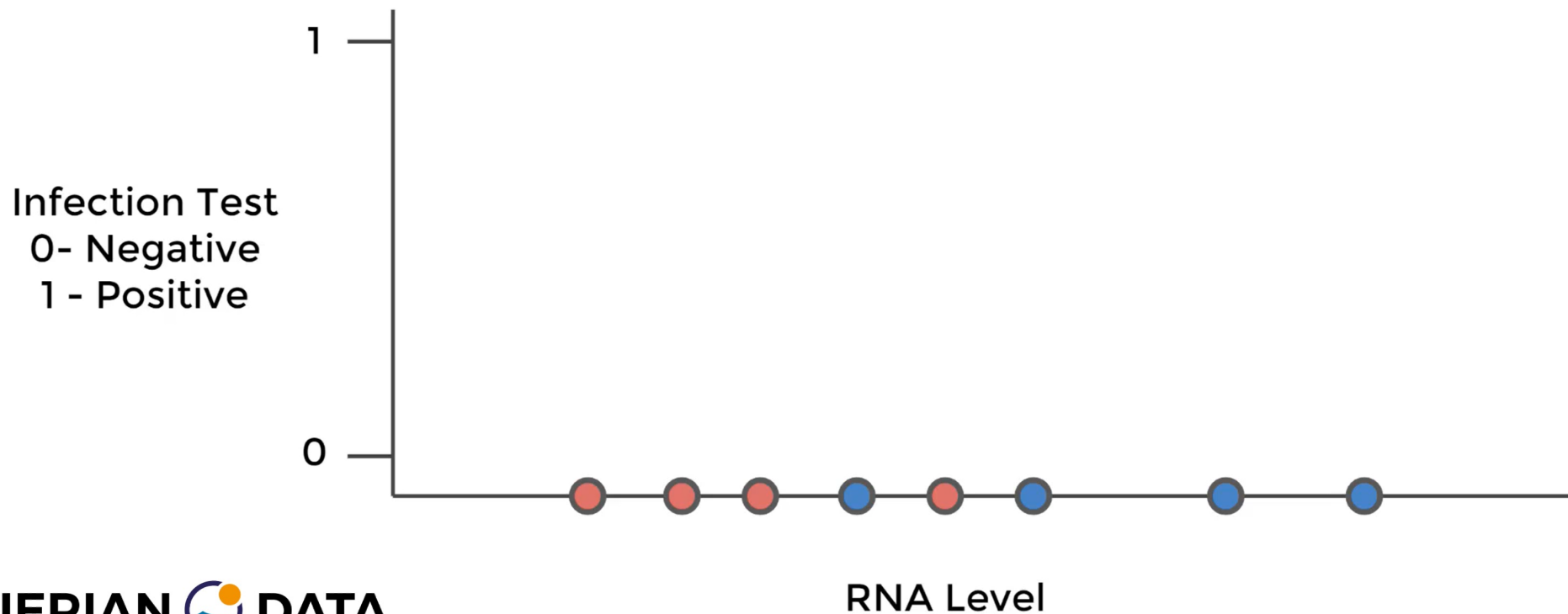




Метрики оценки модели классификации

- Вернёмся к нашему примеру о том, здоров ли пациент

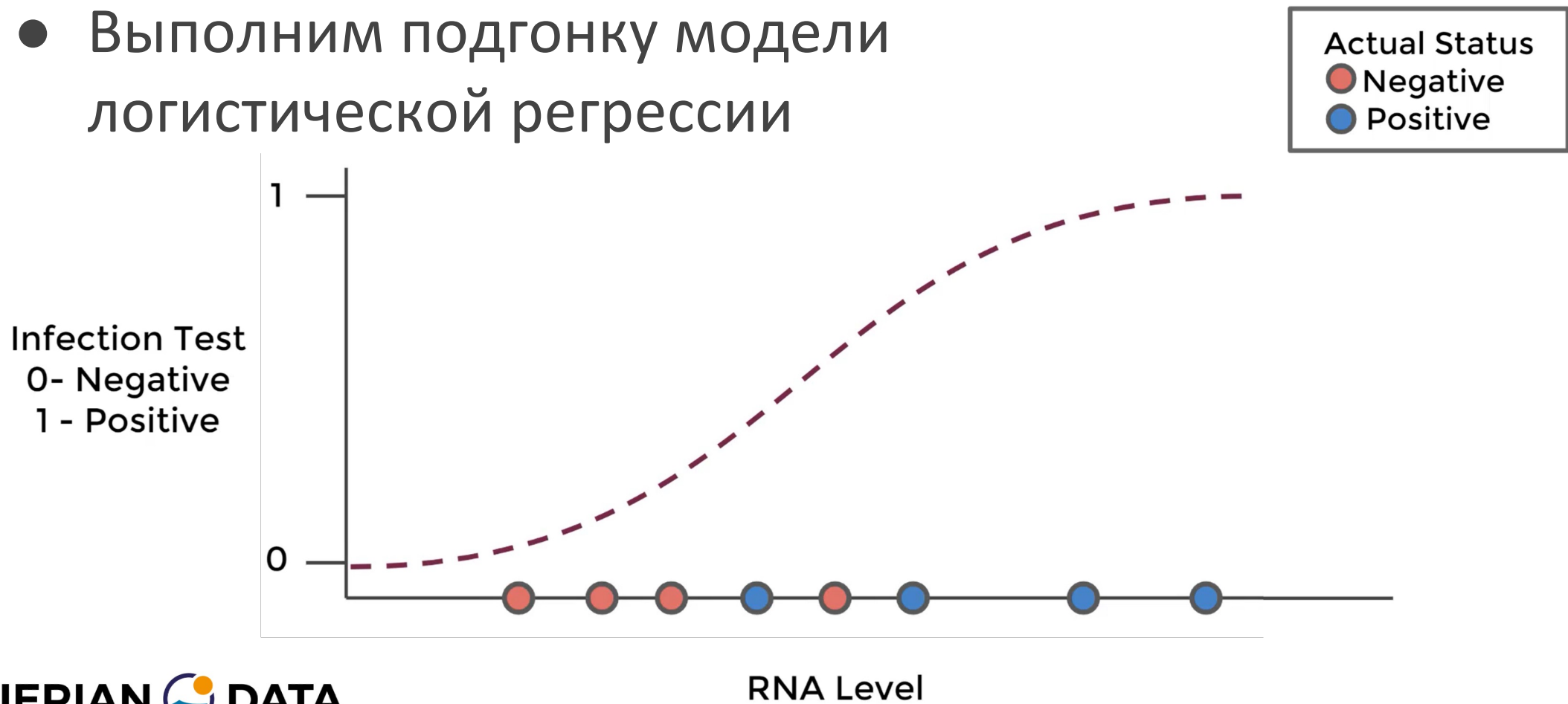
