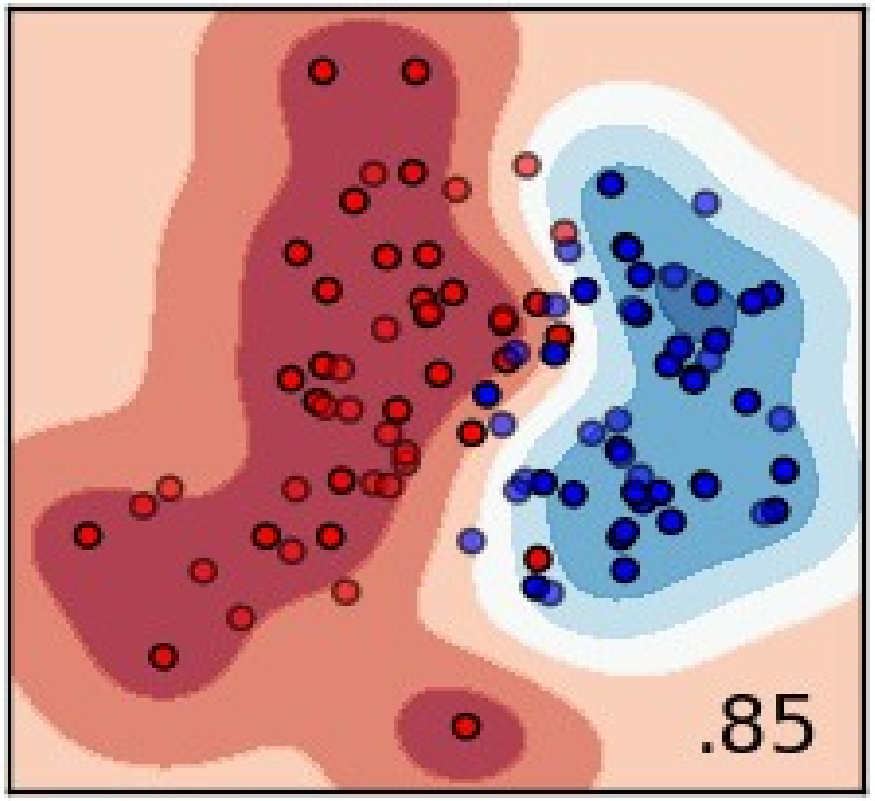
Лекция 3.

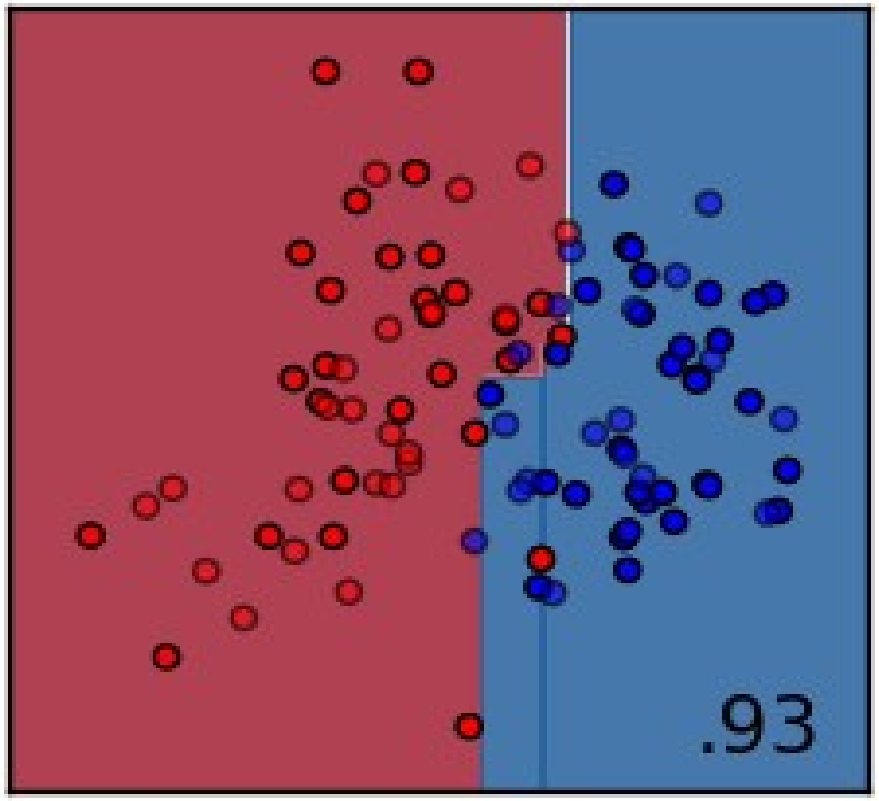
Разделяющие гиперплоскости. Регрессия. Функция штрафа и расчет функции штрафа. Бин-подход. ***Индекс*** Джини.

Павел Владимирович Слипенчук
PavelMSTU@stego.su
ИУ-8

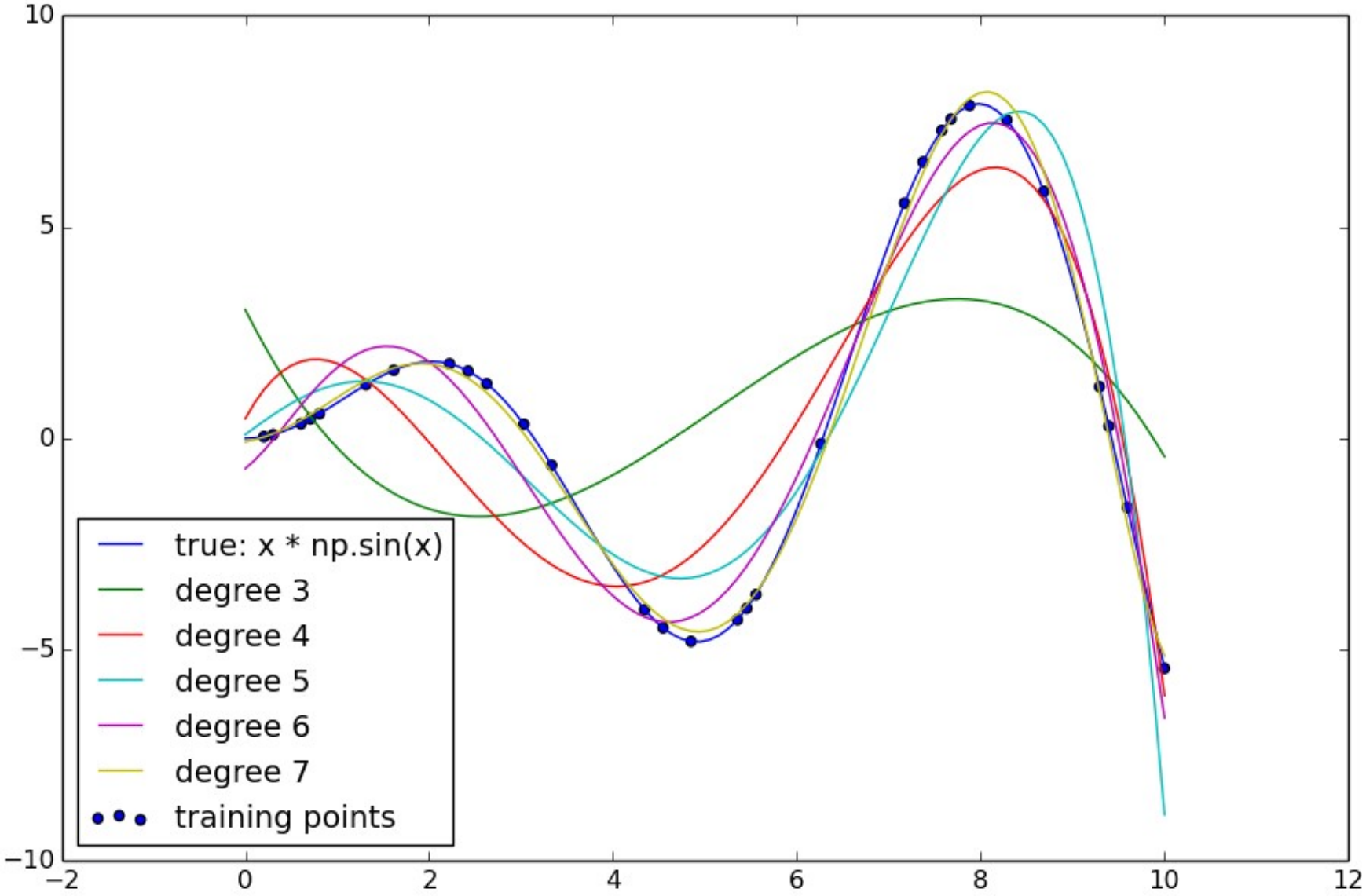
Москва
16.09.2016

Классификация
(Classification)



