Южный федеральный университет

@SINECOR

Задачи классификации

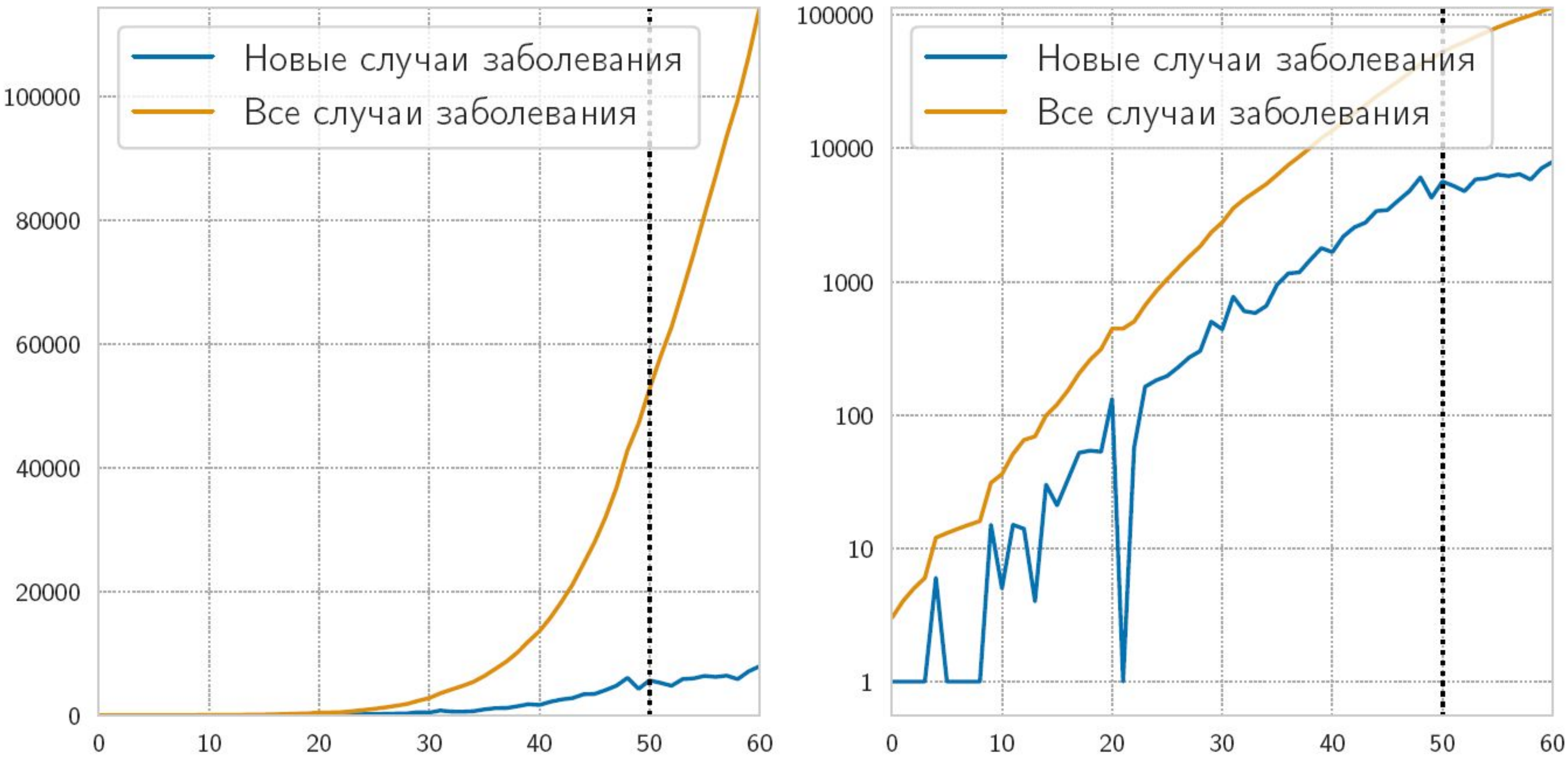
Сергей Николенко

27 ноября 2025 г.