Actual Status
● Negative
● Positive





Метрики оценки модели классификации

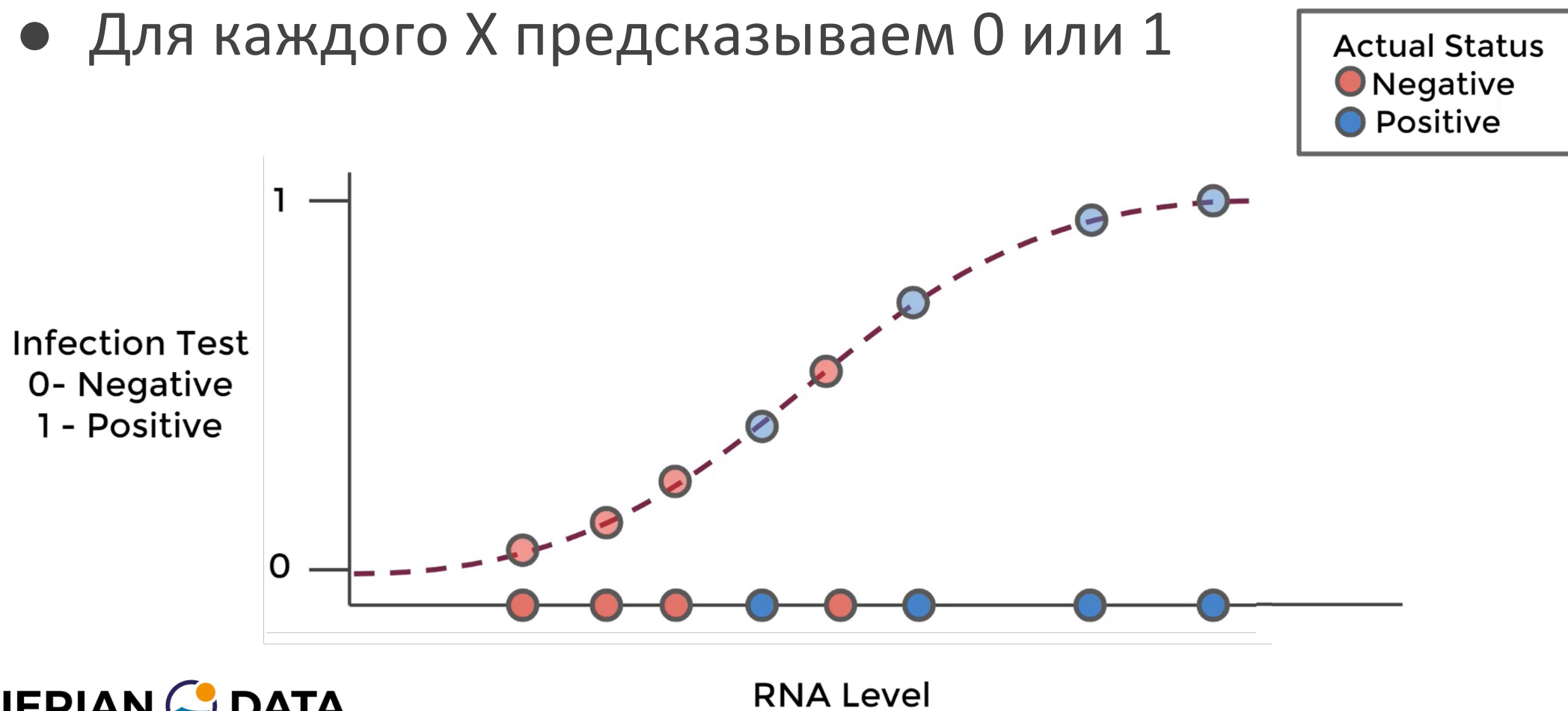
- Выполним подгонку модели логистической регрессии