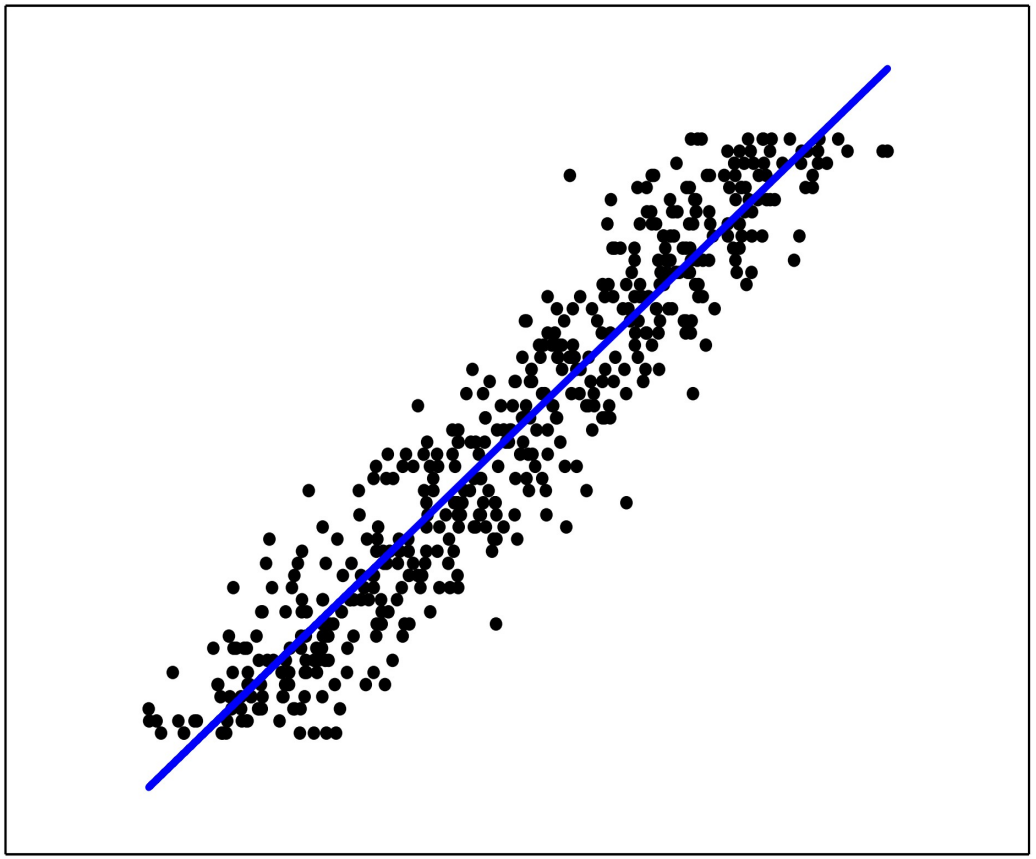


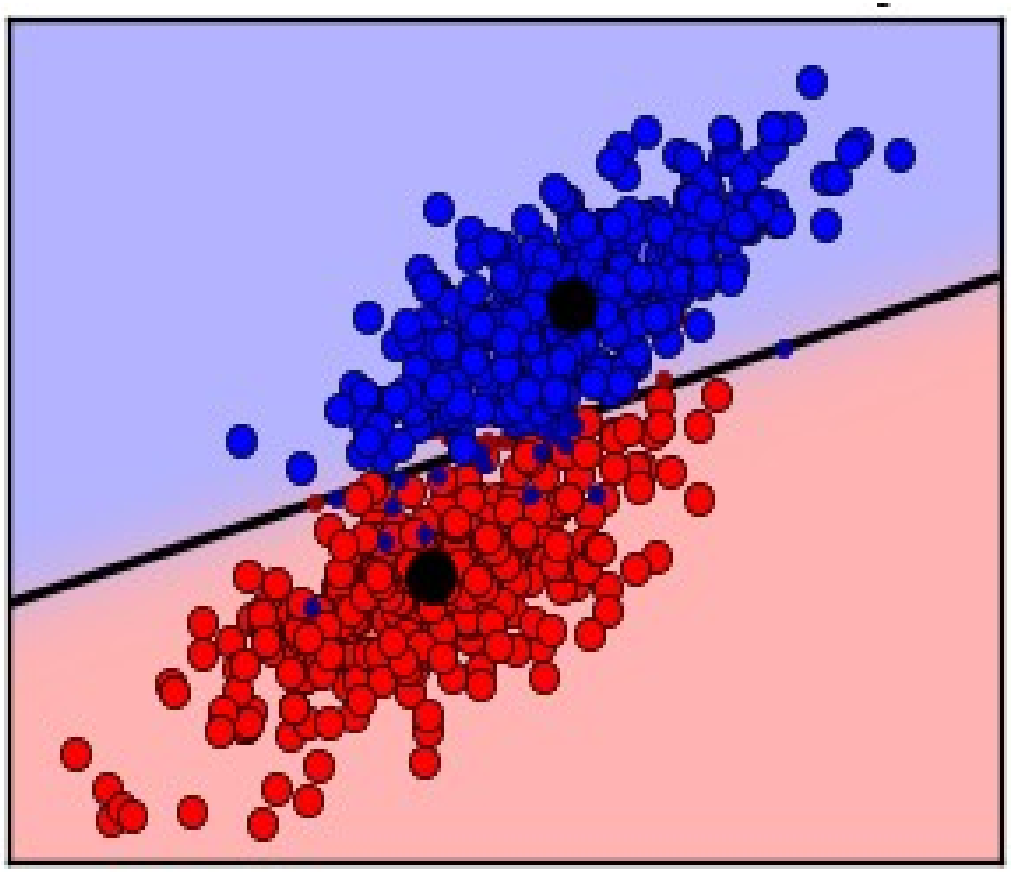
Регрессия
(Regression)



