# КЛАССИФИКАЦИЯ: ДИСКРИМИНАНТ ФИШЕРА

Сергей Николенко

НИУ ВШЭ — Санкт-Петербург
10 февраля 2017 г.

*Random facts*:

- 10 февраля 1355 г., в день памяти святой Схоластики (основательницы первого в Западной Европе женского монастыря), двое студентов Оксфордского университета, Уолтер Спрингхьюз и Роджер де Честерфилд, устроили скандал в таверне Свиндлсток из-за якобы низкого качества напитков; в результате начавшихся беспорядков погибли 63 студента и около 30 местных жителей.

# ВВЕДЕНИЕ В КЛАССИФИКАЦИЮ

# ЗАДАЧА КЛАССИФИКАЦИИ

- Теперь классификация: определить вектор $\mathbf{x}$ в один из $K$ классов $\mathcal{C}_k$.
- В итоге у нас так или иначе всё пространство разобьётся на эти классы.
- Т.е. на самом деле мы ищем *разделяющую поверхность* (decision surface, decision boundary).

# ЗАДАЧА КЛАССИФИКАЦИИ

- Как кодировать? Бинарная задача – очень естественно, переменная $t$, $t = 0$ соответствует $\mathcal{C}_1$, $t = 1$ соответствует $\mathcal{C}_2$.
- Оценку $t$ можно интерпретировать как вероятность (по крайней мере, мы постараемся, чтобы было можно).
- Если несколько классов – удобно $1$-of-$K$:

$$\mathbf{t} = (0, \dots, 0, 1, 0, \dots)^\top.$$

- Тоже можно интерпретировать как вероятности – или пропорционально им.