Метрики оценки модели классификации

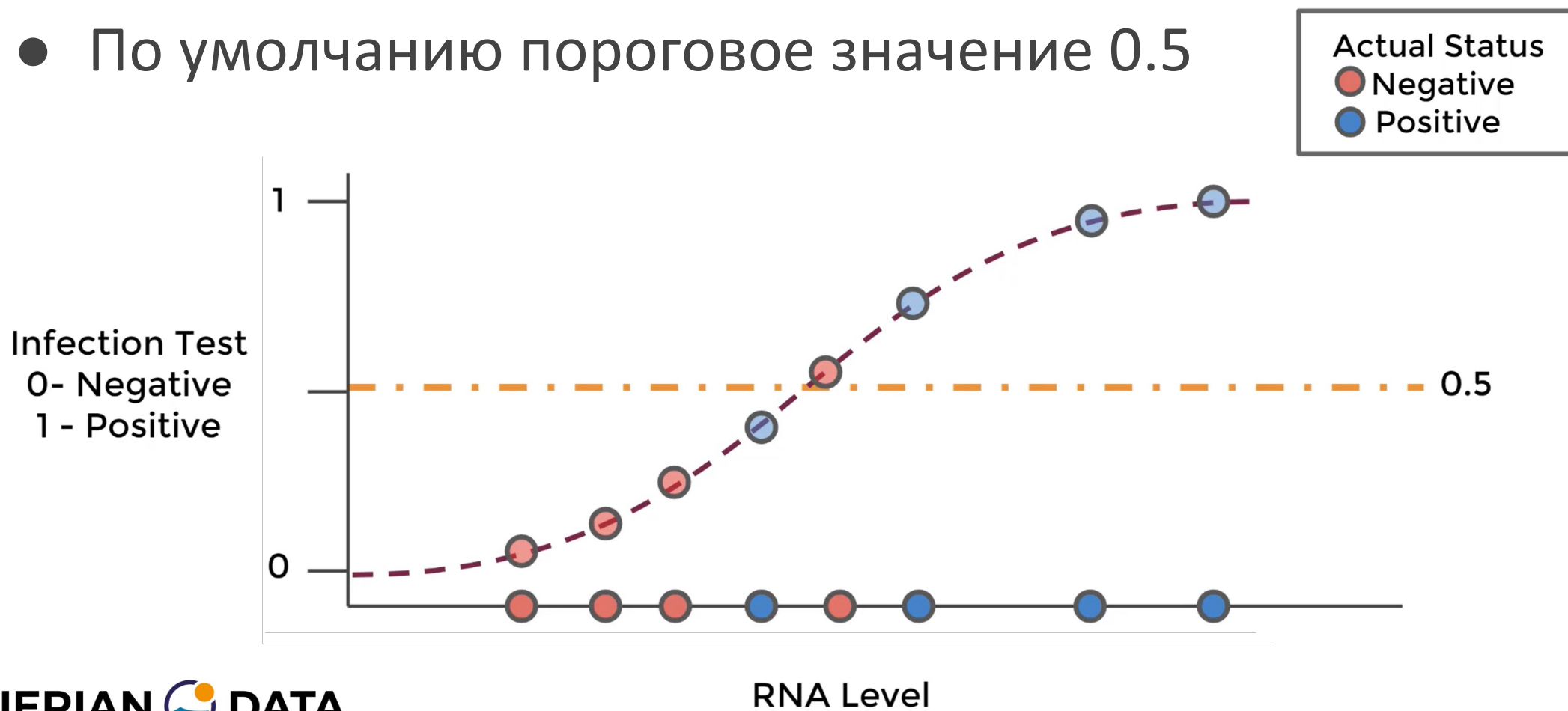
- Для каждого X предсказываем 0 или 1





Метрики оценки модели классификации

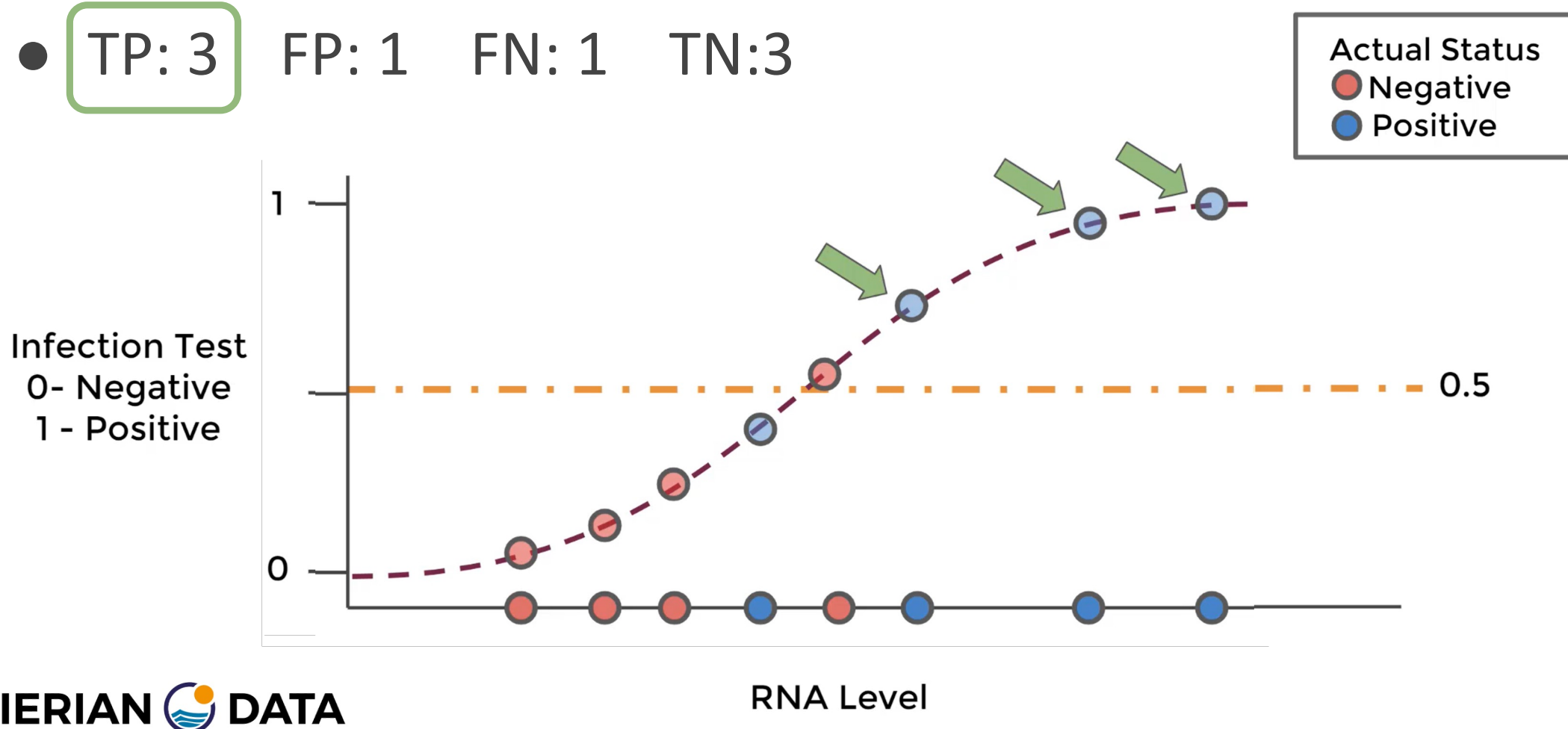
- По умолчанию пороговое значение 0.5





Метрики оценки модели классификации

• TP: 3 FP: 1 FN: 1 TN: 3