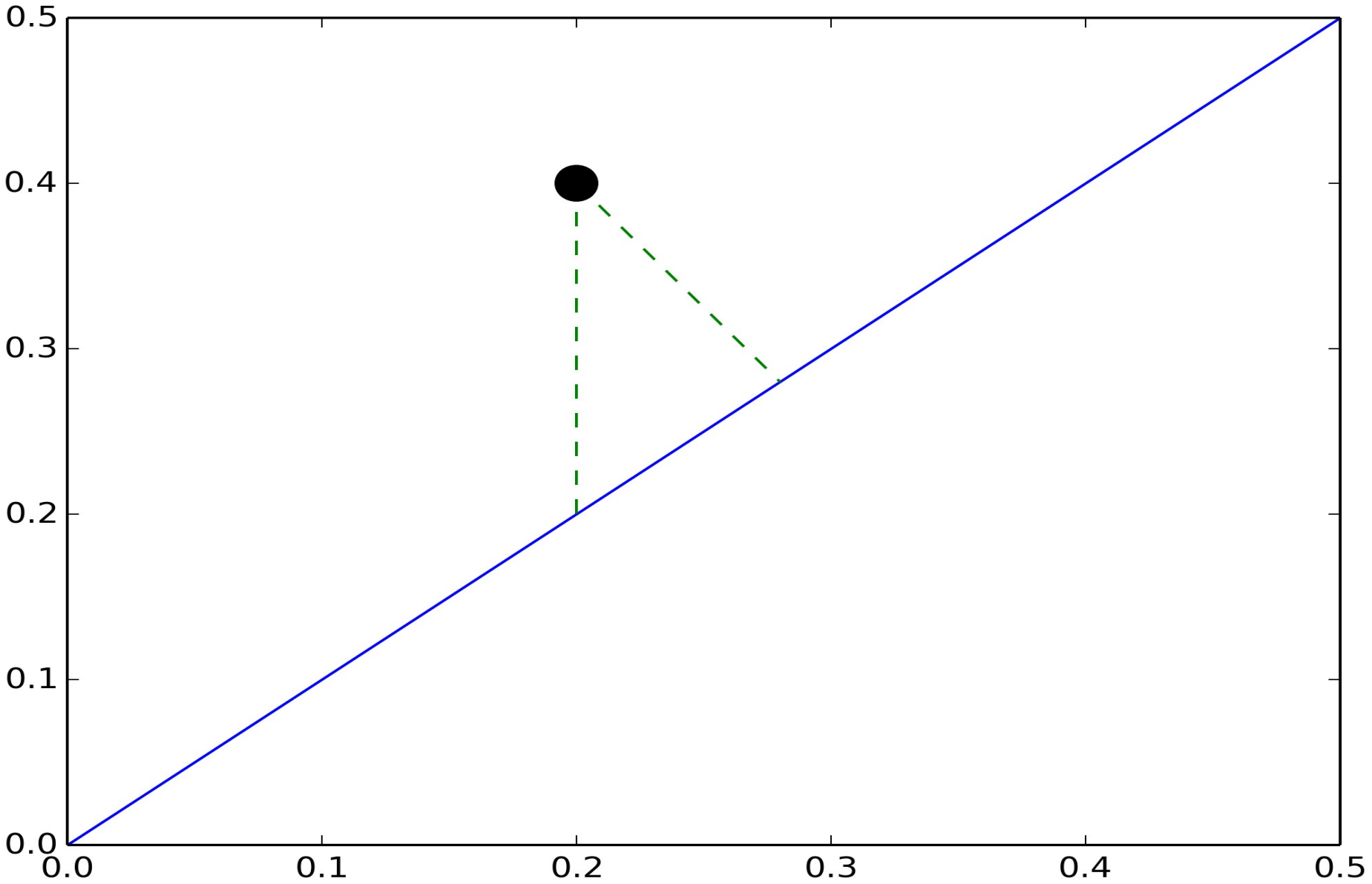
=3=

Линейная регрессия
& линейная классификация





Функция штрафа



Функция штрафа

- Расстояния: Евклида, Минковского (Манхеттенское), ...
- Функция штрафа равно расстоянию от точки до прямой (разделяющей поверхности)
- Регрессия: штраф рассчитывается для всех точек
- Классификация: штраф рассчитывается для точек, попавших в не свой *класс*

Функция штрафа

Задача: минимизация функции штрафа.

- Выбор оптимальной прямой (разделяющей поверхности) из конечного множества.
- Генерация конечного множества гипотез
- Смешанные методы: генерация → выбор → генерация → выбор → …
- Общего решения нет (поэтому существует Machine Learning как инженерное искусство)

Расстояние Махаланобиса

- Δ – ***матрица ковариации***.
- **x** – вектор, содержащий список сравнимых признаков

$$d(x_1, x_2) = \sqrt{(x_1 - x_2)^T \Delta^{-1} (x_1 - x_2)}$$

- За **x1** берут точку. А за **x2** – точку на прямой (на разделяющей поверхности).

  Обозначим прамую за *l*: $d(x_1, l)$

- При каких условиях расстояние Махаланобиса является Евклидовым расстоянием?

Прасанта Чандра Махаланобис

(1893 – 1972)

Частный случай: диагональная матрица ковариаций

- На практике как правило достаточно задать Δ как диагональную матрицу.
- Что это значит?

Частный случай: диагональная матрица ковариаций

- На практике как правило достаточно задать **Δ** как диагональную матрицу.
- Это значит, что все признаки приводяться в “универсальную шкалу” и мы можем находить расстояния между метрами, килограммами, годами и т.д.
- Если **Δ** – единичная матрица, мы имеем евклидово расстояние.
- Что будет, если **Δ** будет “неадекватной”?

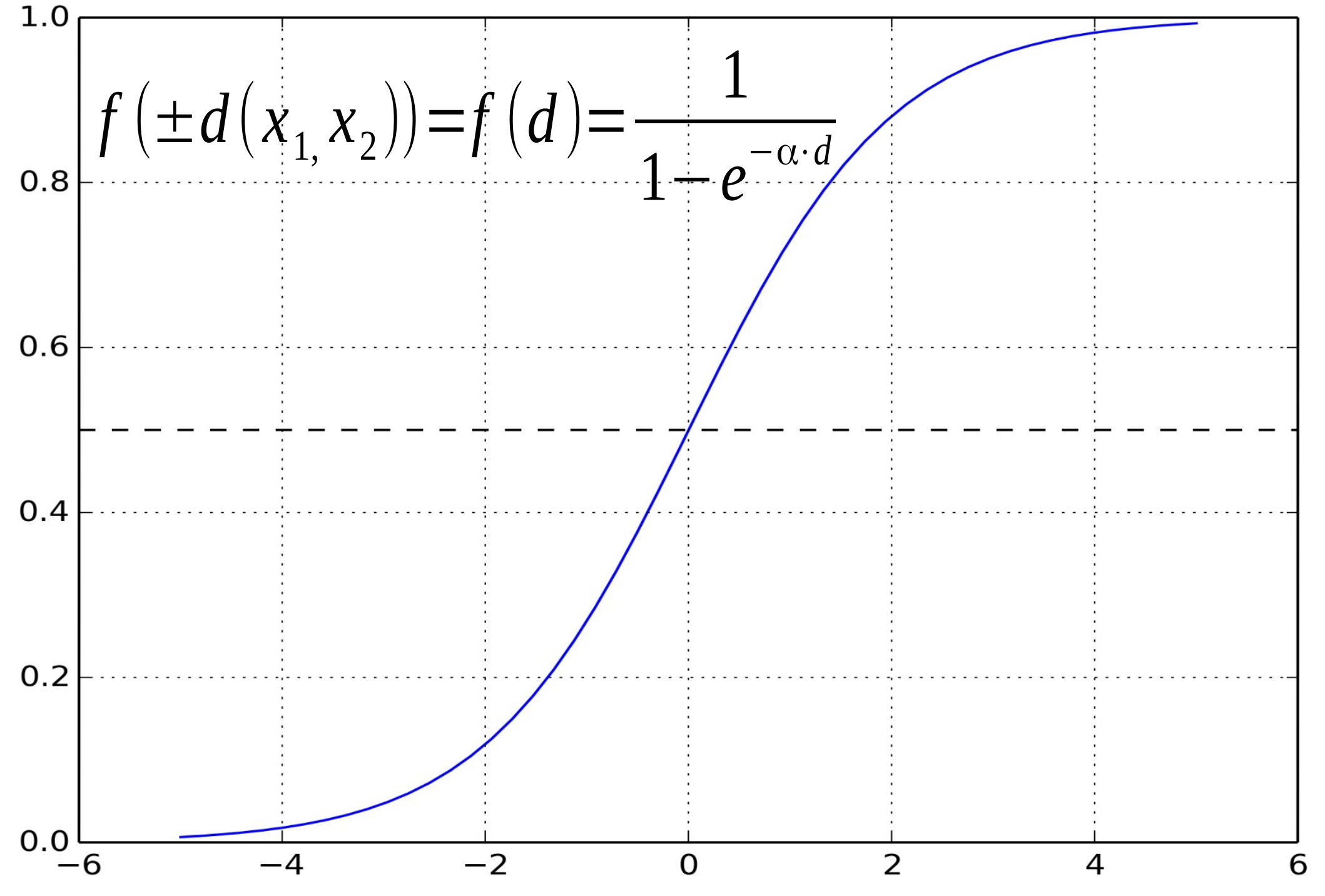
Другие функции штрафа

- **+1** – за каждую точку не в своем классе
- **+1** – за каждую точку “слишком далеко отошедшую от прямой” (на формальном языке математики?)
- **f(d(x1, x2))** – функция от расстояния. Какими должна обладать свойствами эта функция?