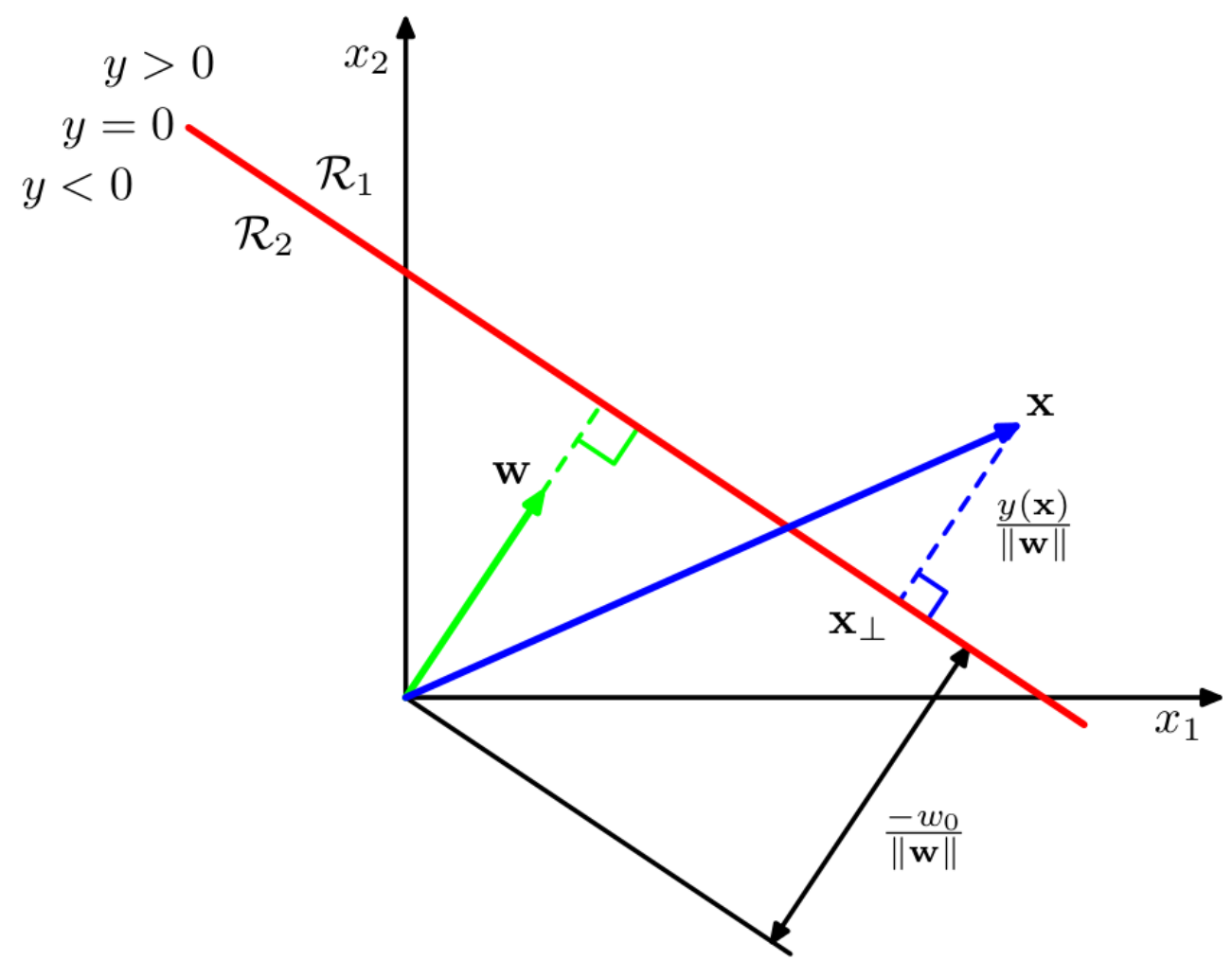
# РАЗДЕЛЯЮЩАЯ ГИПЕРПЛОСКОСТЬ

- Начнём с геометрии: рассмотрим линейную дискриминантную функцию

$$y(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + w_0.$$

- Это гиперплоскость, и $\mathbf{w}$ – нормаль к ней.
- Расстояние от начала координат до гиперплоскости равно $\frac{-w_0}{\|\mathbf{w}\|}$.
- $y(\mathbf{x})$ связано с расстоянием до гиперплоскости: $d = \frac{y(\mathbf{x})}{\|\mathbf{w}\|}$.