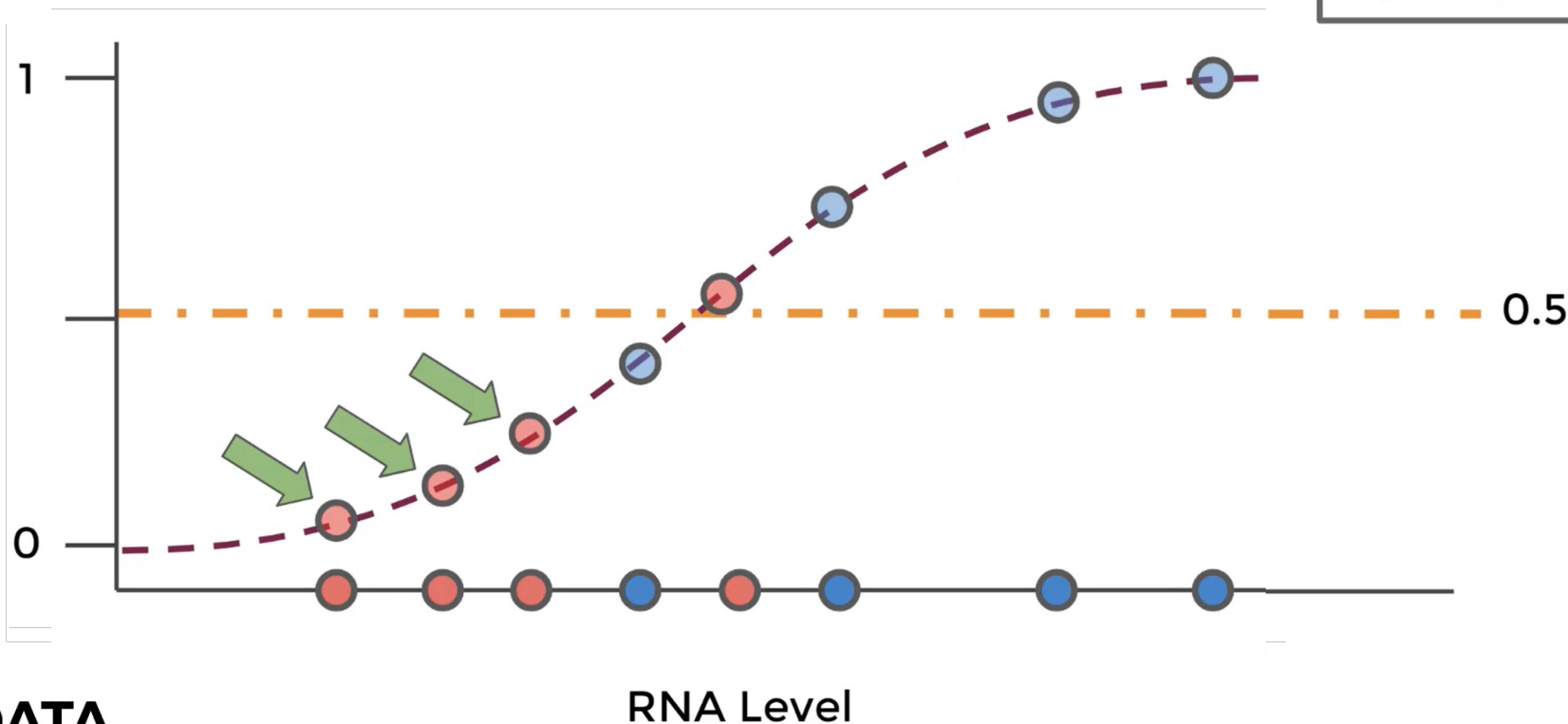


Метрики оценки модели классификации

● TP: 3 FP: 1 FN: 1 **TN:3**

Actual Status
● Negative
● Positive

Infection Test
0 - Negative
1 - Positive



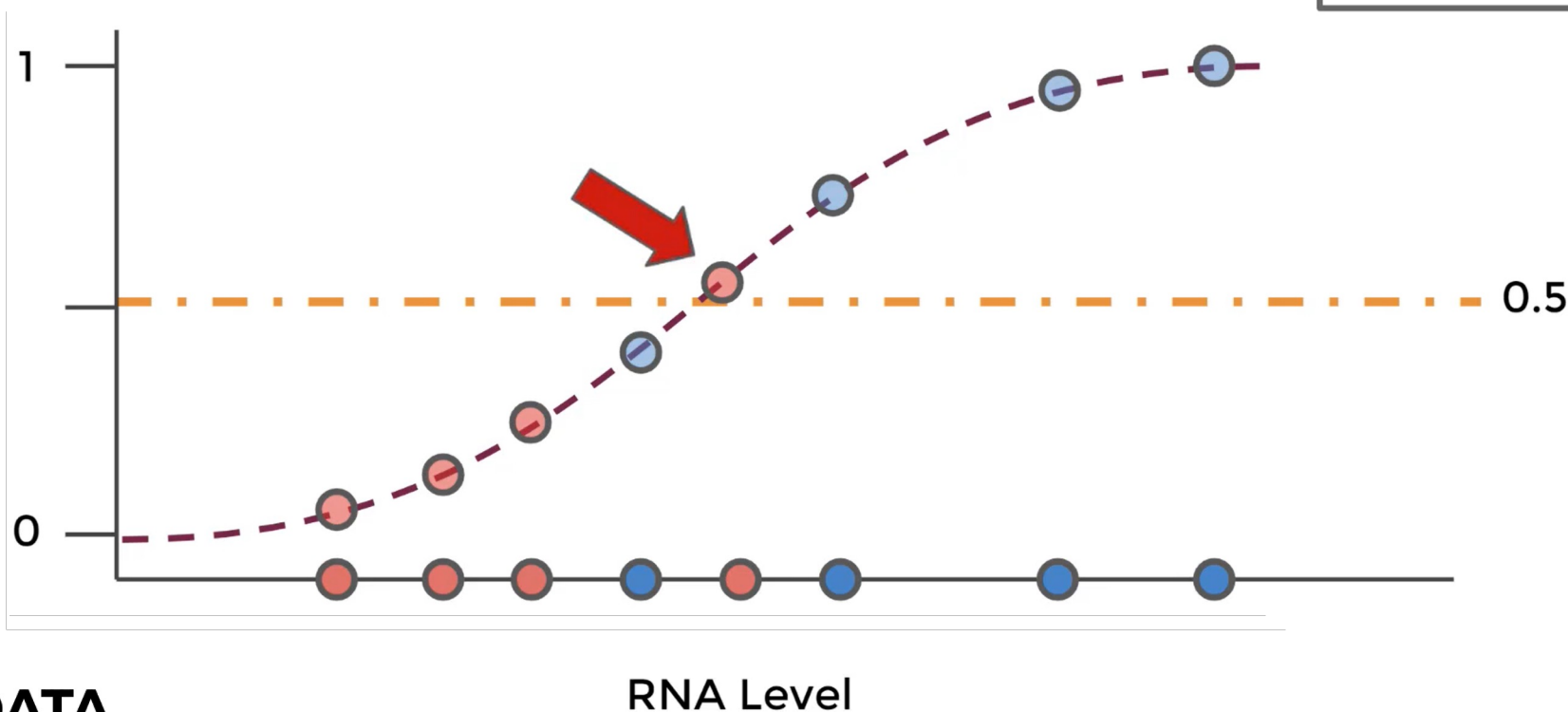


Метрики оценки модели классификации

● TP: 3 **FP: 1** FN: 1 TN: 3

Actual Status
