ВВЕДЕНИЕ В КЛАССИФИКАЦИЮ

Сергей Николенко

СПбГУ — Санкт-Петербург
05 октября 2020 г.

Random facts:

- 5 октября — Всемирный день учителей ООН и День работников уголовного розыска России; именно 5 октября 1918 года НКВД РСФСР создал Центророзыск
- 5 октября — день памяти Мурдока Кульдея, последнего из бардов, который жил возле озера в Аргильшире и почитается как отшельник католической церкви
- 5 октября 1921 г. в Лондоне по инициативе Кэтрин Доусон-Скотт и Джона Голсуорси был учреждён ПЕН-клуб (от слов poet, essayist и novelist)
- 5 октября 1952 г. на XIX съезде ВКП(б) была переименована в КПСС, а над Ленинградом столкнулись Ил-12 и ТС-62
- 5 октября 1962 г. в Великобритании вышел первый сингл *The Beatles* «Love Me Do» и первый фильм о Джеймсе Бонде, «Dr. No», а 5 октября 1969 г. на BBC вышел первый выпуск «Monty Python's Flying Circus»

Коронавирусный пример

Пример: линейная регрессия и коронавирус

- Пример: линейная регрессия и коронавирус
- Как можно было в начале эпидемии предсказать, что будет происходить дальше?
- Давайте посмотрим на пример и на результаты...

Введение в классификацию

Задача классификации

- Теперь классификация: определить вектор $\mathbf{x}$ в один из K классов 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 C_1, $t = 1$ соответствует 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.