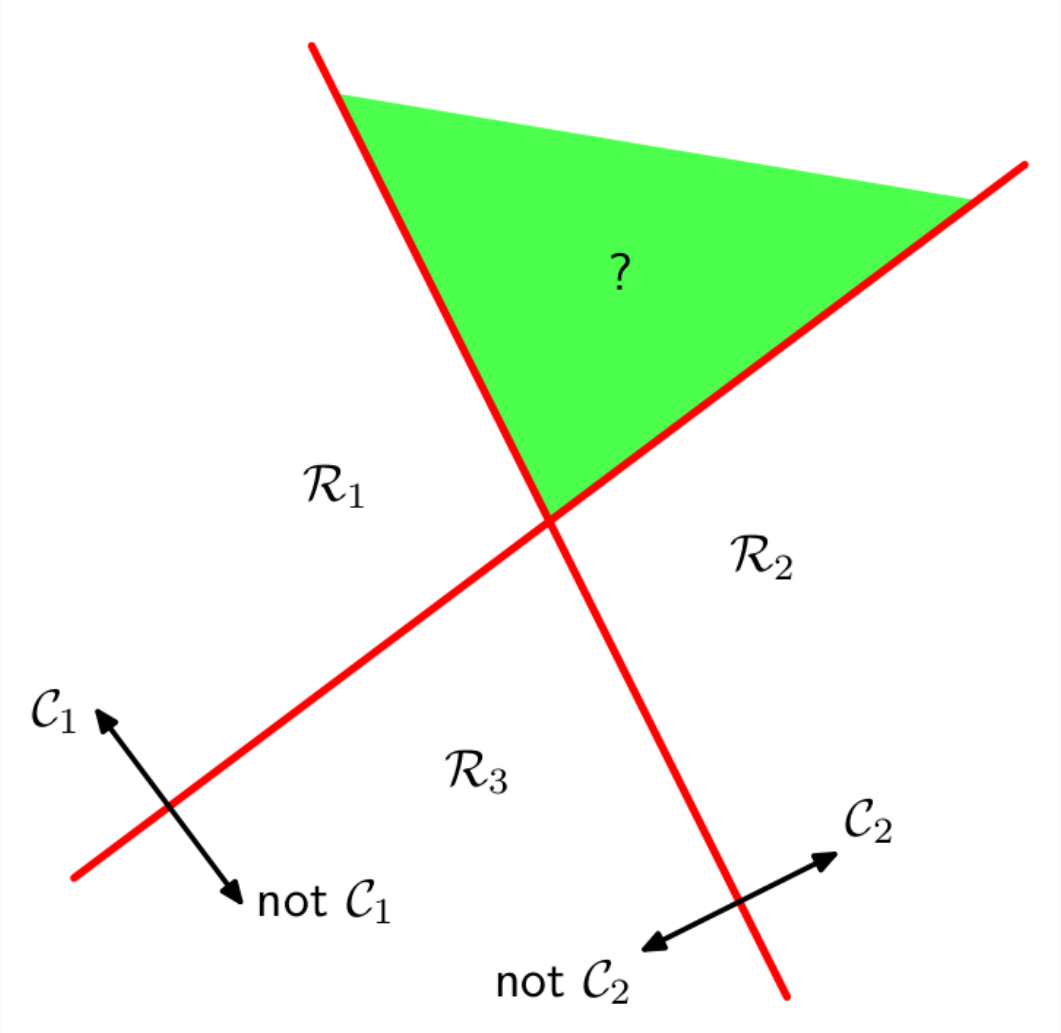
# РАЗДЕЛЯЮЩАЯ ГИПЕРПЛОСКОСТЬ



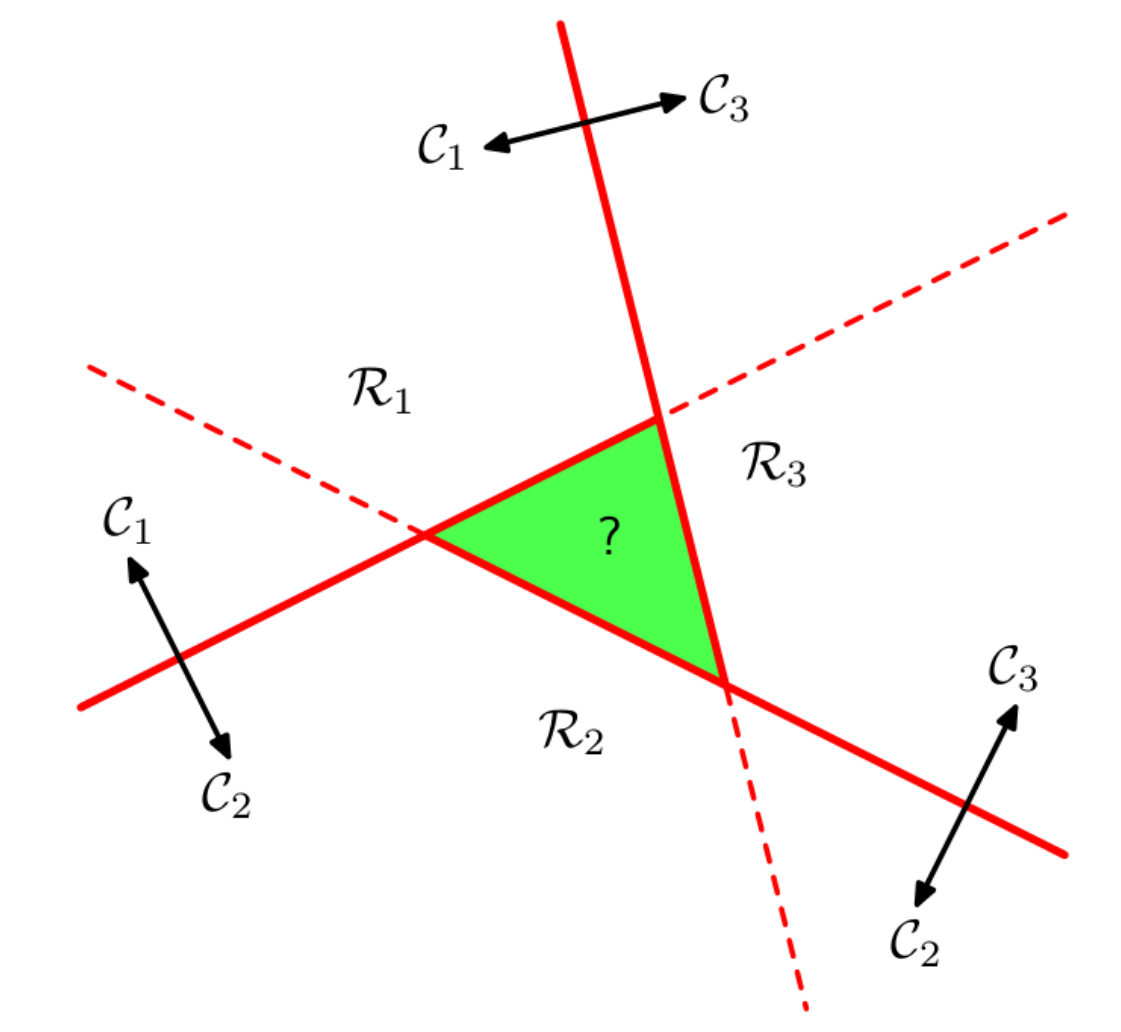
# НЕСКОЛЬКО КЛАССОВ

- С несколькими классами выходит незадача.
- Можно рассмотреть $K$ поверхностей вида «один против всех».
- Можно – $\binom{K}{2}$ поверхностей вида «каждый против каждого».
- Но всё это как-то нехорошо.