● Negative
● Positive

Infection Test
0 - Negative
1 - Positive



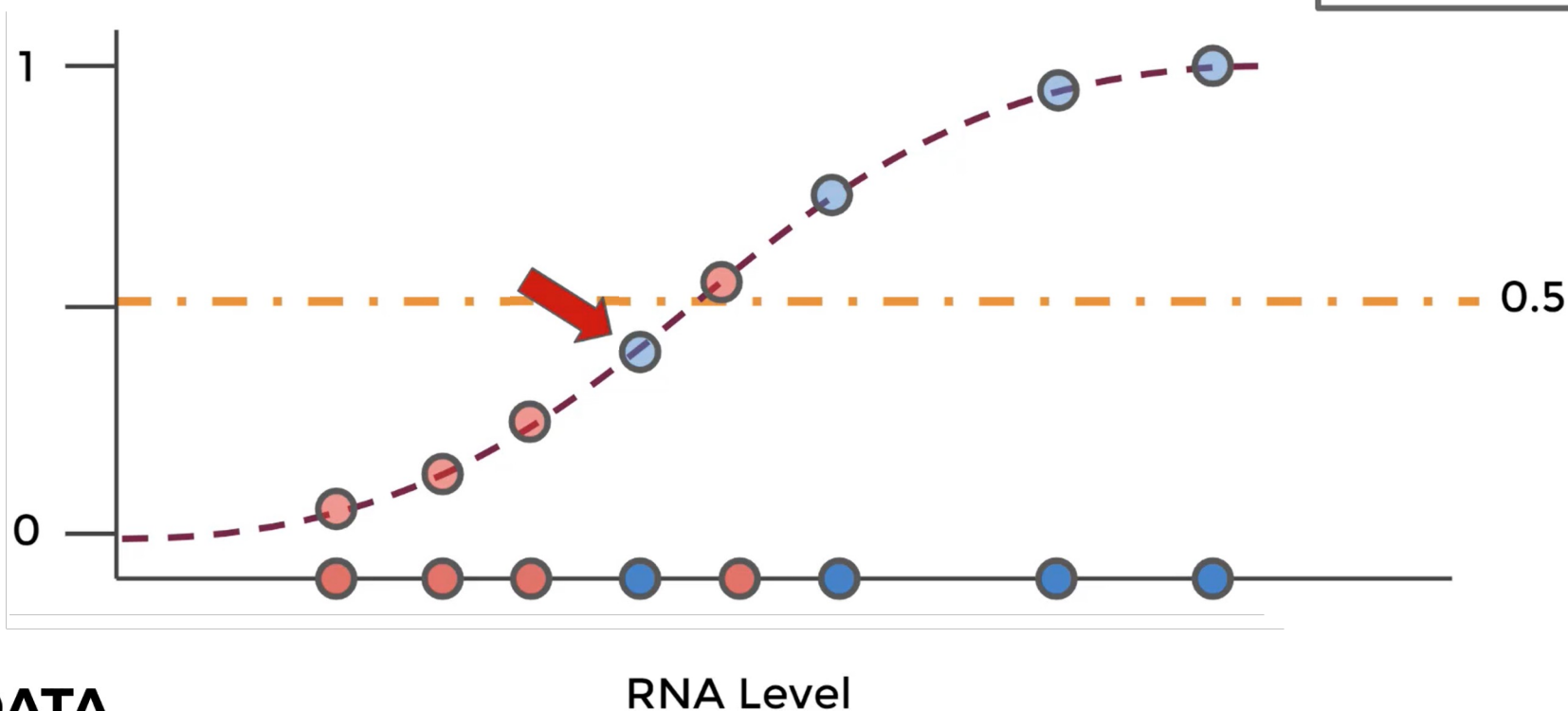


Метрики оценки модели классификации

● TP: 3 FP: 1 **FN: 1** TN: 3

Actual Status
● Negative
● Positive

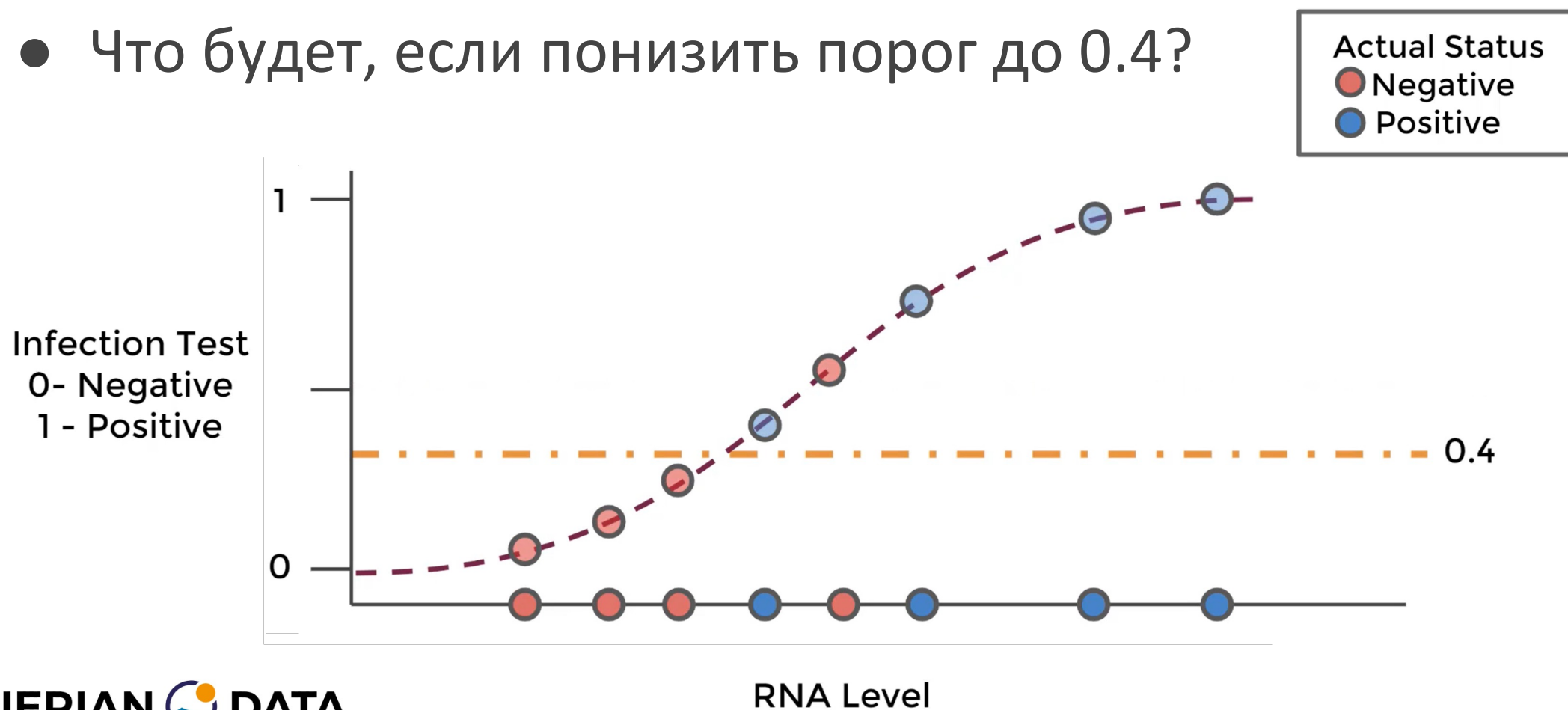
Infection Test
0 - Negative
1 - Positive