Технологии и фронтиры науки о данных

Пример: COVID-19

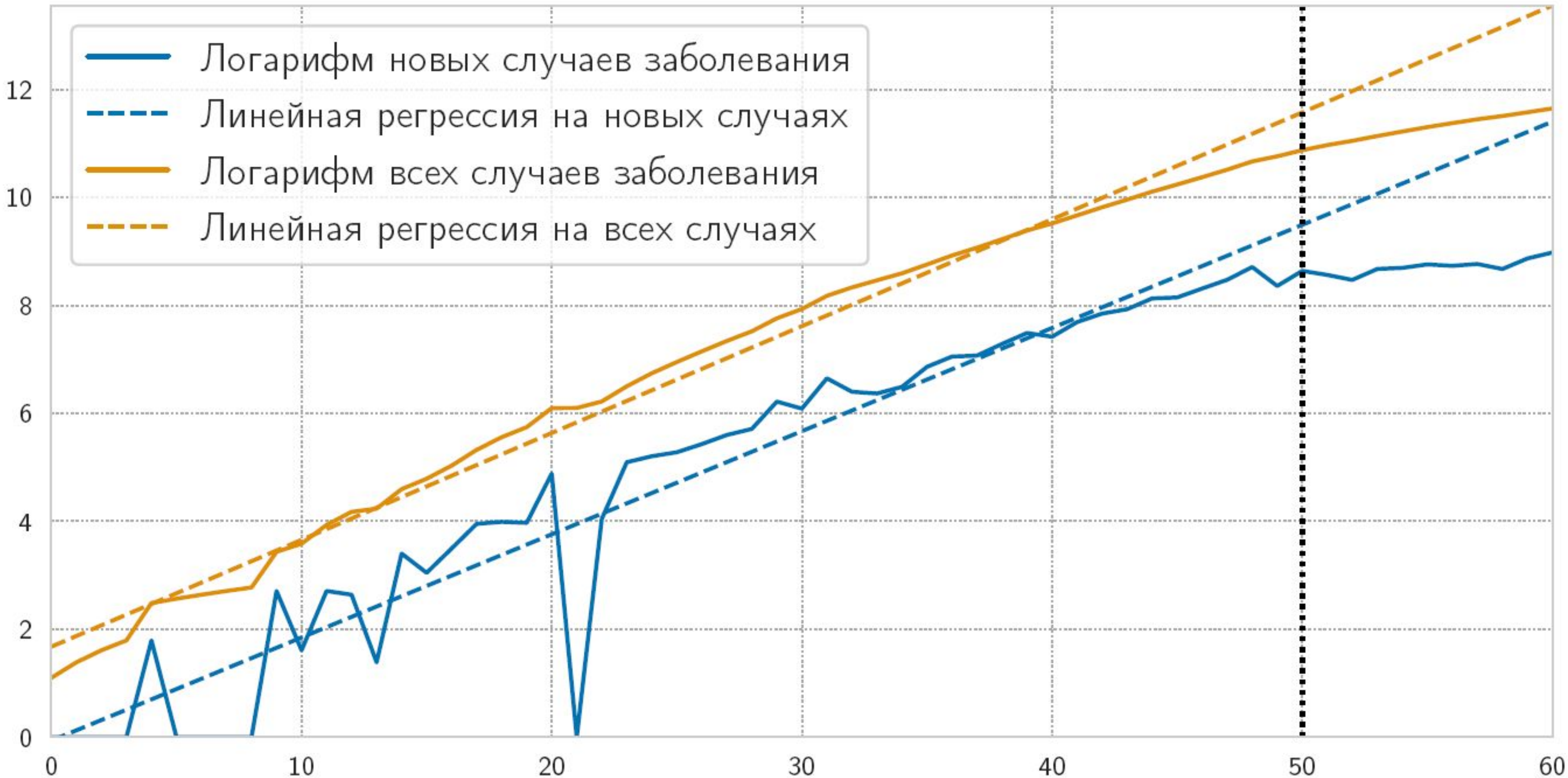
Как предсказать развитие эпидемии?