# НЕСКОЛЬКО КЛАССОВ



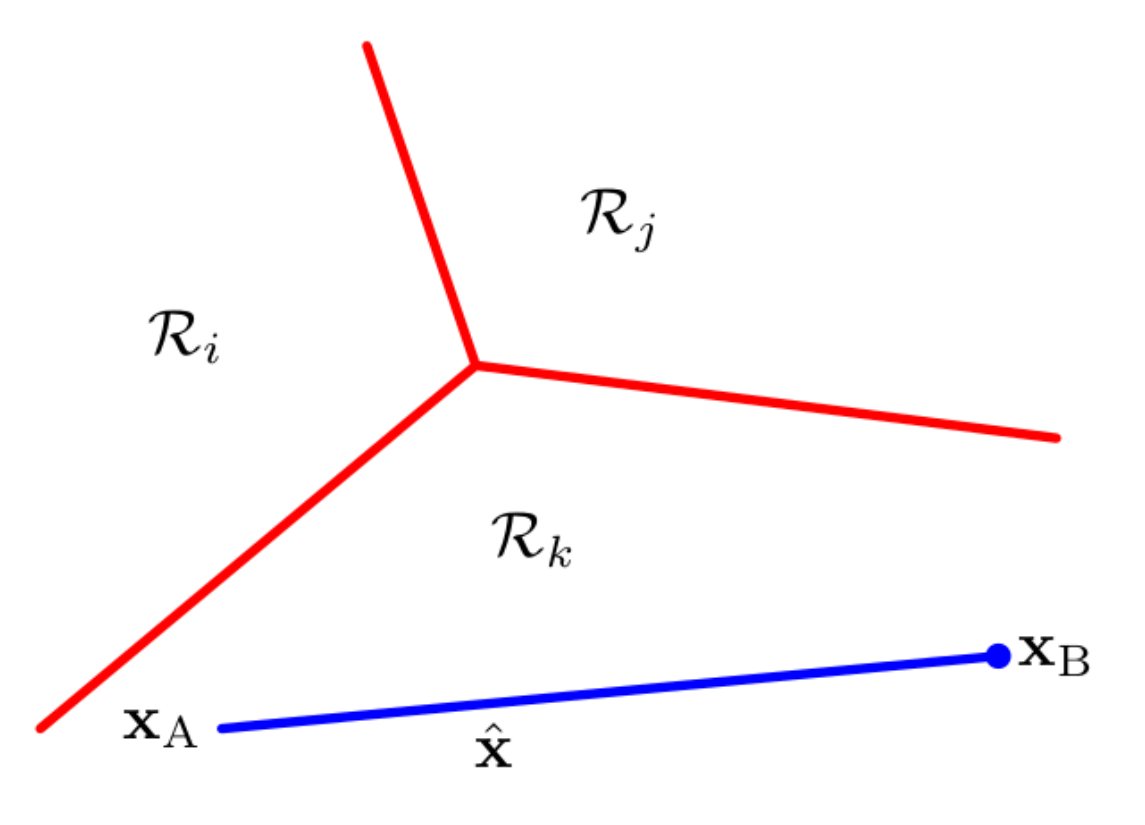
# НЕСКОЛЬКО КЛАССОВ



# НЕСКОЛЬКО КЛАССОВ

- Лучше рассмотреть единый дискриминант из $K$ линейных функций:
$$y_k(\mathbf{x}) = \mathbf{w}_k^\top \mathbf{x} + w_{k0}.$$
- Классифицировать в $\mathcal{C}_k$, если $y_k(\mathbf{x})$ – максимален.
- Тогда разделяющая поверхность между $\mathcal{C}_k$ и $\mathcal{C}_j$ будет гиперплоскостью вида $y_k(\mathbf{x}) = y_j(\mathbf{x})$:
$$\left(\mathbf{w}_k - \mathbf{w}_j\right)^\top \mathbf{x} + \left(w_{k0} - w_{j0}\right).$$

# НЕСКОЛЬКО КЛАССОВ




**Упражнение.** Докажите, что области, соответствующие классам, при таком подходе всегда односвязные и выпуклые.

# МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

- Мы снова можем воспользоваться методом наименьших квадратов: запишем $y_k(\mathbf{x}) = \mathbf{w}_k^\top \mathbf{x} + w_{k0}$ вместе (спрятав свободный член) как

$$\mathbf{y}(\mathbf{x}) = \mathbf{W}^\top \mathbf{x}.$$

- Можно найти $\mathbf{W}$, оптимизируя сумму квадратов; функция ошибки:

$$E_D(\mathbf{W}) = \frac{1}{2}\mathrm{Tr}\left[(\mathbf{X}\mathbf{W} - \mathbf{T})^\top (\mathbf{X}\mathbf{W} - \mathbf{T})\right].$$

- Берём производную, решаем...

# МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

- ...получается привычное

$$\mathbf{W} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{T} = \mathbf{X}^\dagger \mathbf{T},$$

  где $\mathbf{X}^\dagger$ – псевдообратная Мура-Пенроуза.

- Теперь можно найти и дискриминантную функцию:

$$\mathbf{y}(\mathbf{x}) = \mathbf{W}^\top \mathbf{x} = \mathbf{T}^\top \left(\mathbf{X}^\dagger\right)^\top \mathbf{x}.$$

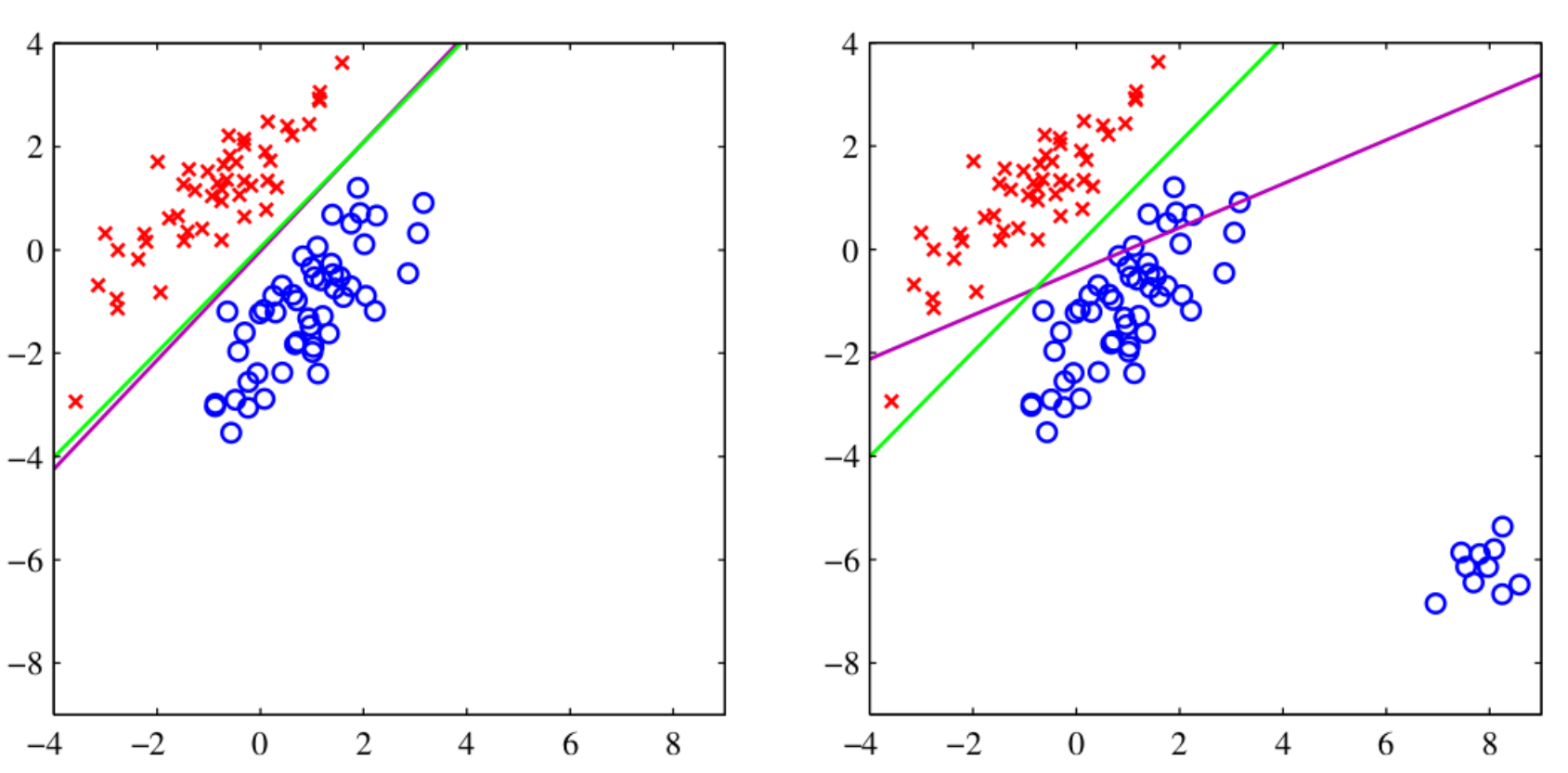
# МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

- Это решение сохраняет линейность.