Метрики оценки модели классификации

- Что будет, если понизить порог до 0.4?



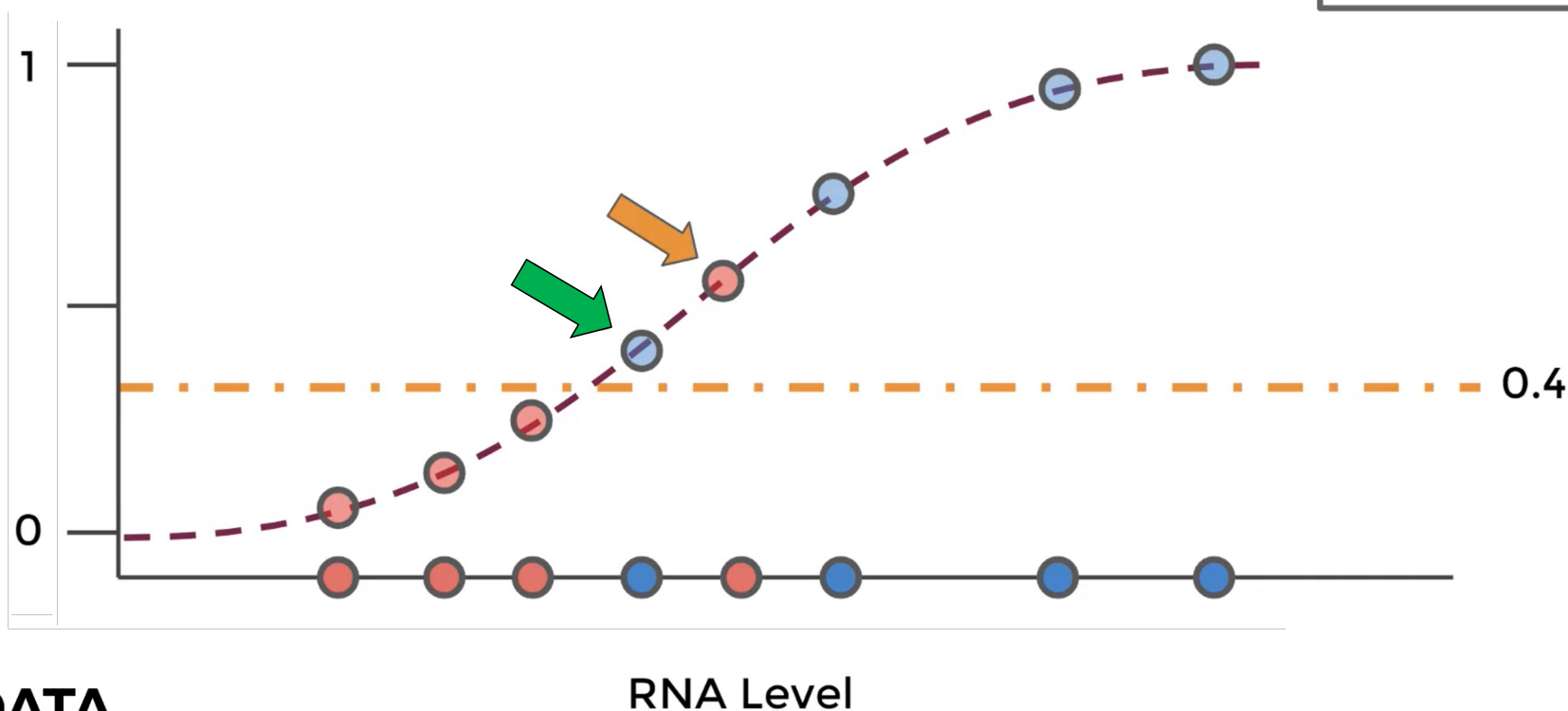


Метрики оценки модели классификации

- TP: 3+1 FP: 1 FN: 0 TN: 3

Actual Status
● Negative
● Positive

Infection Test
0 - Negative
1 - Positive





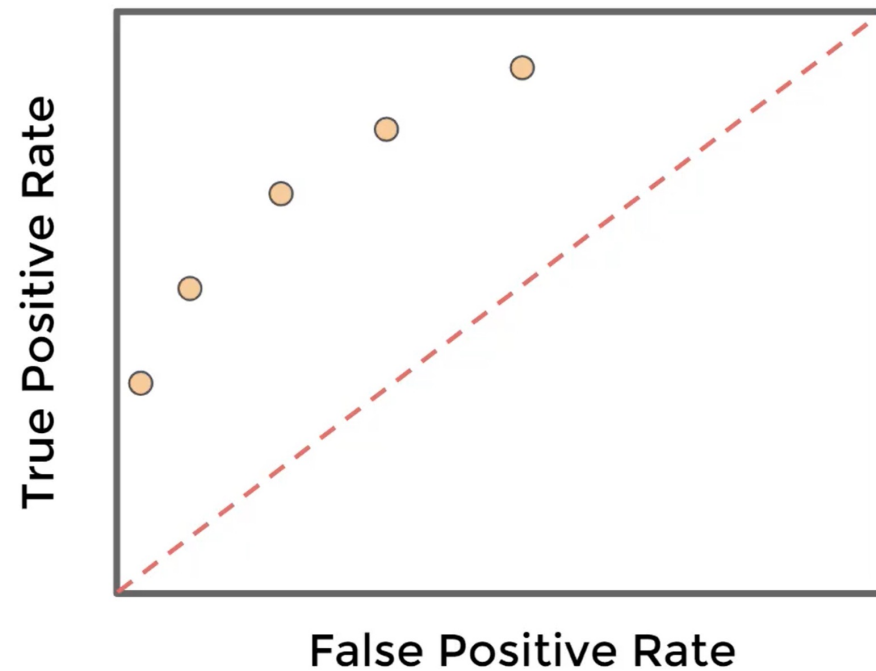
Метрики оценки модели классификации

- В некоторых задачах проще и дешевле согласиться на большее количество ложно-положительных случаев, сокращая ложно-отрицательные.
- Представьте опасный вирус – уж лучше мы получим ложно-положительный тест и сделаем более строгое исследование, чем “пропустим” ложно-отрицательный случай.