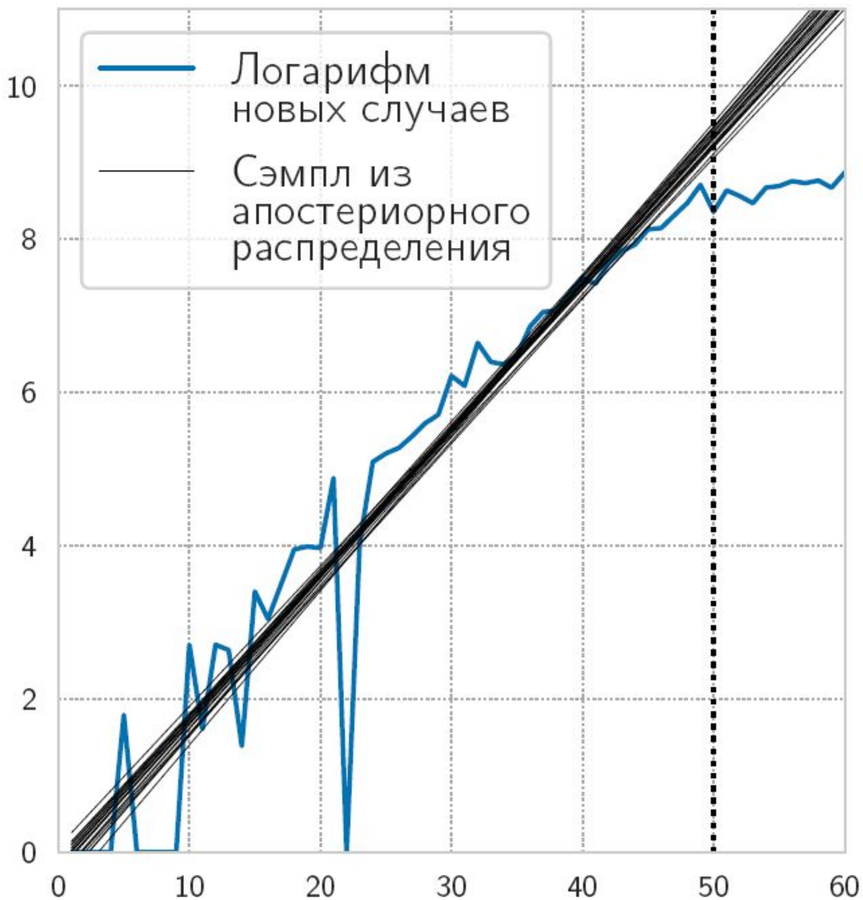
(а) Линейная шкала

(б) Логарифмическая шкала

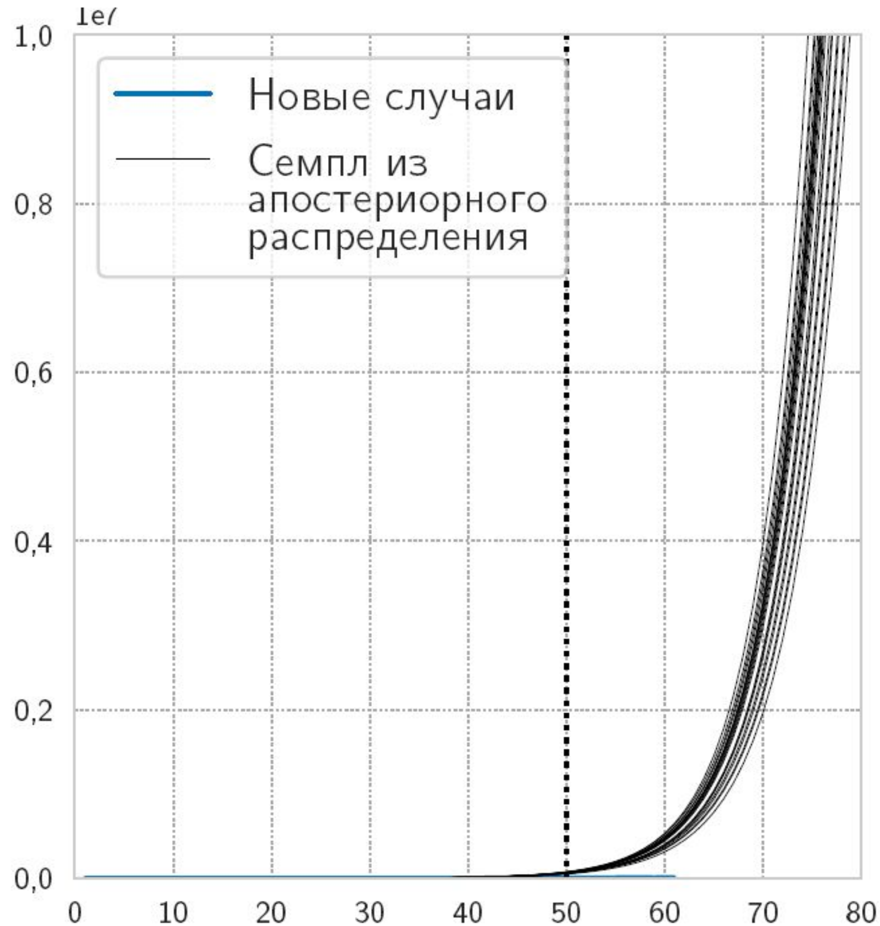
Кажется, что всё просто...