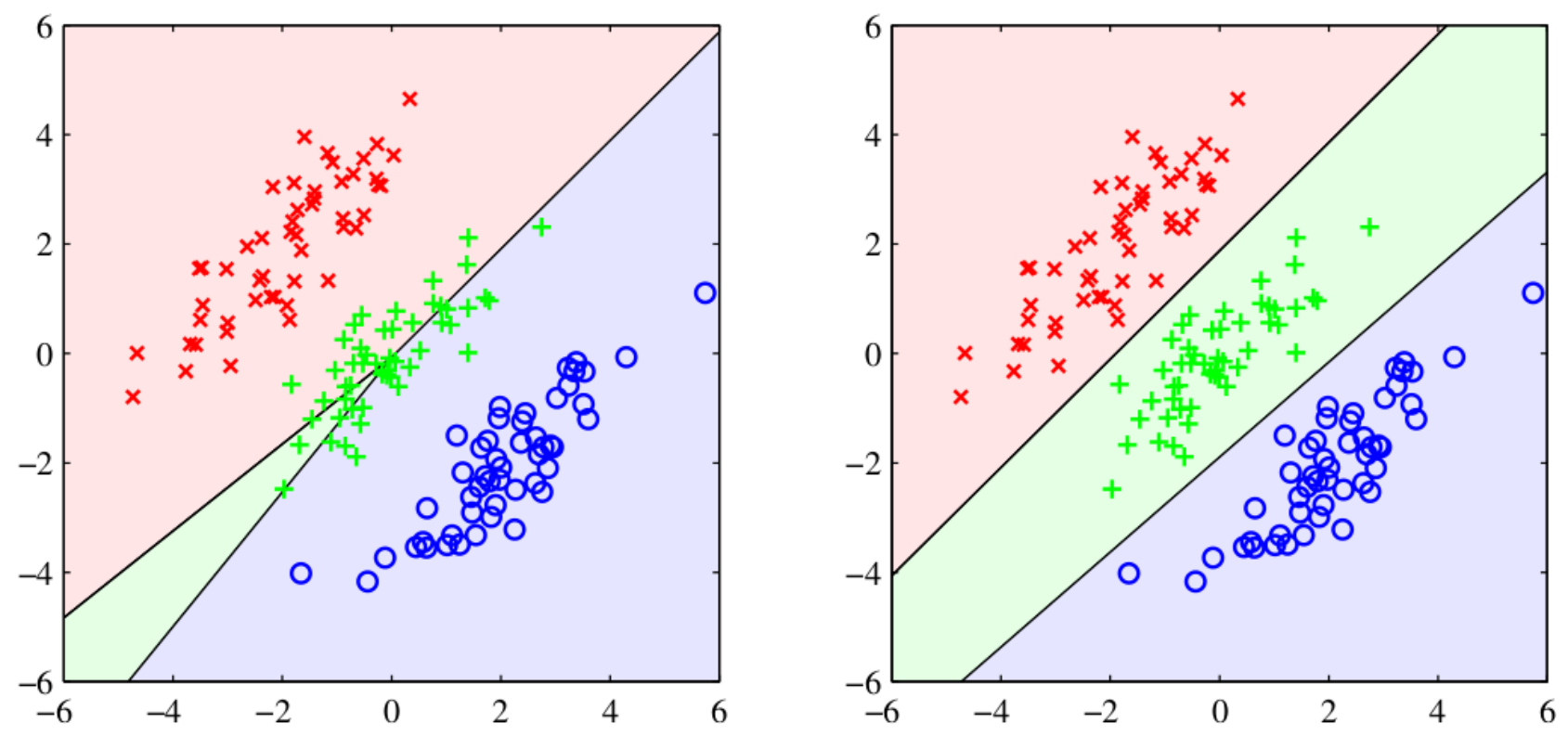
**Упражнение.** Докажите, что в схеме кодирования $1$-of-$K$ предсказания $y_k(\mathbf{x})$ для разных классов при любом $\mathbf{x}$ будут давать в сумме $1$. Почему они всё-таки не будут разумными оценками вероятностей?

- Проблемы наименьших квадратов:
  - outliers плохо обрабатываются;
  - «слишком правильные» предсказания добавляют штраф.