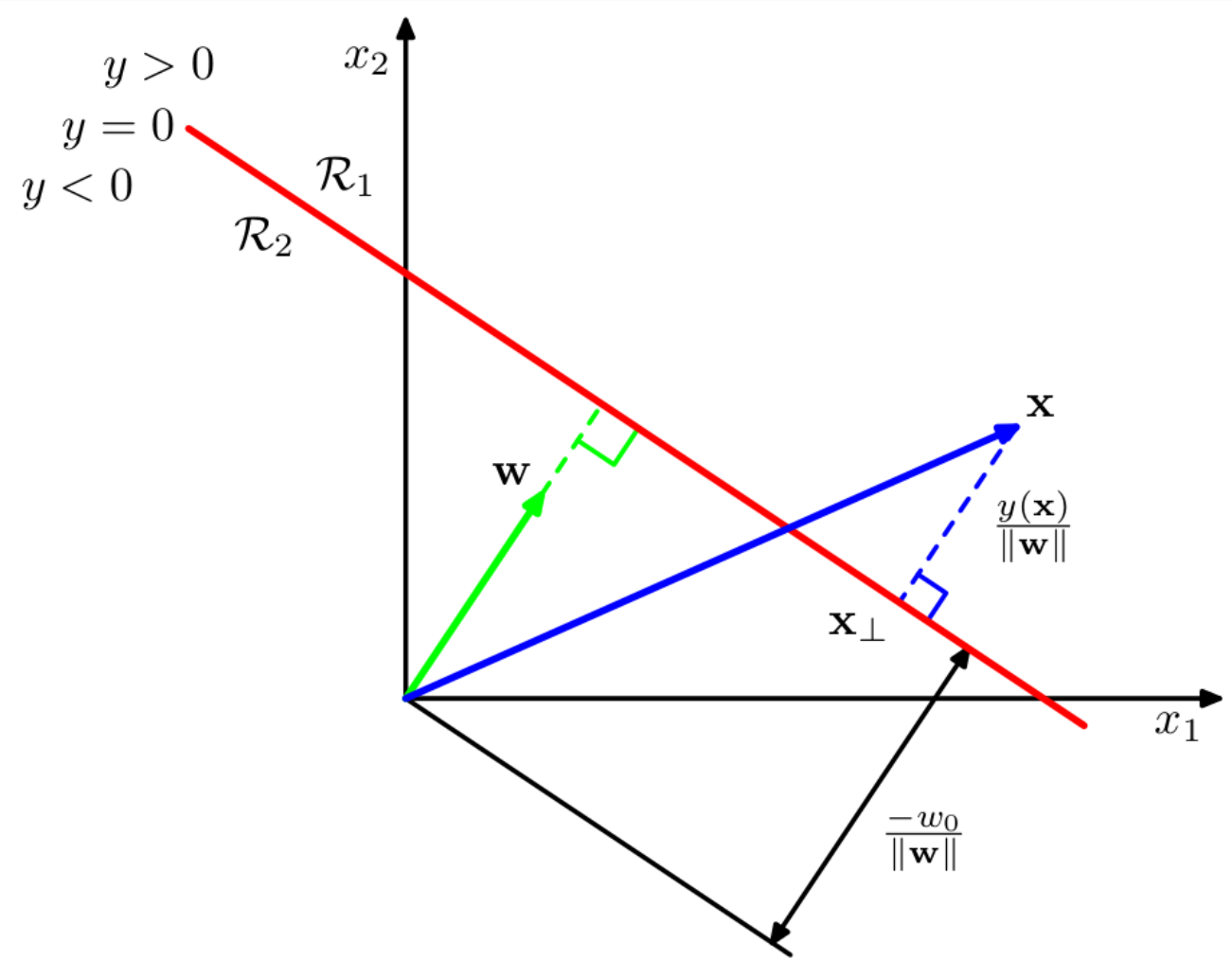
Разделяющая гиперплоскость

- 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: $d = \frac{y(\mathbf{x})}{\|\mathbf{w}\|}$.