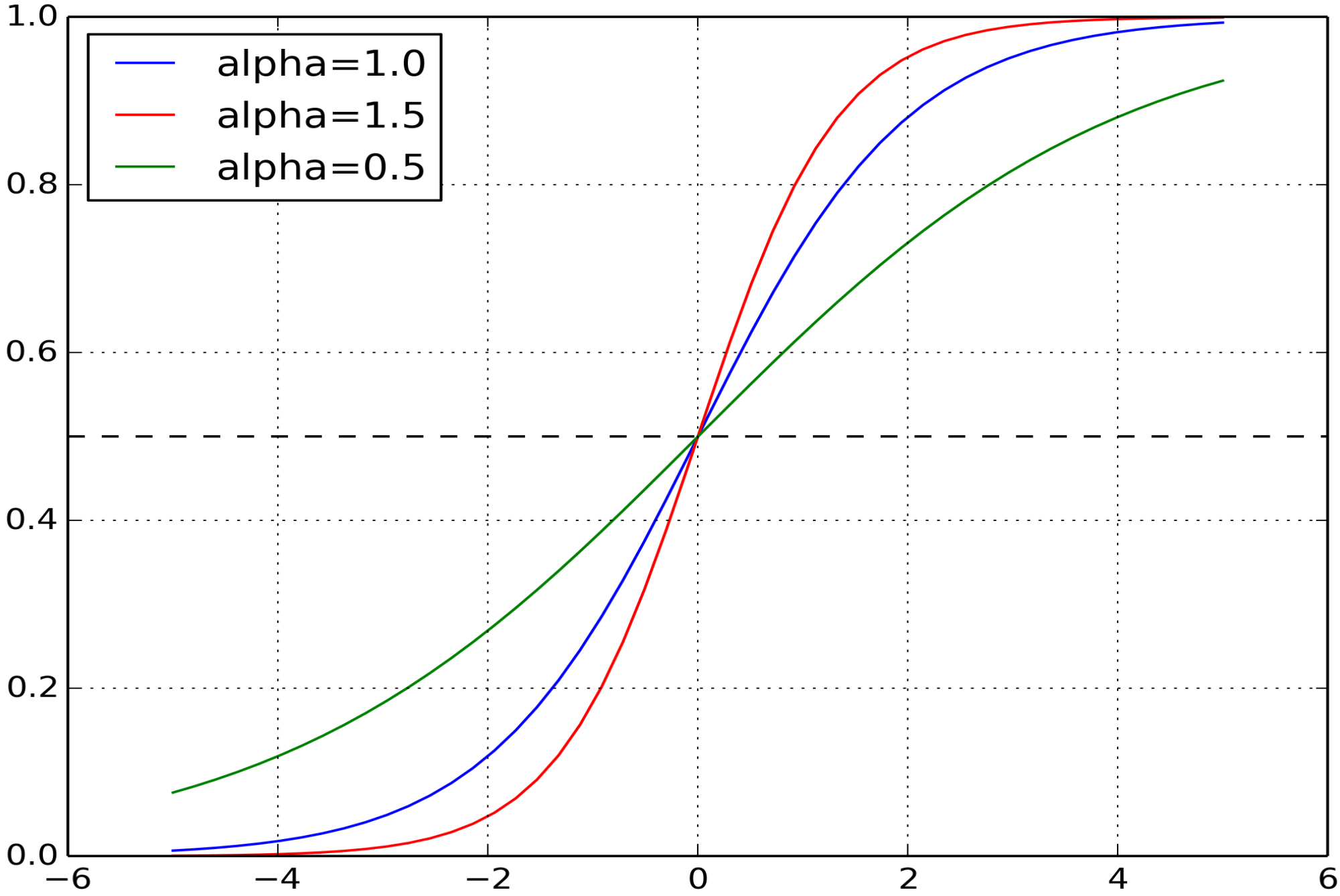
Другие функции штрафа

- **+1** – за каждую точку не в своем классе
- **+1 –** за каждую точку “слишком далеко отошедшую от прямой”
  Т.е. $d\left(\boldsymbol{x}_{1,}\boldsymbol{x}_{2}\right)\geq\alpha$
- **f(d(x1, x2))** – функция от расстояния. Должна быть монотонно неубывающей и положительно определена на множестве всевозможных **d(x1, x2)**

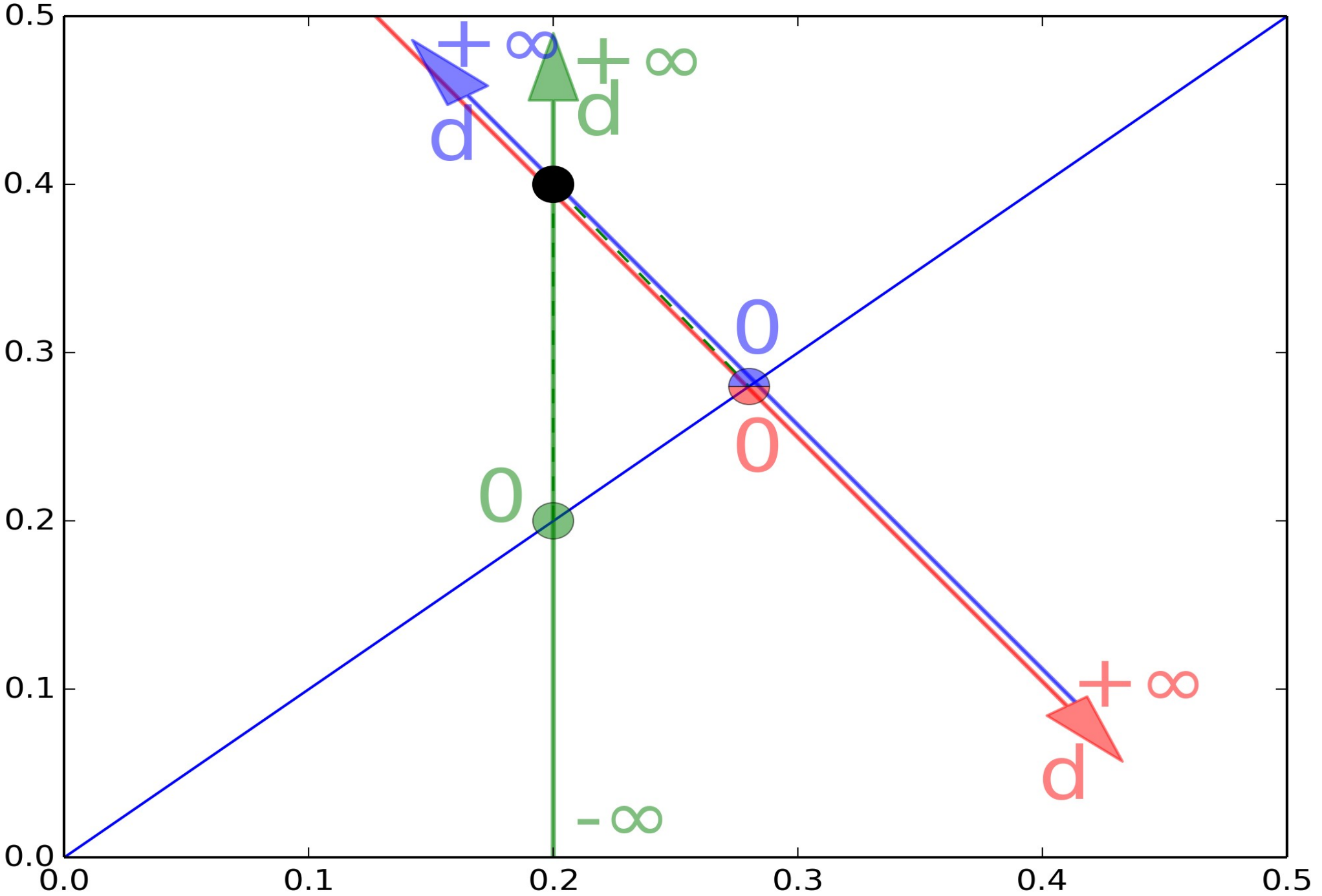
Логистическая функция



Логистическая функция



Логистическая функция



Логистическая функция

- Логистическая функция – функция вероятности. 1.0 – однозначно фрод. 0.0 – однозначно не фрод. 0.5 – “подбрасывание мотетки”.
- Ноль оси d должен быть на разделяющей поверхности.
- В отличие от Евклидова расстояния выбросы с большим расстоянием слабо портят модель.
- Как задать функцию штрафа?