Метрики оценки модели классификации

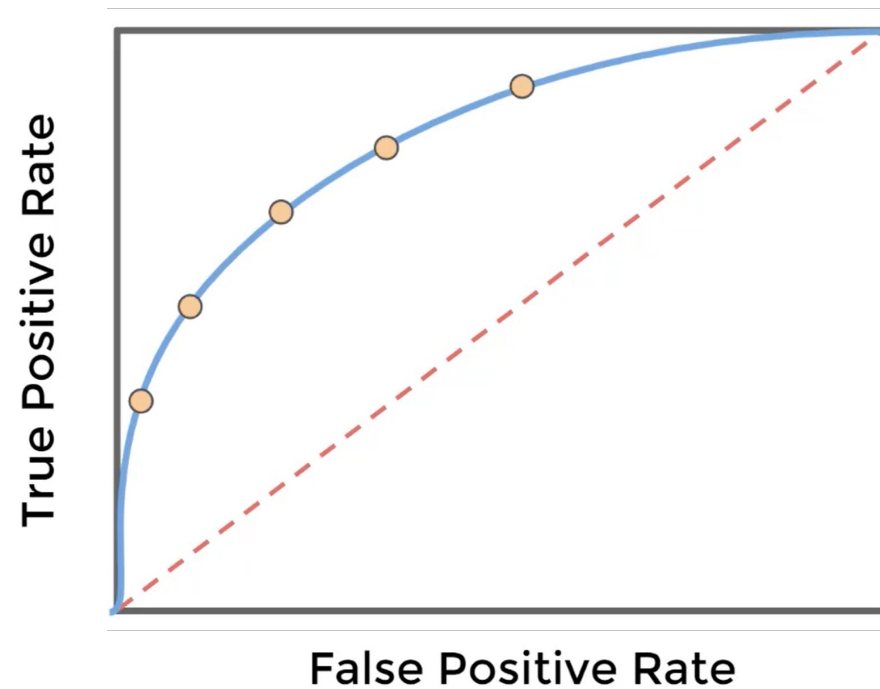
- Нанесём на график True Positive v.s. False Positive для различных пороговых значений для ROC-кривой.





Метрики оценки модели классификации

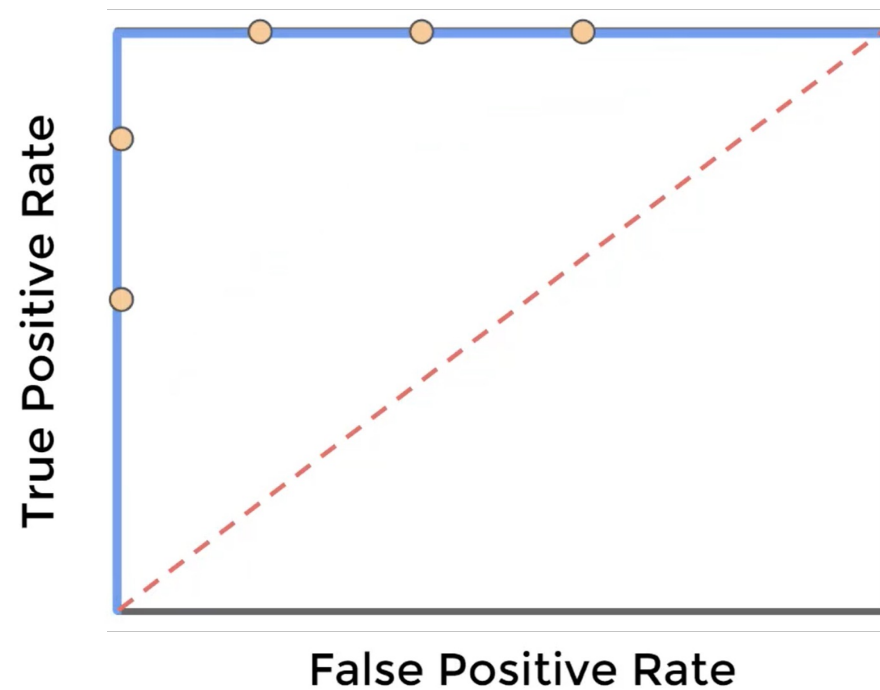
- Меняя пороговое значение, мы можем регулировать True Positive и False Positive.





Метрики оценки модели классификации

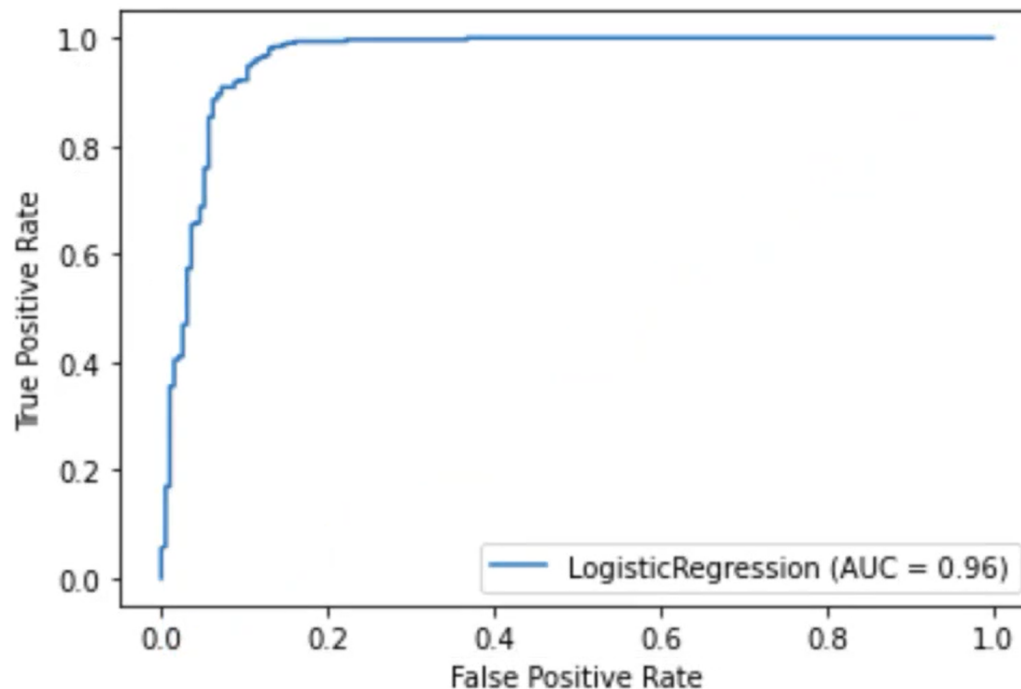
- Идеальная модель будет иметь нулевой FPR
- Красная пунктирная линия - случайное угадывание.