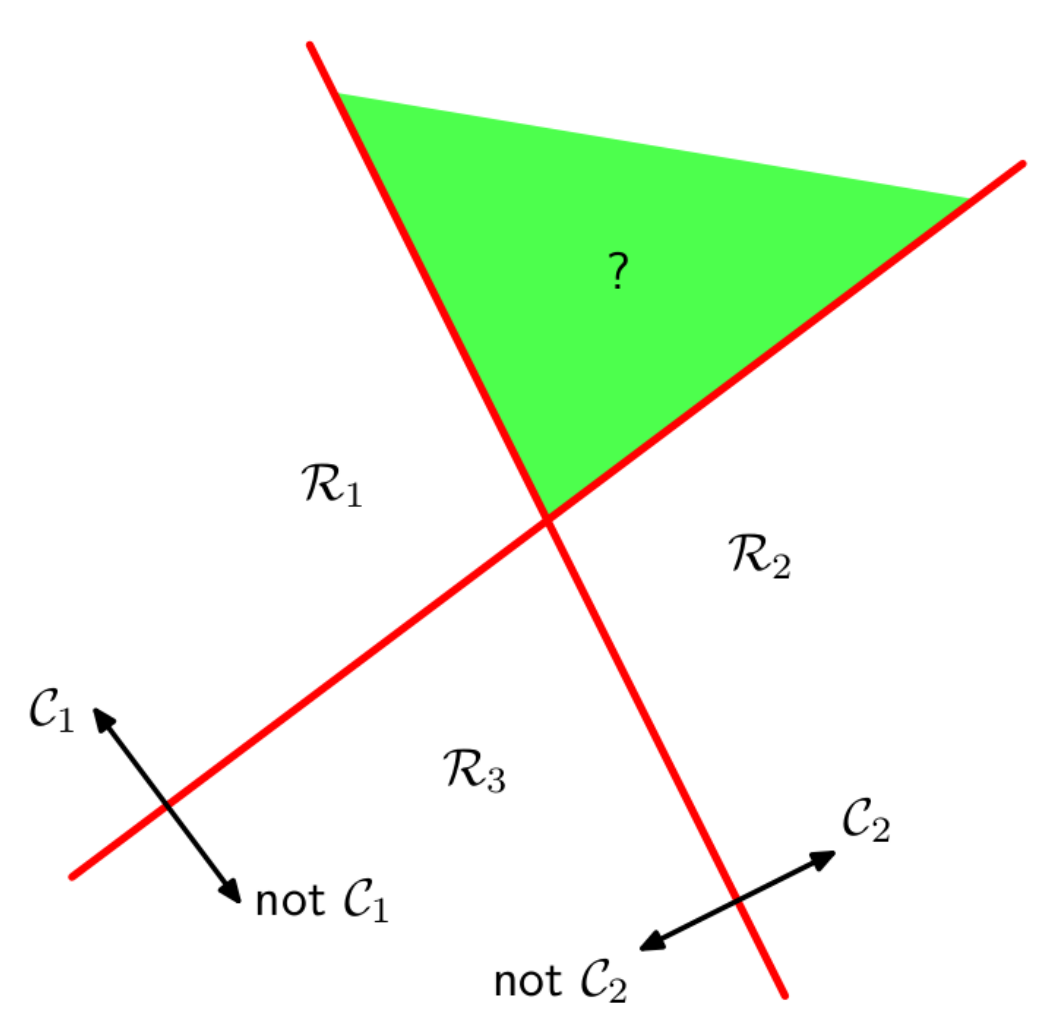
Разделяющая гиперплоскость



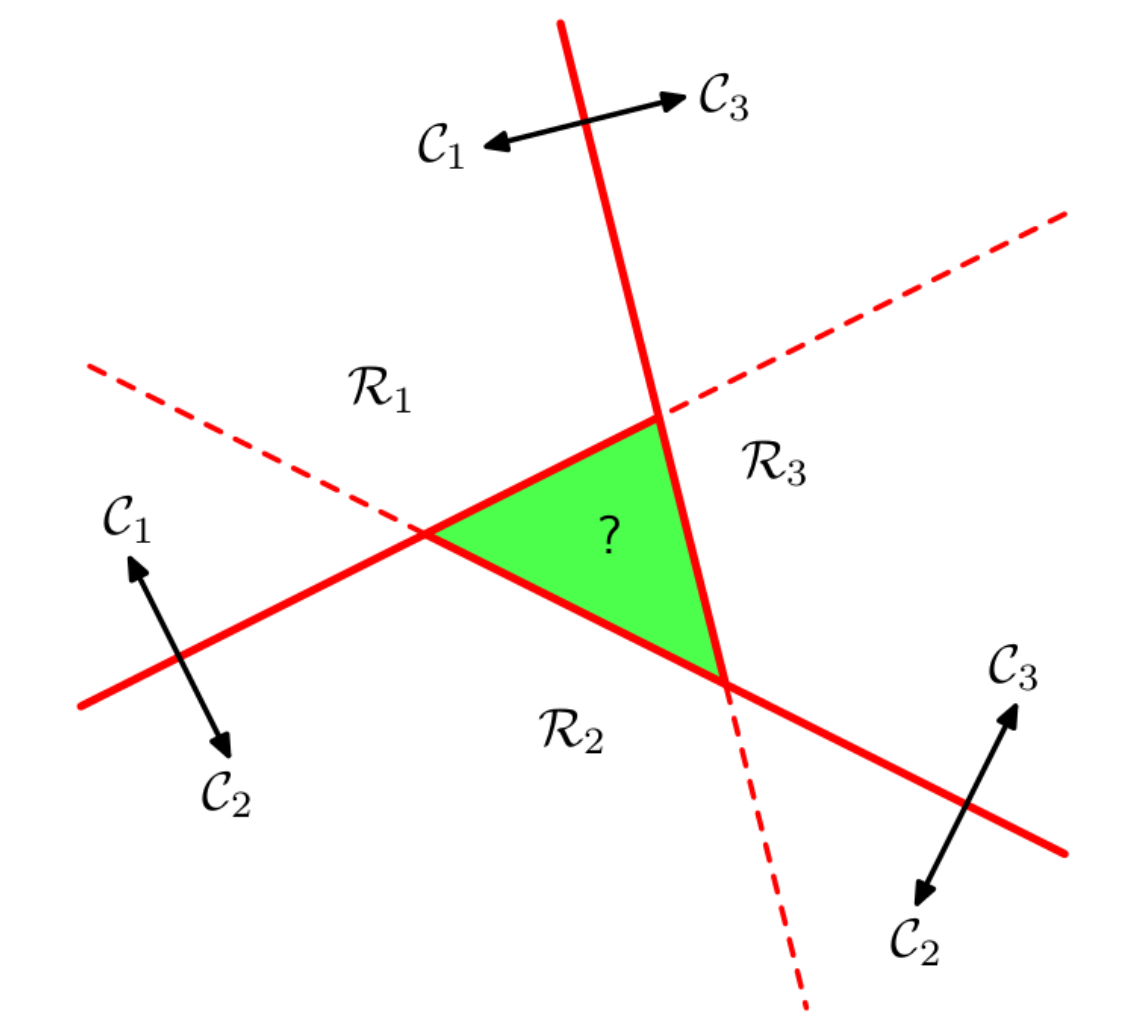
Несколько классов

- 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

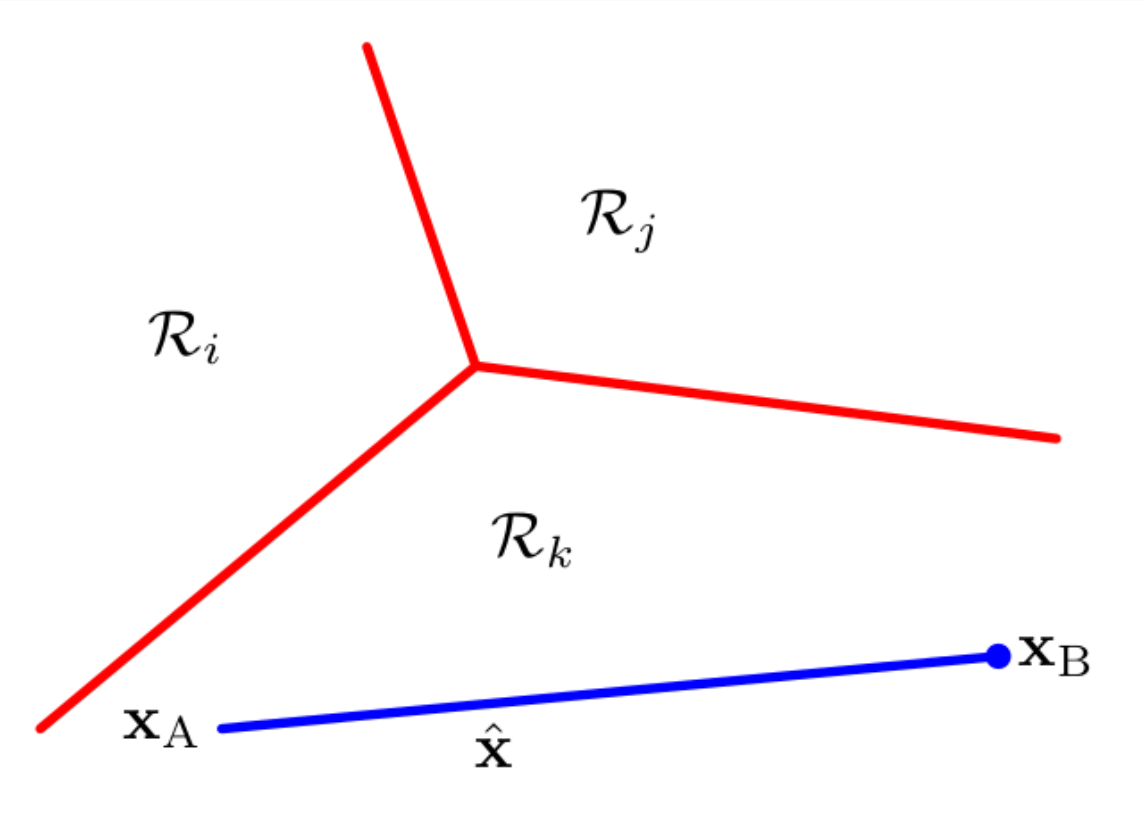
- Лучше рассмотреть единый дискриминант из K линейных функций:

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

- Классифицировать в C_k, если $y_k(\mathbf{x})$ – максимален.
- Тогда разделяющая поверхность между C_k и 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:

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

- 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:

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

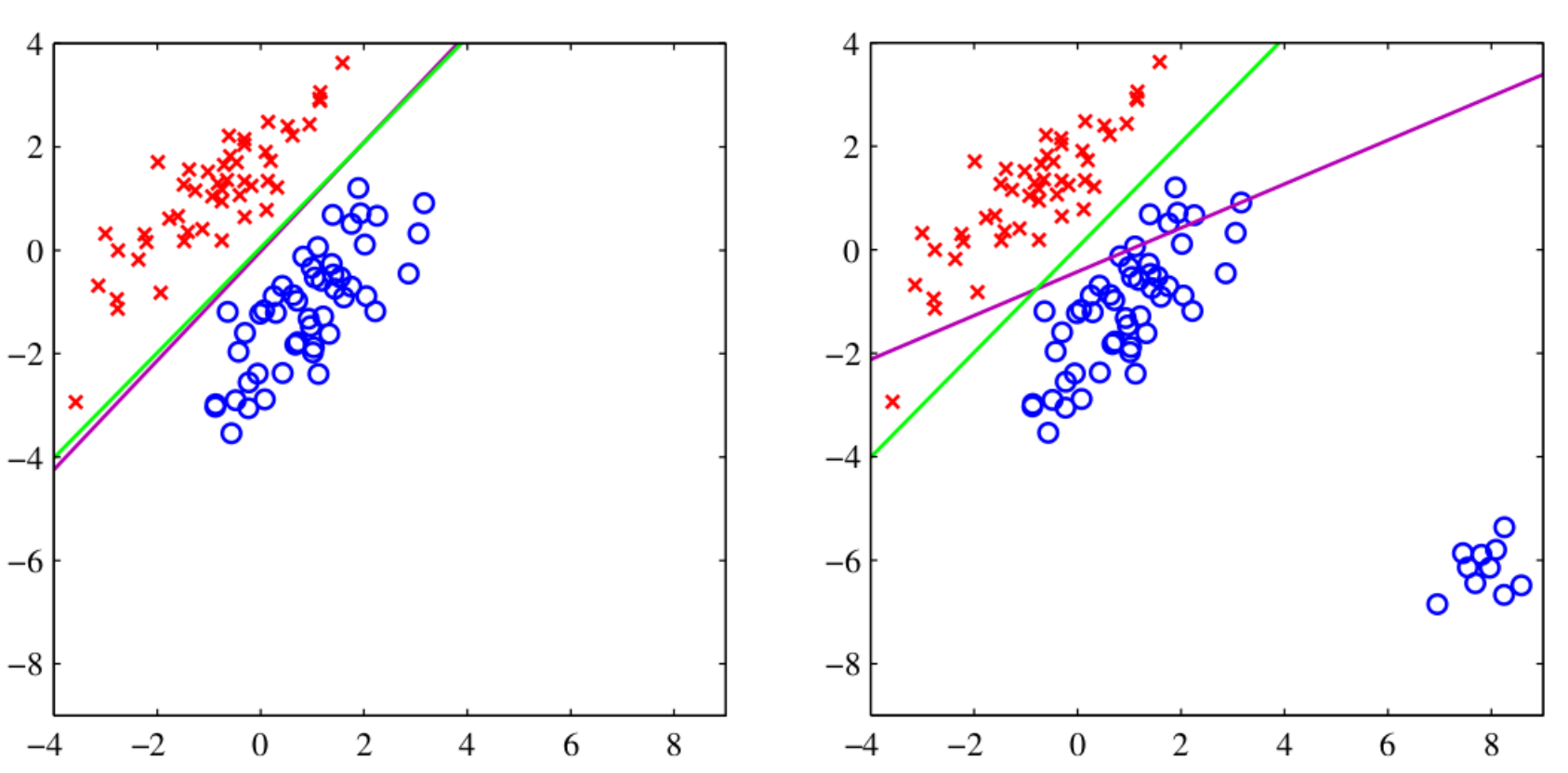
Метод наименьших квадратов

- Это решение сохраняет линейность.