Логистическая функция

- Функция штрафа. Варианты

$$S=\sum\left(y_i-f\left(\boldsymbol{x_i},\boldsymbol{l}\right)\right)$$

$$S=\sum\left\|\left|\left(y_i-f\left(\boldsymbol{x_i},\boldsymbol{l}\right)\right)\right|>0.5\right\|$$

$$S=\sum\left\|\left|\left(y_i-f\left(\boldsymbol{x_i},\boldsymbol{l}\right)\right)\right|>\alpha+0.5\right\|$$

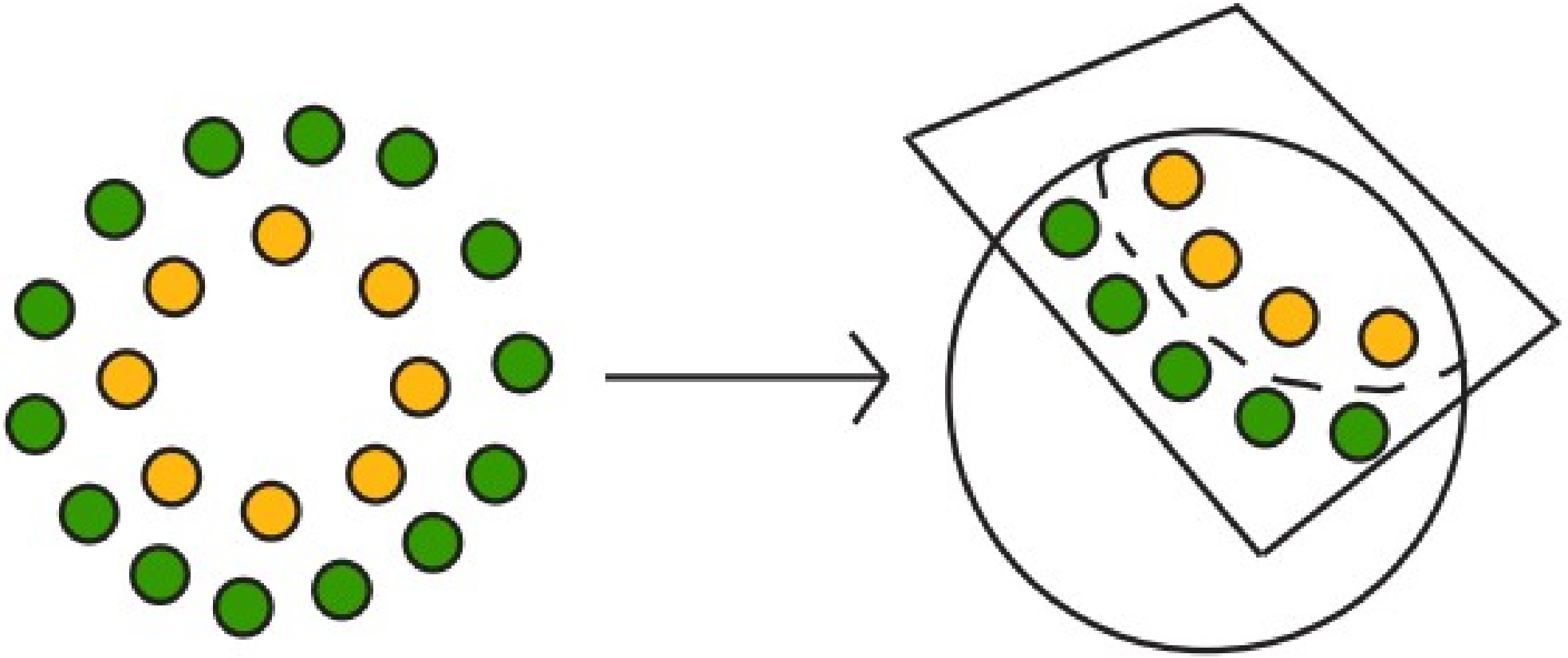
Усложняем ситуацию

- Вместо разделяющей *гиперплоскости* задали *гиперповерхность*. Что делать?

Усложняем ситуацию

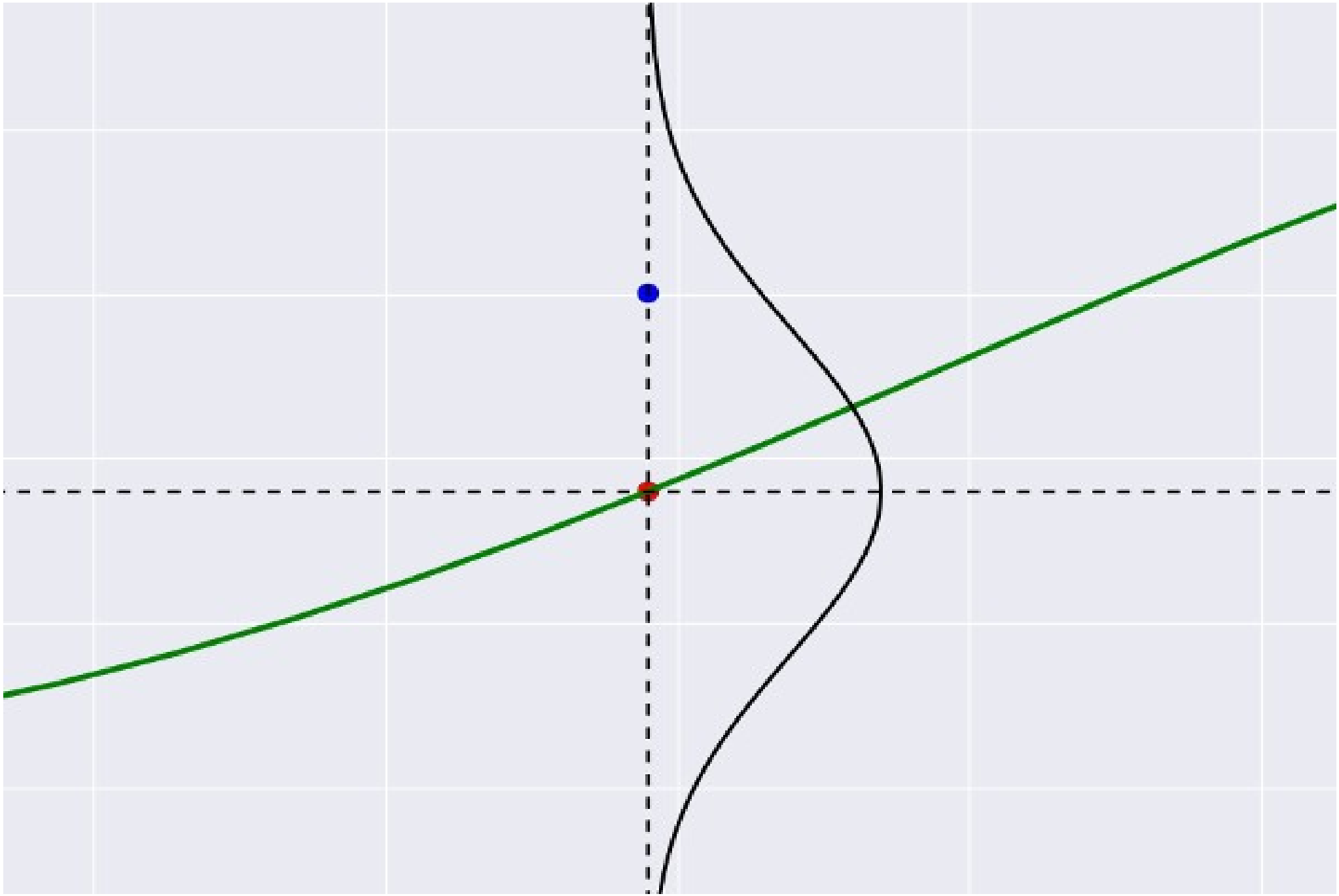
- Так же строим перпендикуляр к ***гиперповерхности***.