...но нет

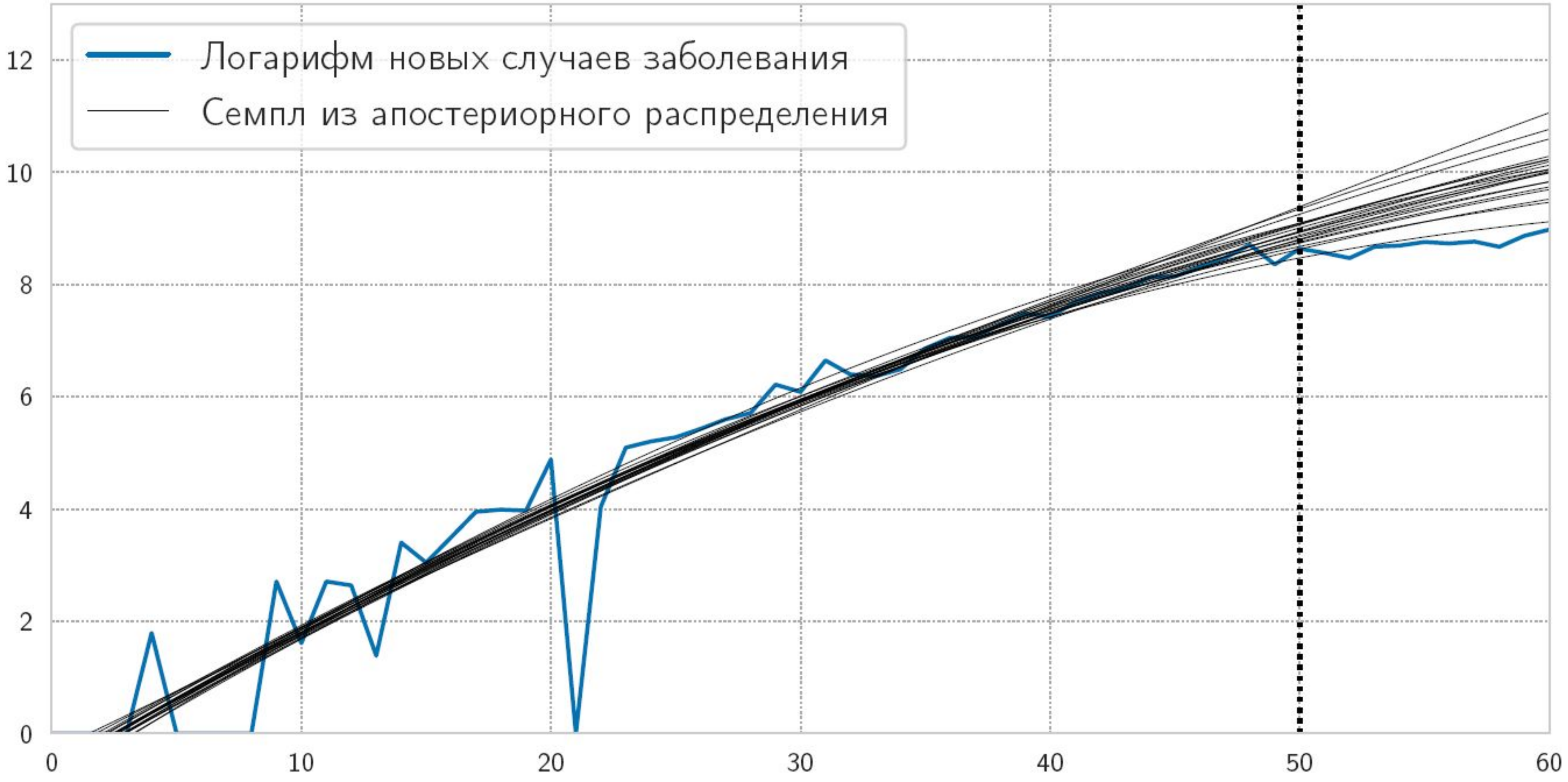


(а) Логарифмическая шкала

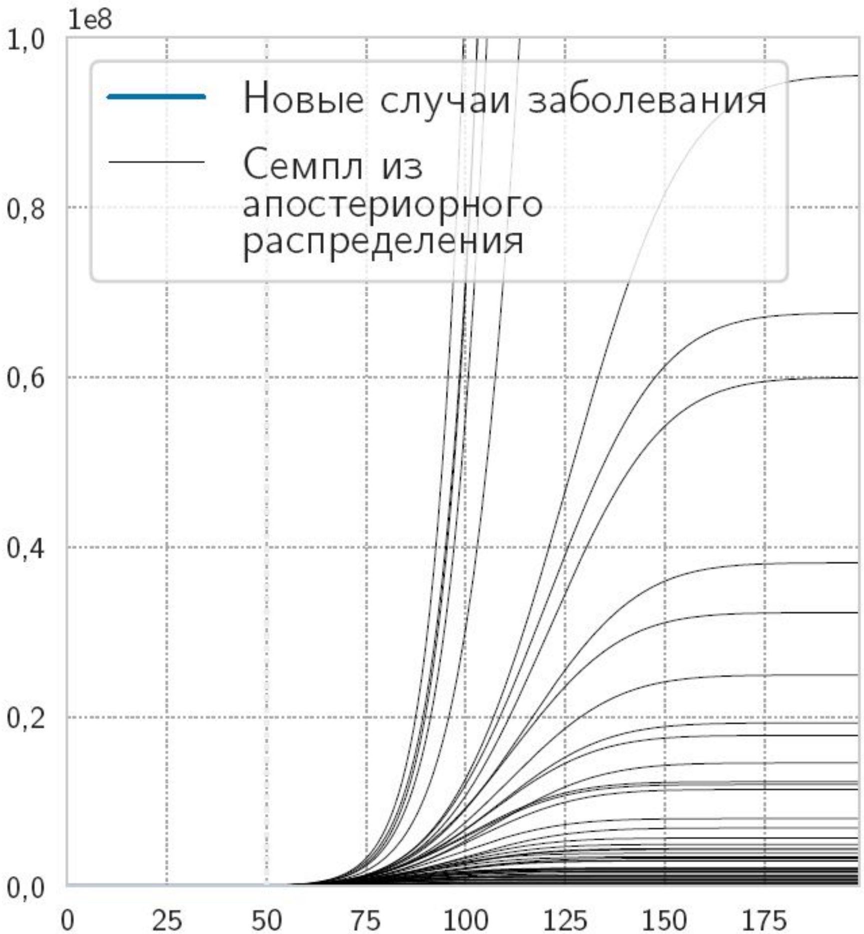


(б) Линейная шкала

А если квадратичная функция в логарифме?



Да тоже как-то не очень понятно...