# ПРОБЛЕМЫ НАИМЕНЬШИХ КВАДРАТОВ



# ПРОБЛЕМЫ НАИМЕНЬШИХ КВАДРАТОВ



# ПРОБЛЕМЫ НАИМЕНЬШИХ КВАДРАТОВ

- Почему так? Почему наименьшие квадраты так плохо работают?

# ПРОБЛЕМЫ НАИМЕНЬШИХ КВАДРАТОВ

- Почему так? Почему наименьшие квадраты так плохо работают?
- Они предполагают гауссовское распределение ошибки.
- Но, конечно, распределение у бинарных векторов далеко не гауссово.

# ЛИНЕЙНЫЙ ДИСКРИМИНАНТ ФИШЕРА

- Другой взгляд на классификацию: 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.
- Давайте посмотрим на классификацию с этих позиций и попробуем добиться оптимальности в каком-то смысле.

# ЛИНЕЙНЫЙ ДИСКРИМИНАНТ ФИШЕРА

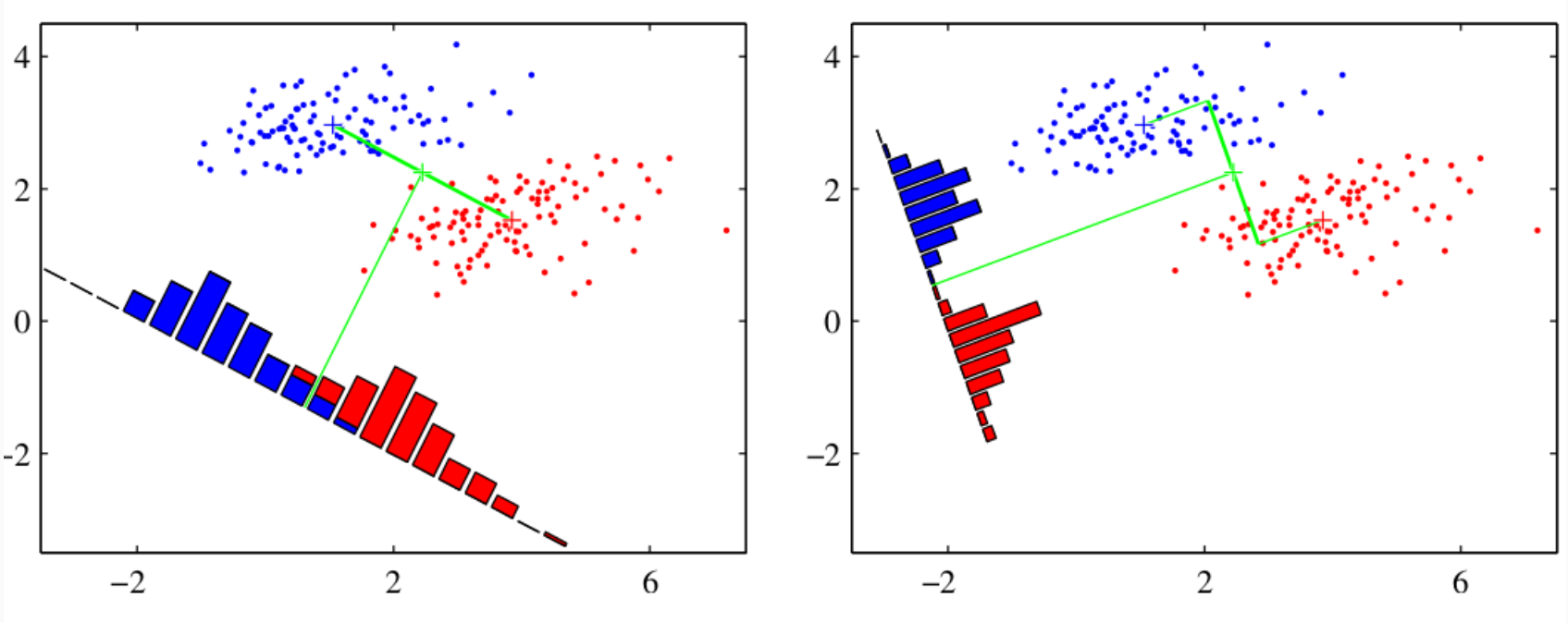
- Рассмотрим два класса $\mathcal{C}_1$ и $\mathcal{C}_2$ с $N_1$ и $N_2$ точками.
- Первая идея – надо найти серединный перпендикуляр между центрами кластеров

$$\mathbf{m}_1 = \frac{1}{N_1} \sum_{\mathcal{C}_1} \mathbf{x}, \text{ и } \mathbf{m}_2 = \frac{1}{N_2} \sum_{\mathcal{C}_2} \mathbf{x},$$

т.е. максимизировать $\mathbf{w}^\top (\mathbf{m}_2 - \mathbf{m}_1)$.