Идея Метода опорных векторов (Support vector machine, SVM)

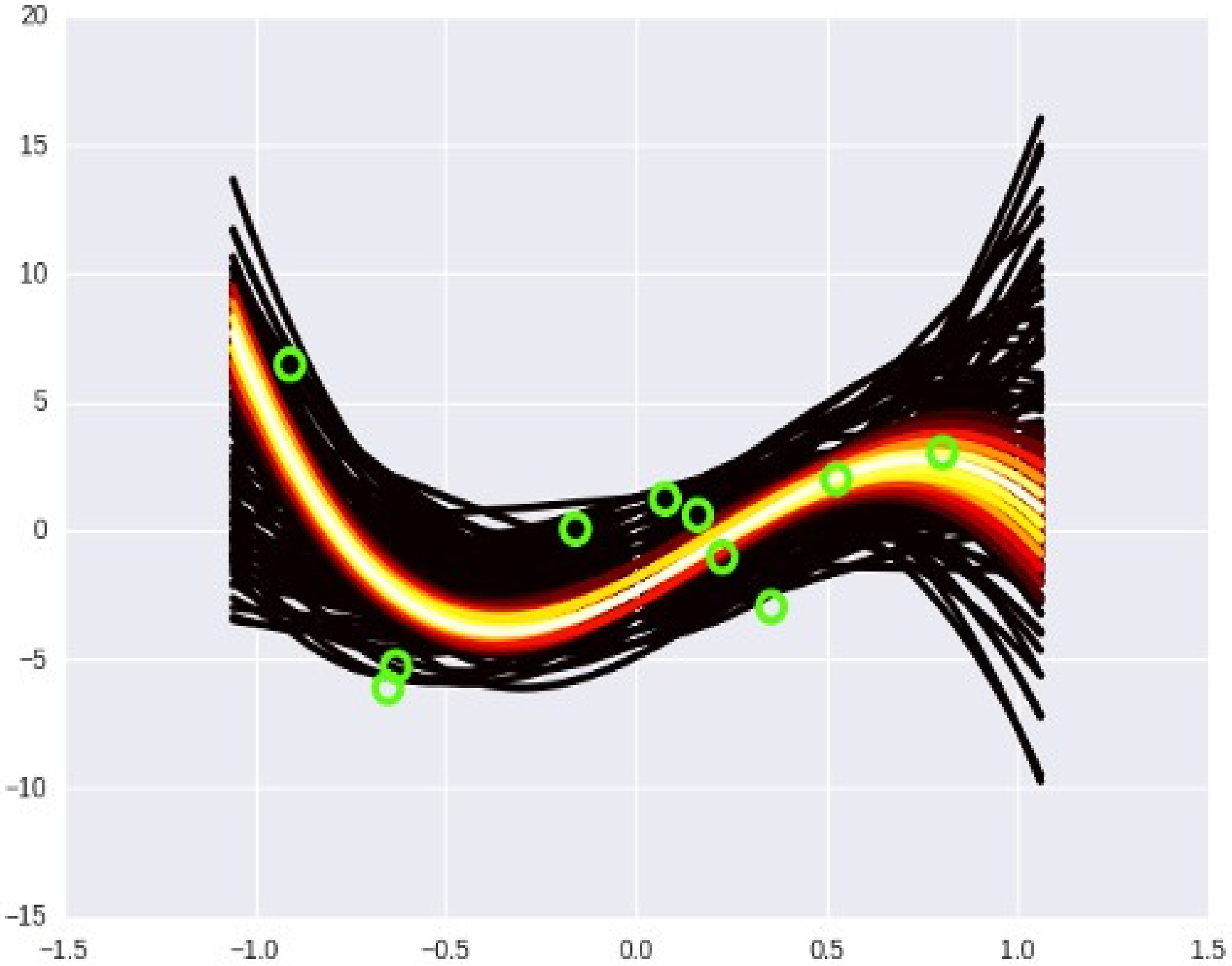


Байесовский подход

- Мы высчитываем совокупность различных решений, для каждого находим “*вероятность истинности*” этого решения.
- Задавать “вероятность истинности” можно различными способами (в частности сумма всех вероятностей может быть больше 1)



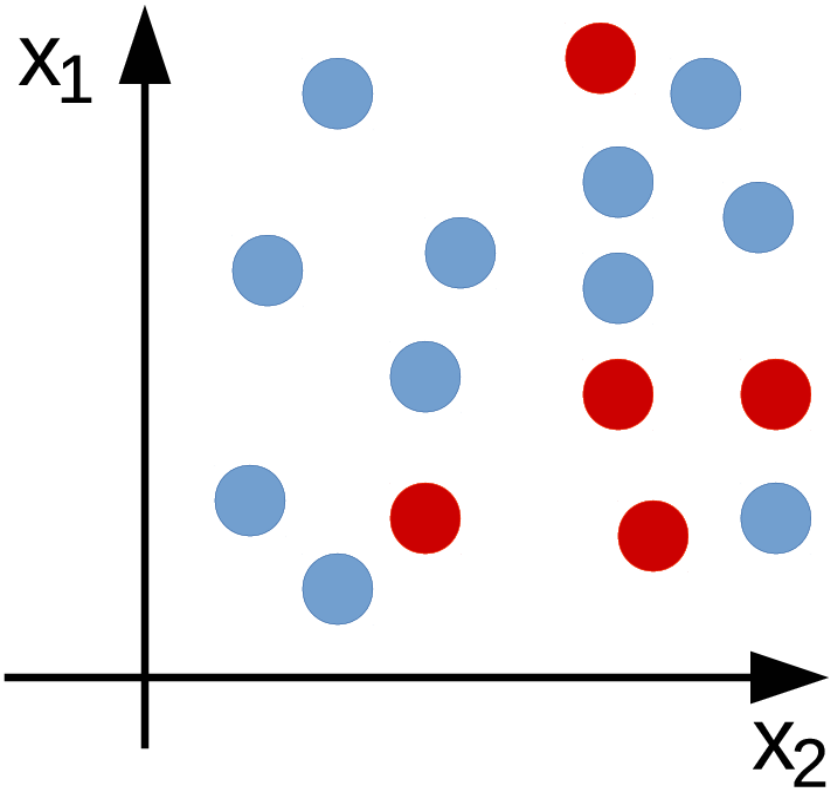
- Копилефт: https://habrahabr.ru/post/276355/