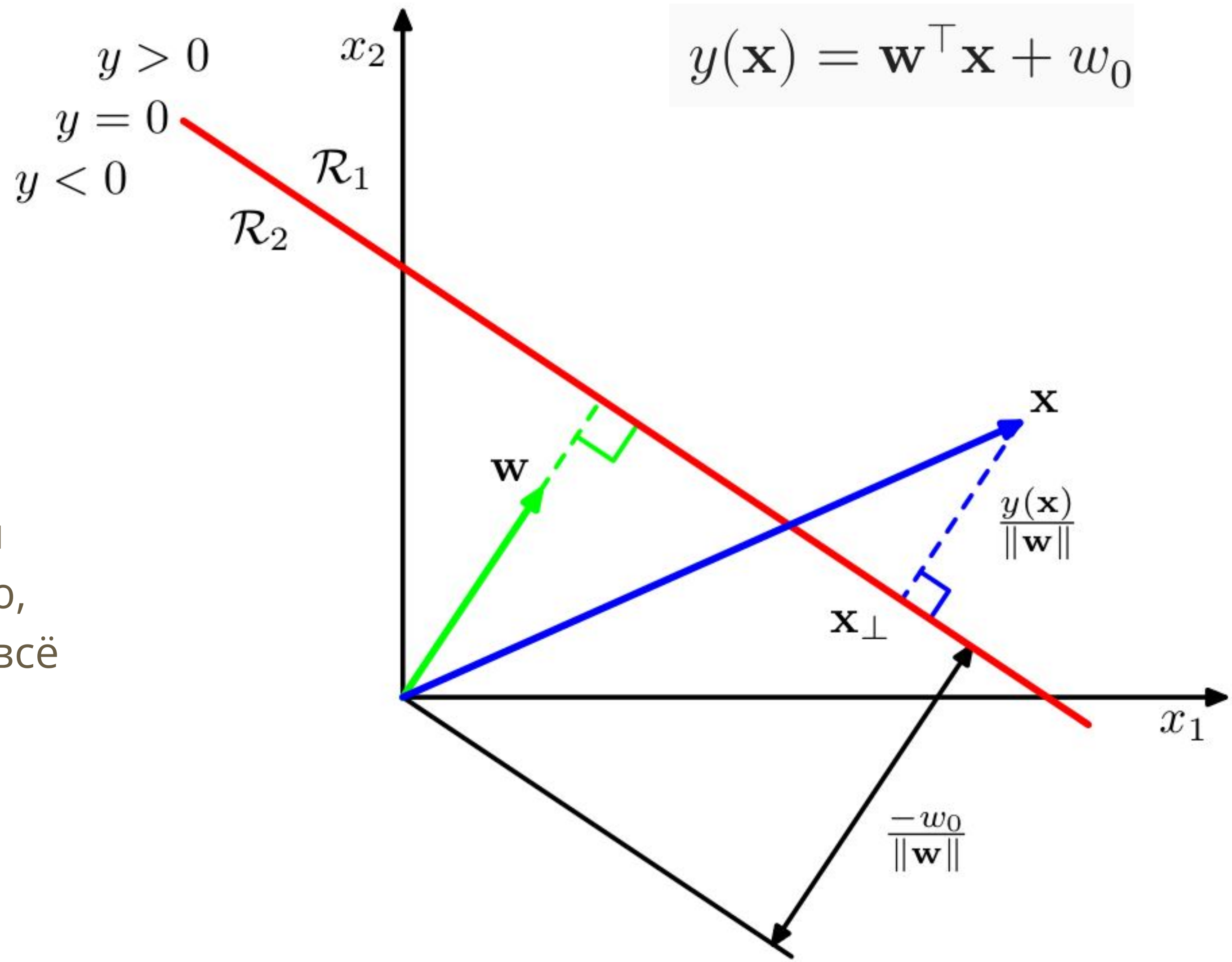
(а) Несколько примеров

(б) Гистограмма значений насыщения

Геометрия классификации

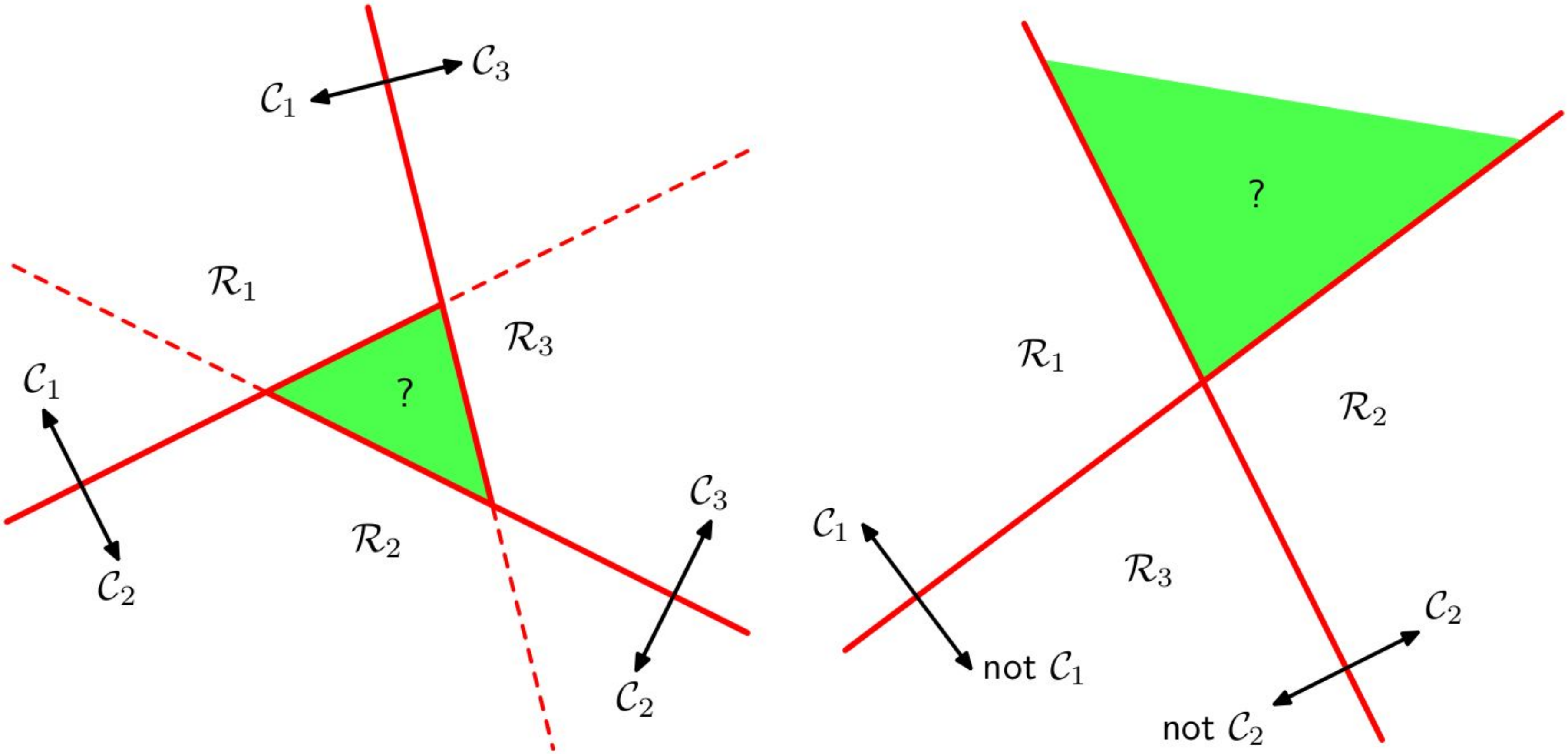
Линейная разделяющая поверхность

- С линейной разделяющей поверхностью, казалось бы, всё просто…



Линейная разделяющая поверхность

- …но возникают странные проблемки в случае нескольких классов
- Что делать?..