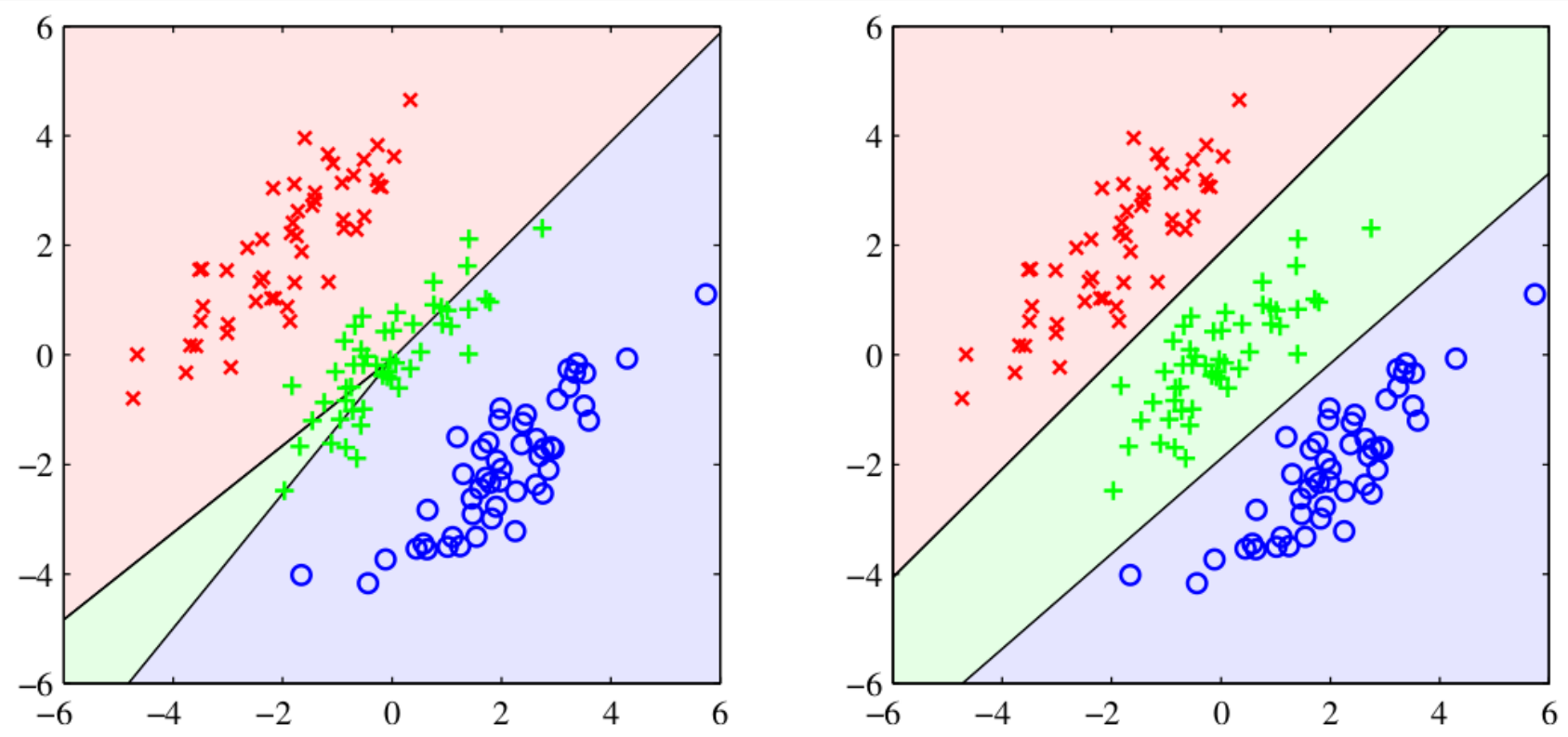
Упражнение. Докажите, что в схеме кодирования 1-of-K предсказания $y_k(\mathbf{x})$ для разных классов при любом $\mathbf{x}$ будут давать в сумме 1. Почему они всё-таки не будут разумными оценками вероятностей?