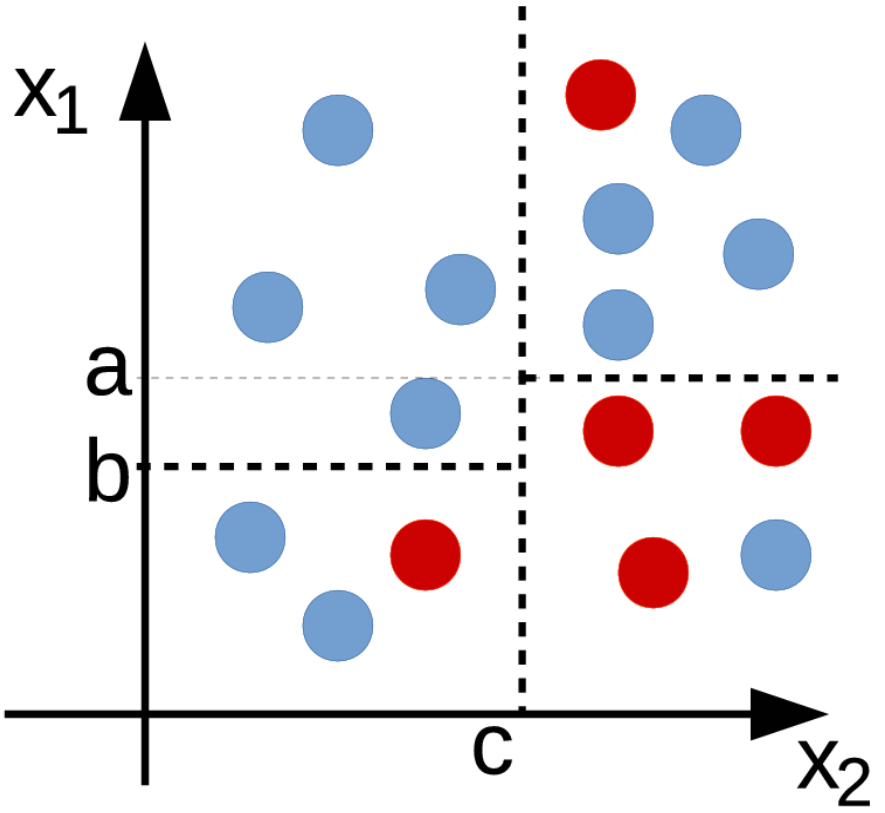


- Копилефт: https://habrahabr.ru/post/276355/

Bin-подход

(bin – карман, корзиночка, мешок для сбора хмеля и т.д)



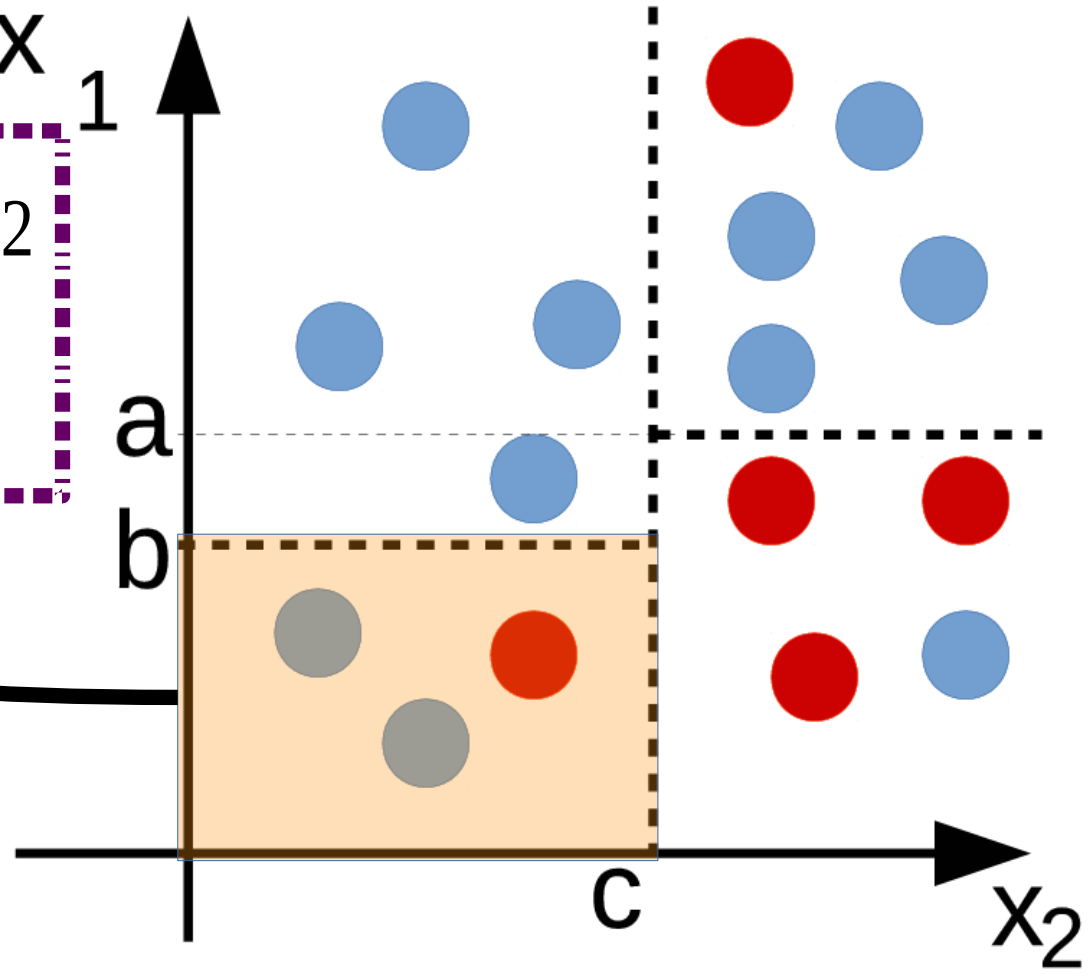


Индекс Джини*

= Индекс “загрязненности”

*не путать с ***коэфициэнтом*** Джини

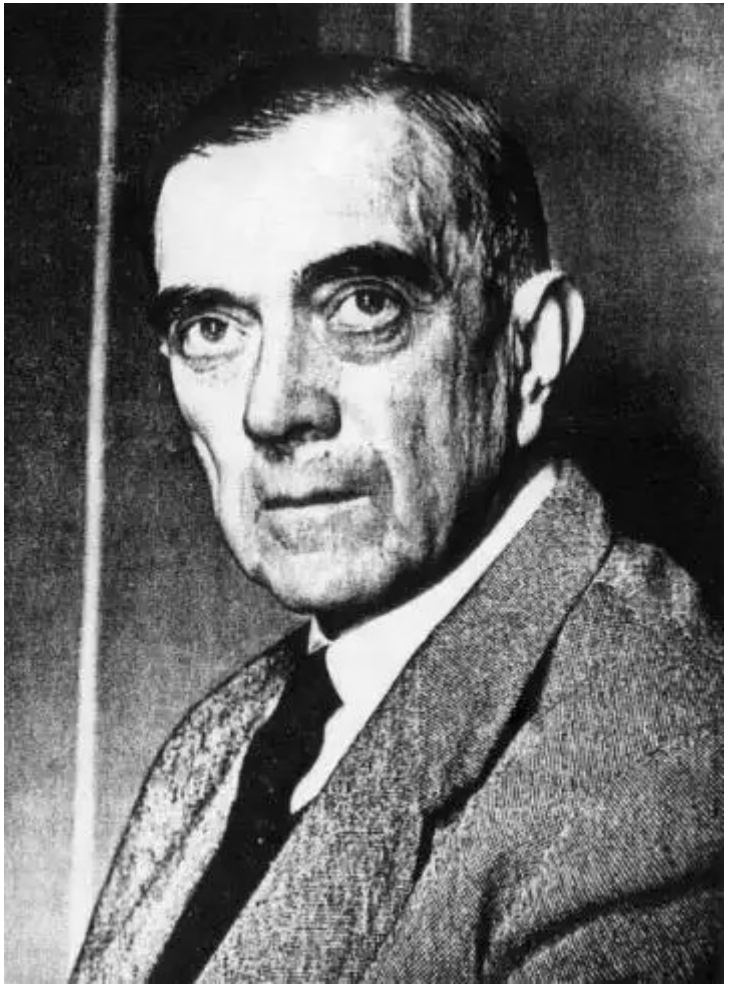
$$\delta = 1 - \sum_{\forall c_i \in C} \|c = c_i\|^2$$



- Анализ малых данных: “Знакомьтесь, Джини”