Метрики оценки модели классификации

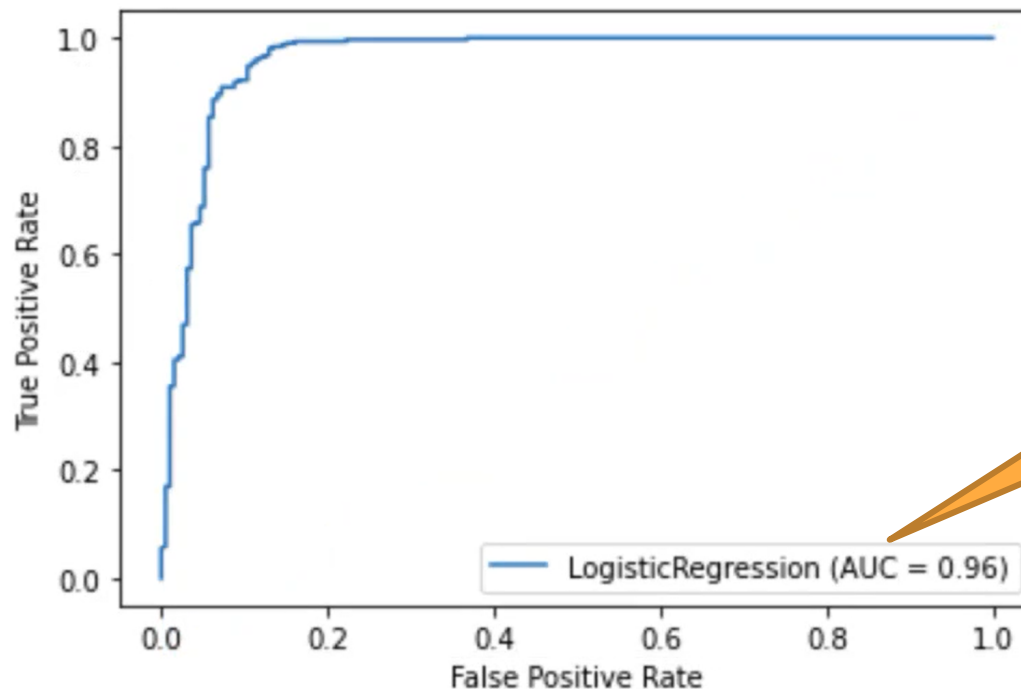
- Зачастую при меньшем количестве данных эти графики могут не выглядеть гладкими.





Метрики оценки модели классификации

- Зачастую при меньшем количестве данных эти графики могут не выглядеть гладкими.

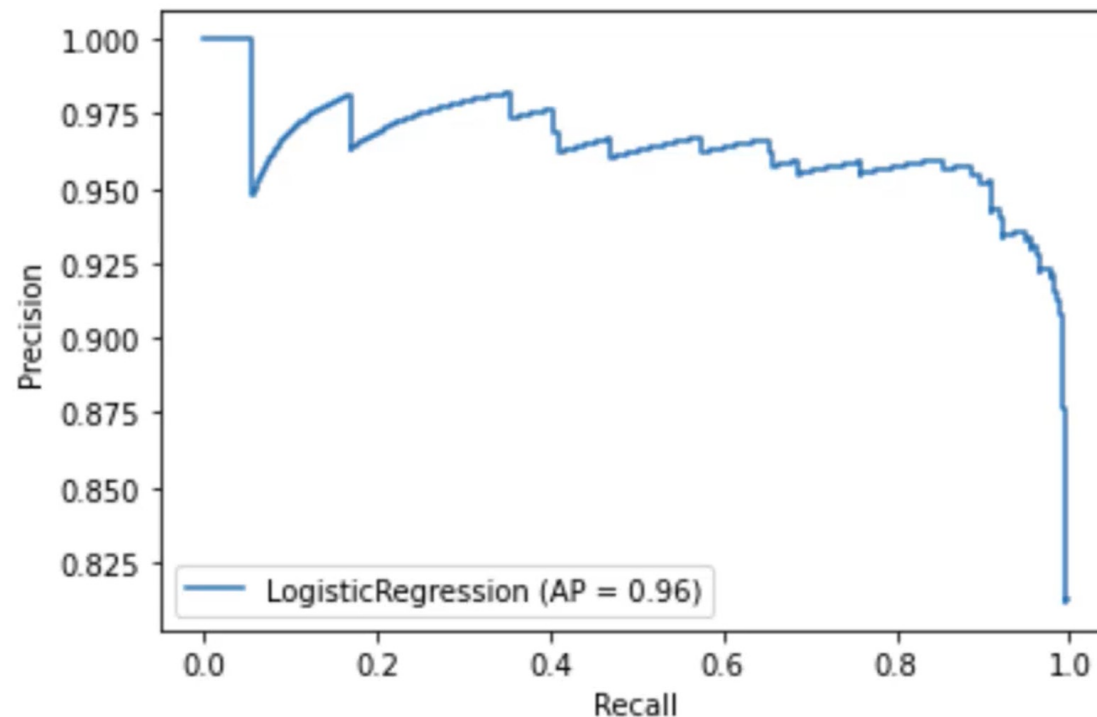


AUC = Area
Under the
Curve



Метрики оценки модели классификации

- Ещё можно визуализировать график precision v.s. recall





Логистическая регрессия в Scikit-Learn

Часть 3: метрики оценки модели



Логистическая регрессия – много-классовые задачи

Часть 1: данные и модель



Логистическая регрессия – много-классовые задачи

Часть 2: обучение и оценка модели



Проект по логистической регрессии

Обзор



Проект по логистической регрессии

Решения