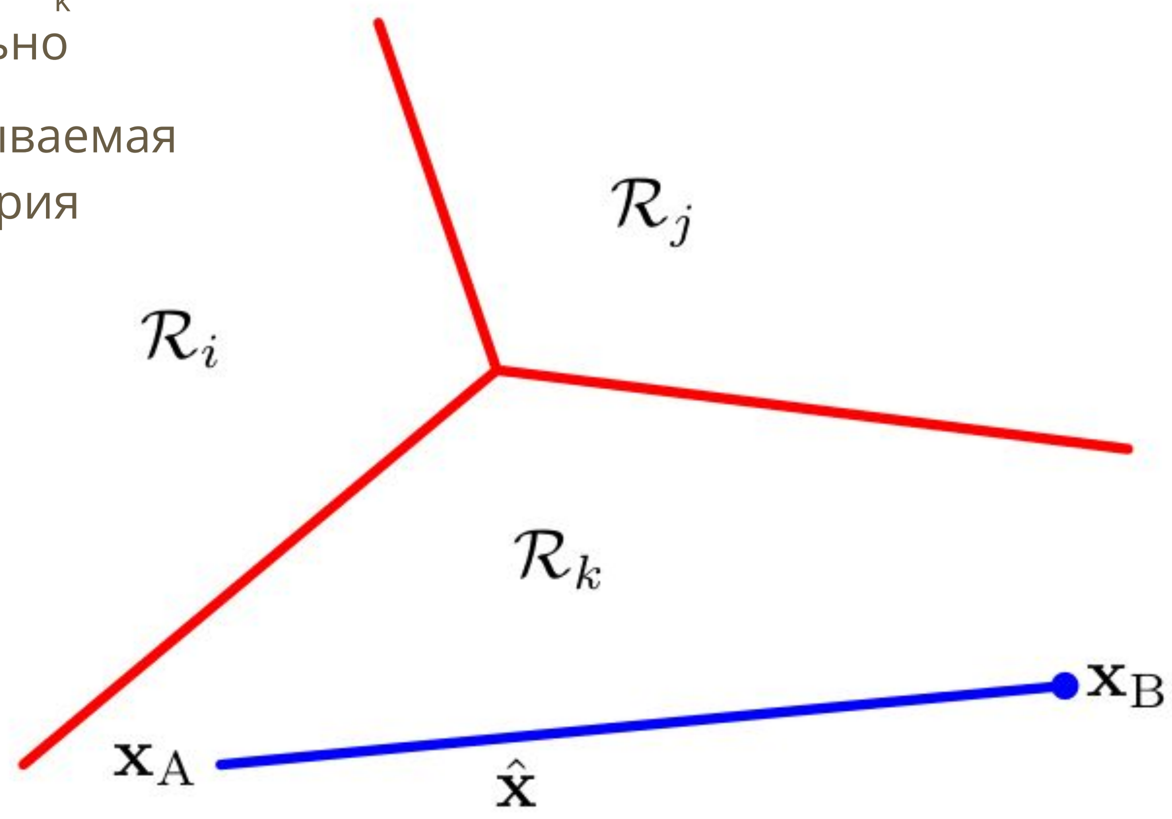


Линейная разделяющая поверхность

- Лучше рассмотреть единый дискриминант из K линейных функций:

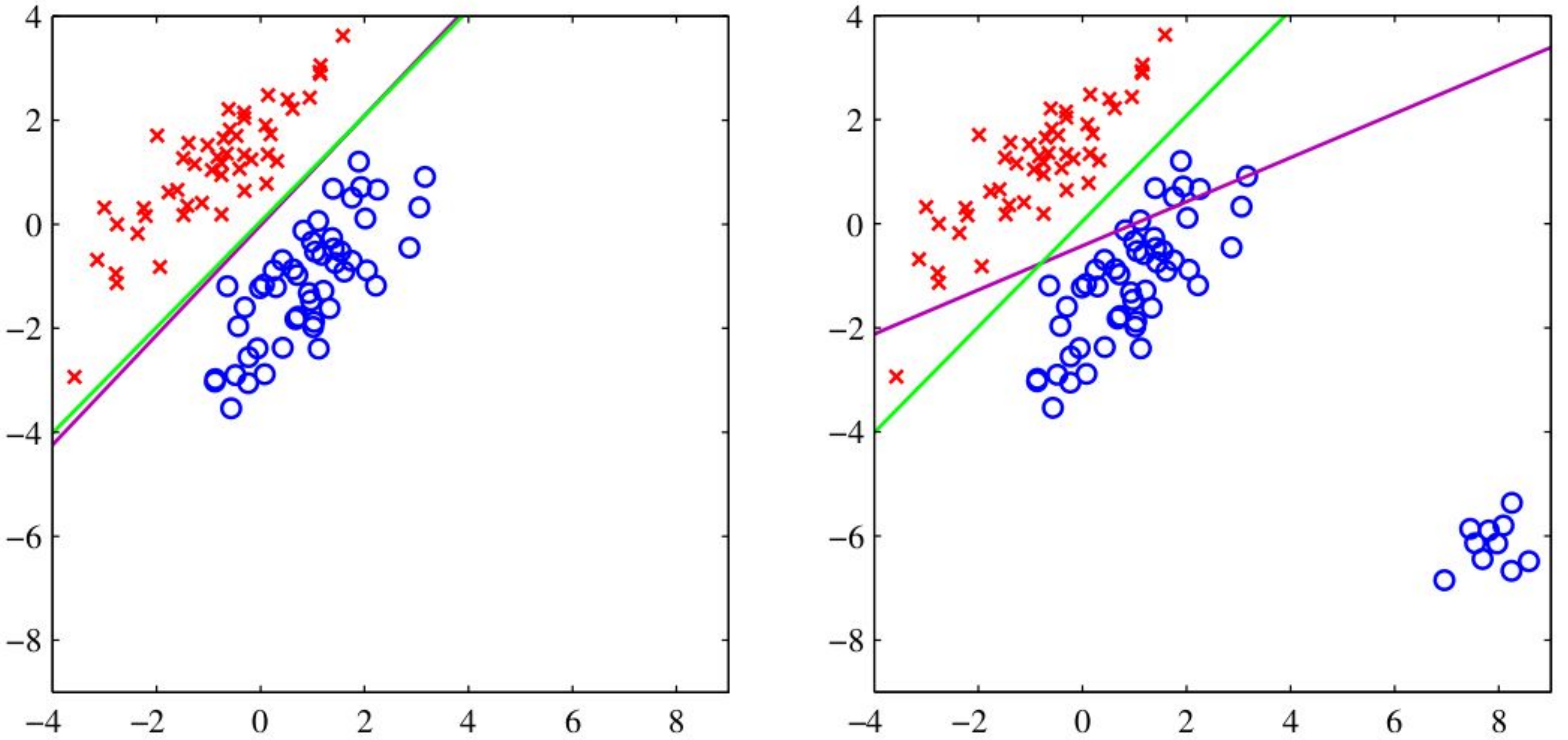
$$y_k(\mathbf{x}) = \mathbf{w}_k^\top \mathbf{x} + w_{k0}.$$

- Классифицировать в C_k, если $y_k(\mathbf{x})$ максимально
- Получается так называемая тропическая геометрия