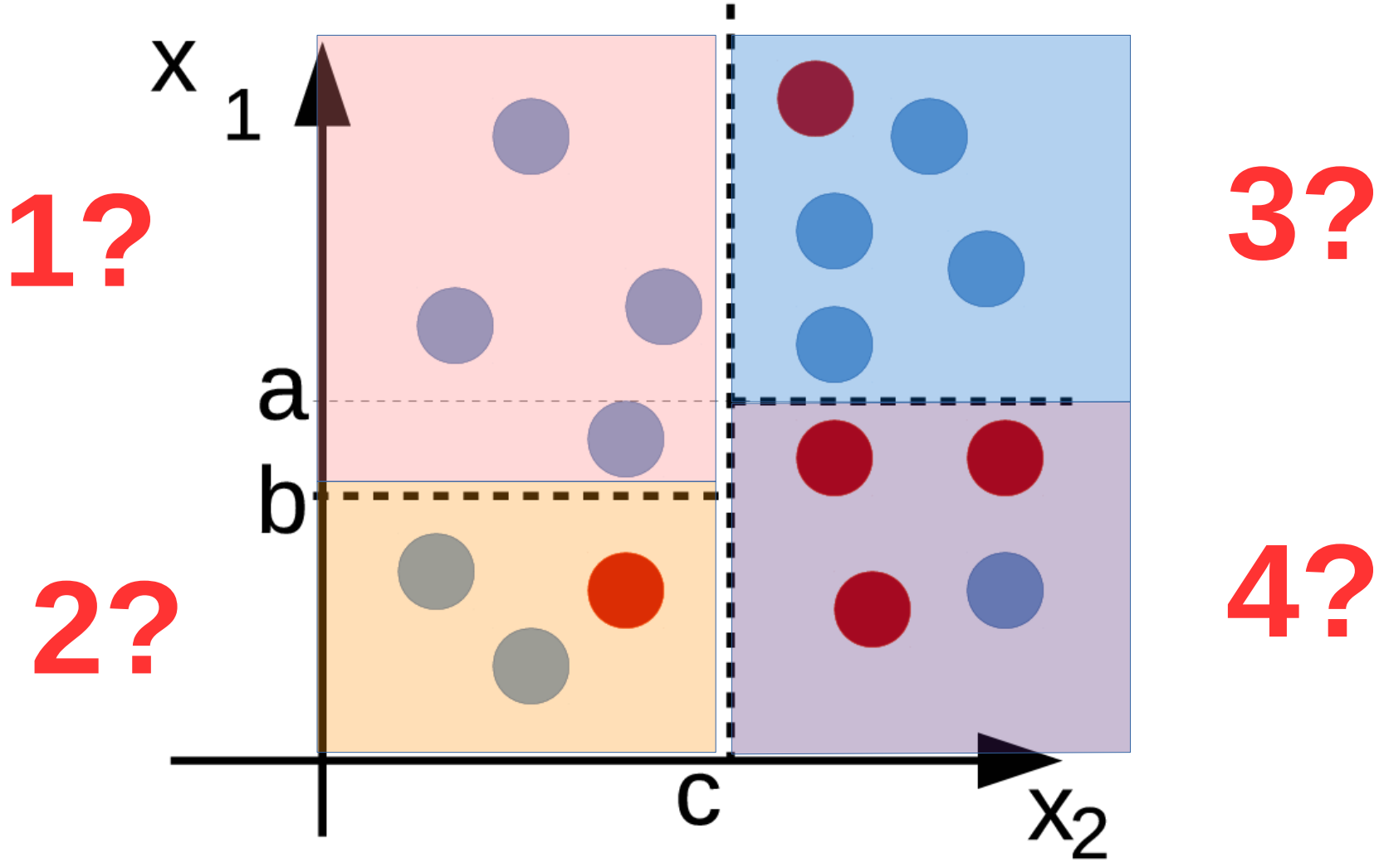
Corrado Gini

- Математик, экономист, социолог.
- Эксперт по статистике
- Автор коэффициента Джини (для кривой Лоренца)
- Автор книги «Научные основы фашизма» (1927)



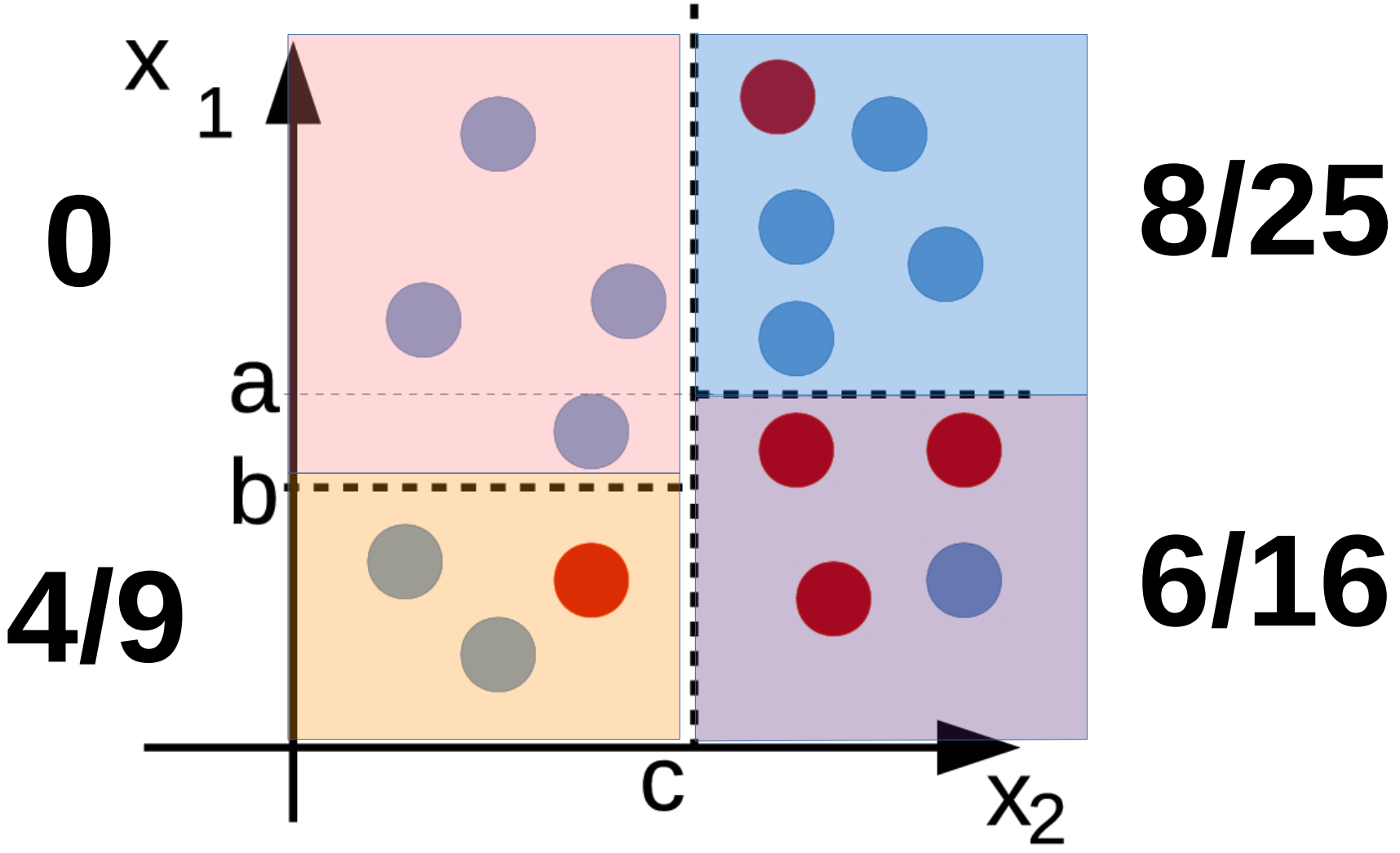
1884 – 1965

Индекс Джини



- Анализ малых данных: “Знакомьтесь, Джини”

Индекс Джини



- Анализ малых данных: “Знакомьтесь, Джини”

Bin-подход и индекс Джини

- **Зачем вообще нужно проводить разбиновку и высчитывать индекс Джини?**
- **“Крупными” или “мелкими” должны быть бины?**
- **Как проводить разбиновку не на плоскости, а в R3 (3 признака) или в Rn (n признаков)? Что из себя представляют бины?**
- **Важно ли расстояние для разбиновки?**
- **Как алгебраически задаются бины?**

ДЗ. *Настроить систему.*

- Советую Linux
- PyCharm
- Python 2.7 – приоритетнее.
- Numpy, scipy
- Matplotlib
- Pandas
- Scikit-learn
- Учебник по Python: М.Саммерфильд (M.Summerfield) “Python на практике”
- Запустить тестовый код: http://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html