- Надо ещё добавить ограничение $\|\mathbf{w}\| = 1$, но всё равно не ахти как работает.

# ЛИНЕЙНЫЙ ДИСКРИМИНАНТ ФИШЕРА




Чем левая картинка хуже правой?

# ЛИНЕЙНЫЙ ДИСКРИМИНАНТ ФИШЕРА

- Слева больше дисперсия каждого кластера.
- Идея: минимизировать перекрытие классов, оптимизируя и проекцию расстояния, и дисперсию.
- Выборочные дисперсии в проекции: для $y_n = \mathbf{w}^\top \mathbf{x}_n$

$$s_1 = \sum_{n \in \mathcal{C}_1} (y_n - m_1)^2 \text{ и } s_1 = \sum_{n \in \mathcal{C}_2} (y_n - m_2)^2 .$$

# ЛИНЕЙНЫЙ ДИСКРИМИНАНТ ФИШЕРА

- Критерий Фишера:

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

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

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

(between-class covariance и within-class covariance).

- Дифференцируя по $\mathbf{w}$...

# ЛИНЕЙНЫЙ ДИСКРИМИНАНТ ФИШЕРА

- ...получим, что $J(\mathbf{w})$ максимален при

$$\left(\mathbf{w}^\top \mathbf{S}_B \mathbf{w}\right) \mathbf{S}_W \mathbf{w} = \left(\mathbf{w}^\top \mathbf{S}_W \mathbf{w}\right) \mathbf{S}_B \mathbf{w}.$$

- Т.к. $\mathbf{S}_B = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^\top$, $\mathbf{S}_B \mathbf{w}$ всё равно будет в направлении $\mathbf{m}_2 - \mathbf{m}_1$, а длина $\mathbf{w}$ нас не интересует.
- Поэтому получается

$$\mathbf{w} \propto \mathbf{S}_W^{-1} (\mathbf{m}_2 - \mathbf{m}_1).$$

- В итоге мы выбрали направление проекции, и осталось только разделить данные на этой проекции.

# ЛИНЕЙНЫЙ ДИСКРИМИНАНТ ФИШЕРА

- Любопытно, что дискриминант Фишера тоже можно получить из наименьших квадратов.
- Давайте для класса $\mathcal{C}_1$ выберем целевое значение $\frac{N_1+N_2}{N_1}$, а для класса $\mathcal{C}_2$ возьмём $-\frac{N_1+N_2}{N_2}$.

**Упражнение.** Докажите, что при таких целевых значениях наименьшие квадраты – это дискриминант Фишера.

# ЛИНЕЙНЫЙ ДИСКРИМИНАНТ ФИШЕРА

- А что будет с несколькими классами? Рассмотрим $\mathbf{y} = \mathbf{W}^\top \mathbf{x}$, обобщим внутреннюю дисперсию как

$$\mathbf{S}_W = \sum_{k=1}^{K} \mathbf{S}_k = \sum_{k=1}^{K} \sum_{n \in \mathcal{C}_k} (\mathbf{x}_n - \mathbf{m}_k)(\mathbf{x}_n - \mathbf{m}_k)^\top .$$

- Чтобы обобщить внешнюю (межклассовую) дисперсию, просто возьмём остаток полной дисперсии

$$\mathbf{S}_T = \sum_n (\mathbf{x}_n - \mathbf{m})(\mathbf{x}_n - \mathbf{m})^\top ,$$

$$\mathbf{S}_B = \mathbf{S}_T - \mathbf{S}_W .$$

# ЛИНЕЙНЫЙ ДИСКРИМИНАНТ ФИШЕРА

- Обобщить критерий можно разными способами, например:

$$J(\mathbf{W}) = \mathrm{Tr}\left[\mathbf{s}_W^{-1}\mathbf{s}_B\right],$$

где $\mathbf{s}$ – ковариации в пространстве проекций на $\mathbf{y}$:

$$\mathbf{s}_W = \sum_{k=1}^{K}\sum_{n\in\mathcal{C}_k}\left(\mathbf{y}_n - \mu_k\right)\left(\mathbf{y}_n - \mu_k\right)^\top,$$

$$\mathbf{s}_B = \sum_{k=1}^{K} N_k\left(\mu_k - \mu\right)\left(\mu_k - \mu\right)^\top,$$

где $\mu_k = \frac{1}{N_k}\sum_{n\in\mathcal{C}_k}\mathbf{y}_n$.

# СПАСИБО!

Спасибо за внимание!