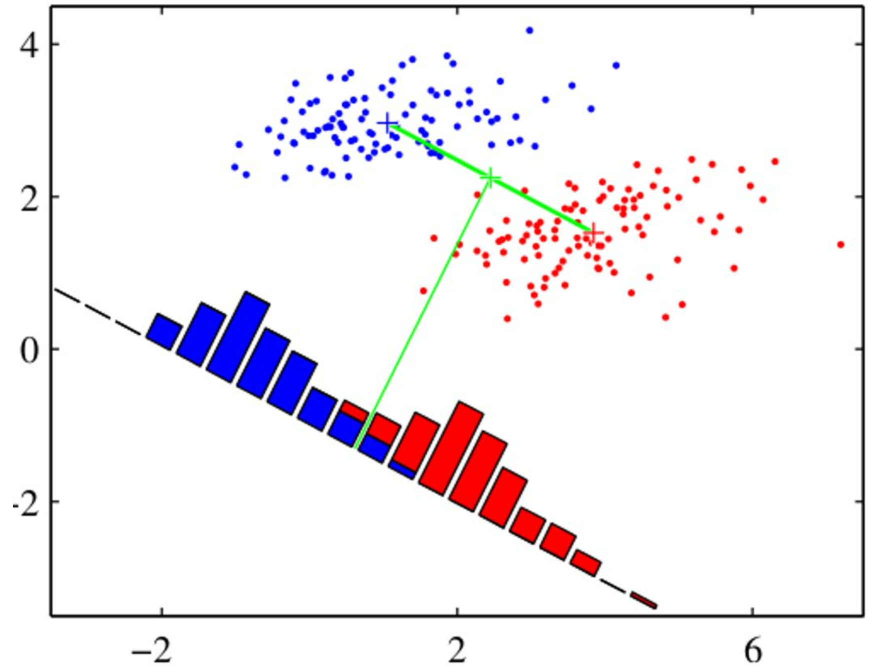
Классификация методом наименьших квадратов

- Линейной регрессией можно классифицировать! Но не нужно :)



Линейный дискриминант Фишера

- Пример, который проясняет геометрию происходящего
- В линейном случае мы хотим спроецировать точки в размерность 1 (на нормаль разделяющей гиперплоскости) так, чтобы в этой размерности 1 они хорошо разделялись
- Т.е. классификация — это такой метод радикального сокращения размерности
- Как добиться оптимальности в этом смысле? Первая идея — найти серединный перпендикуляр между центрами наборов точек
- Но она как-то не ахти работает
- А в чём проблема?



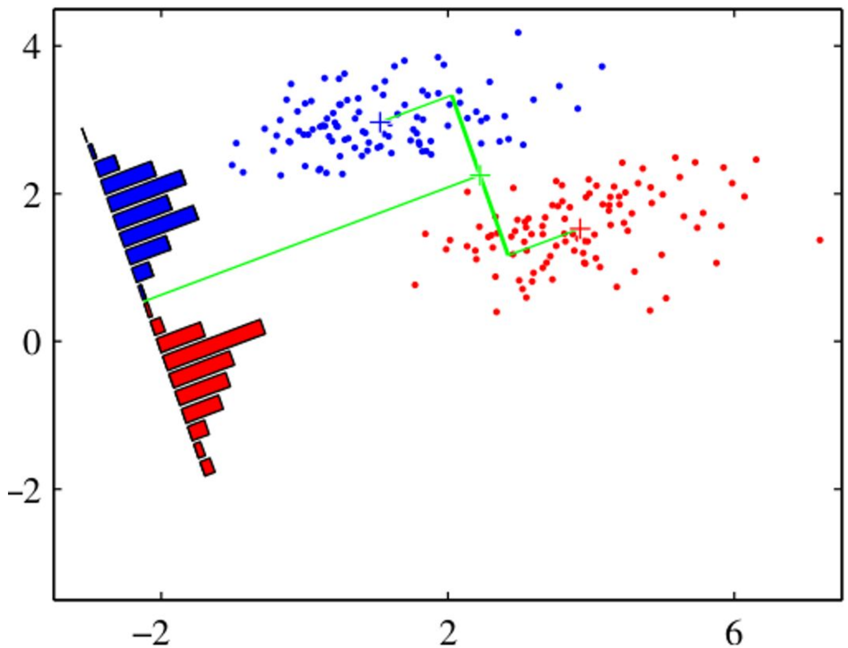
Линейный дискриминант Фишера

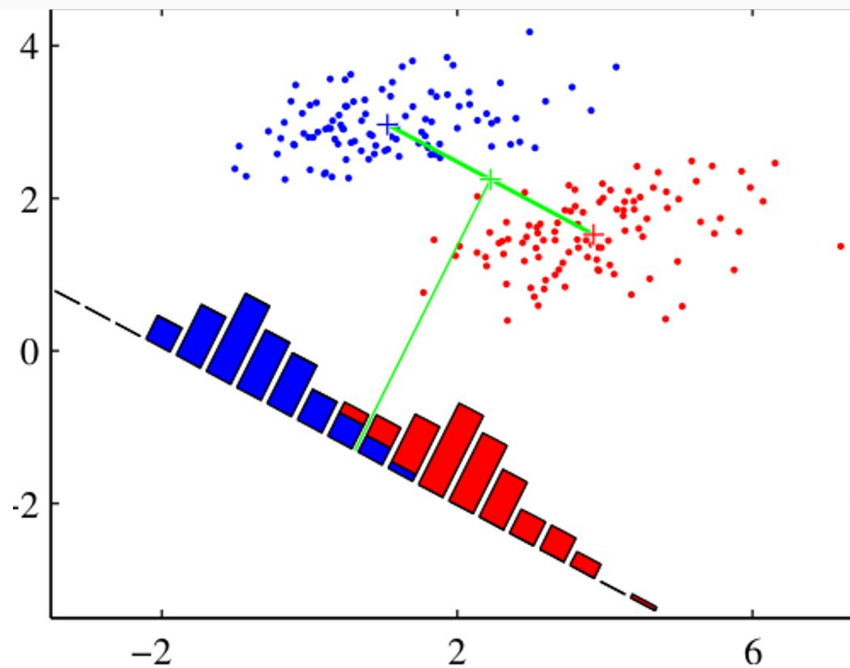
- “Хорошо разделялись” — это не только про большое расстояние, но и про маленькую дисперсию каждого класса! Отсюда *критерий Фишера*:

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} = \frac{\mathbf{w}^\top \mathbf{S}_B \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_W \mathbf{w}}, \text{ где}$$

$$\mathbf{S}_B = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^\top,$$

$$\mathbf{S}_W = \sum_{n \in C_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^\top + \sum_{n \in C_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^\top.$$





Порождающие модели и классификация

Оптимальный байесовский классификатор

- Классификация через порождающие модели: давайте каждому классу сопоставим плотность $p(\mathbf{x}|C_k)$, найдём априорные распределения $p(C_k)$, будем искать $p(C_k|\mathbf{x})$ по теореме Байеса
- Для двух классов:

$$p(C_1 \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C_1)p(C_1)}{p(\mathbf{x} \mid C_1)p(C_1) + p(\mathbf{x} \mid C_2)p(C_2)}.$$

- Для нескольких классов:

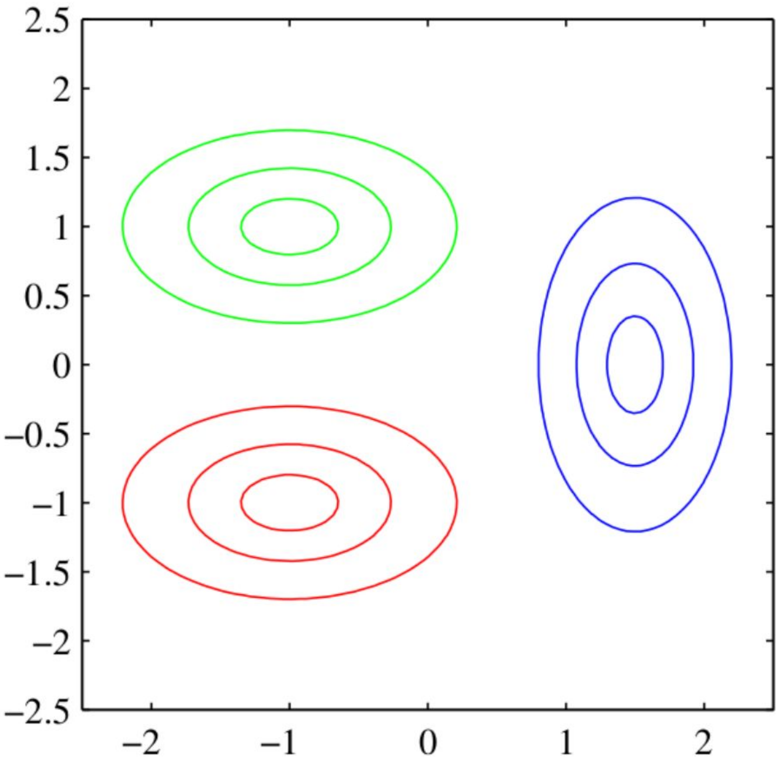
$$p(C_k \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C_k)p(C_k)}{\sum_j p(\mathbf{x} \mid C_j)p(C_j)} = \frac{e^{a_k}}{\sum_j e^{a_j}}.$$

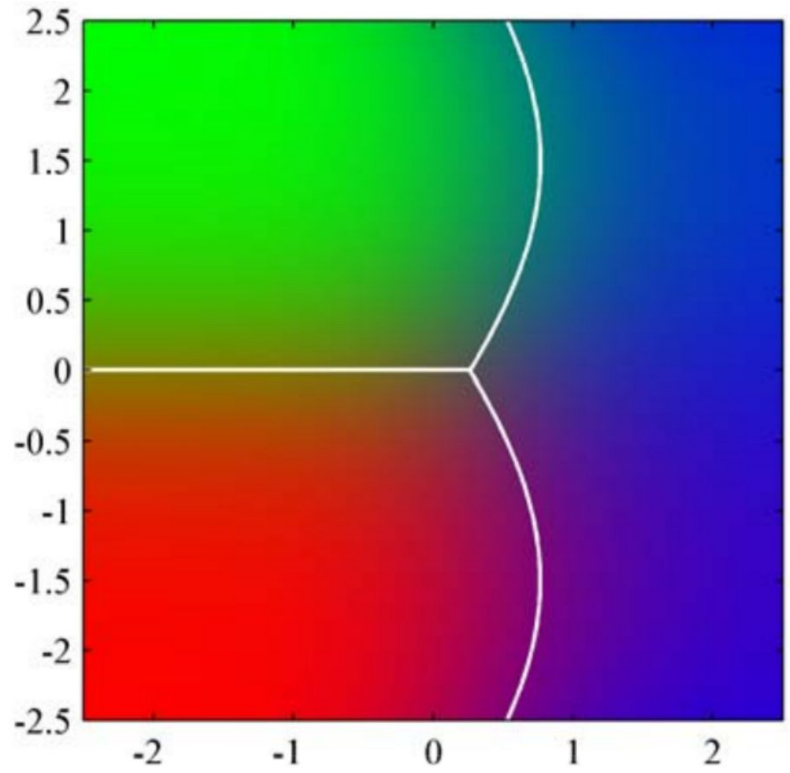
- Это и есть те самые логистический сигмоид и softmax, и к ним мы ещё вернёмся
- Это оптимальный байесовский классификатор, он будет работать всегда, для любых $p(\mathbf{x}|C_k)$

Оптимальный байесовский классификатор

- А если рассмотреть нормальные распределения в классах, то получатся LDA (linear discriminant analysis) и QDA (quadratic discriminant analysis)

$$p(\mathbf{x} \mid C_k) = N(\mathbf{x} \mid \mu_k, \Sigma)$$





Южный федеральный университет

Спасибо за внимание!

@SINECOR