- 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

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

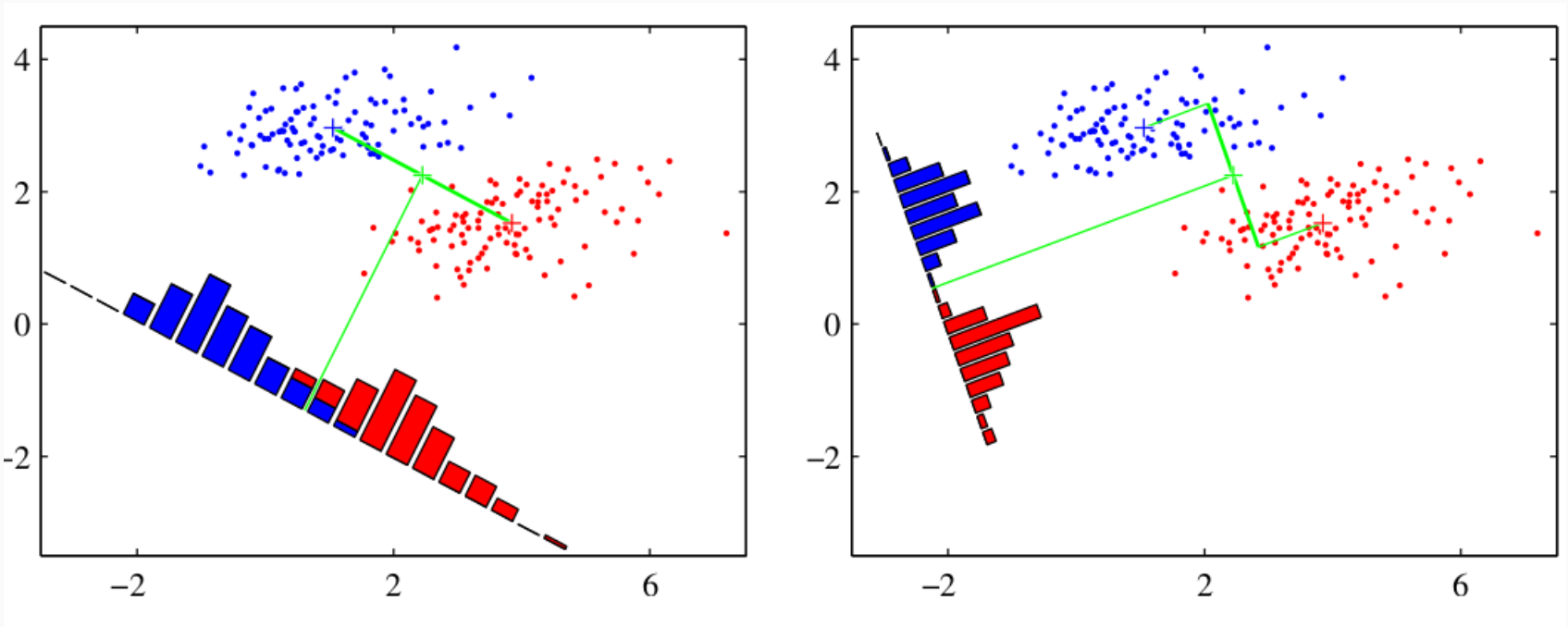
- Рассмотрим два класса C_1 и C_2 с N_1 и N_2 точками.
- Первая идея – надо найти серединный перпендикуляр между центрами кластеров

$$\mathbf{m}_1 = \frac{1}{N_1}\sum_{C_1}\mathbf{x}, \text{ и } \mathbf{m}_2 = \frac{1}{N_2}\sum_{C_2}\mathbf{x},$$

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 $y_n = \mathbf{w}^\top \mathbf{x}_n$

$$s_1 = \sum_{n \in C_1} (y_n - m_1)^2 \text{ и } s_1 = \sum_{n \in C_2} (y_n - m_2)^2.$$

Линейный дискриминант Фишера

- Критерий Фишера:

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

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

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

(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 C_1 выберем целевое значение $\frac{N_1+N_2}{N_1}$, а для класса 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

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

- 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 $\mathbf{y}$:

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

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

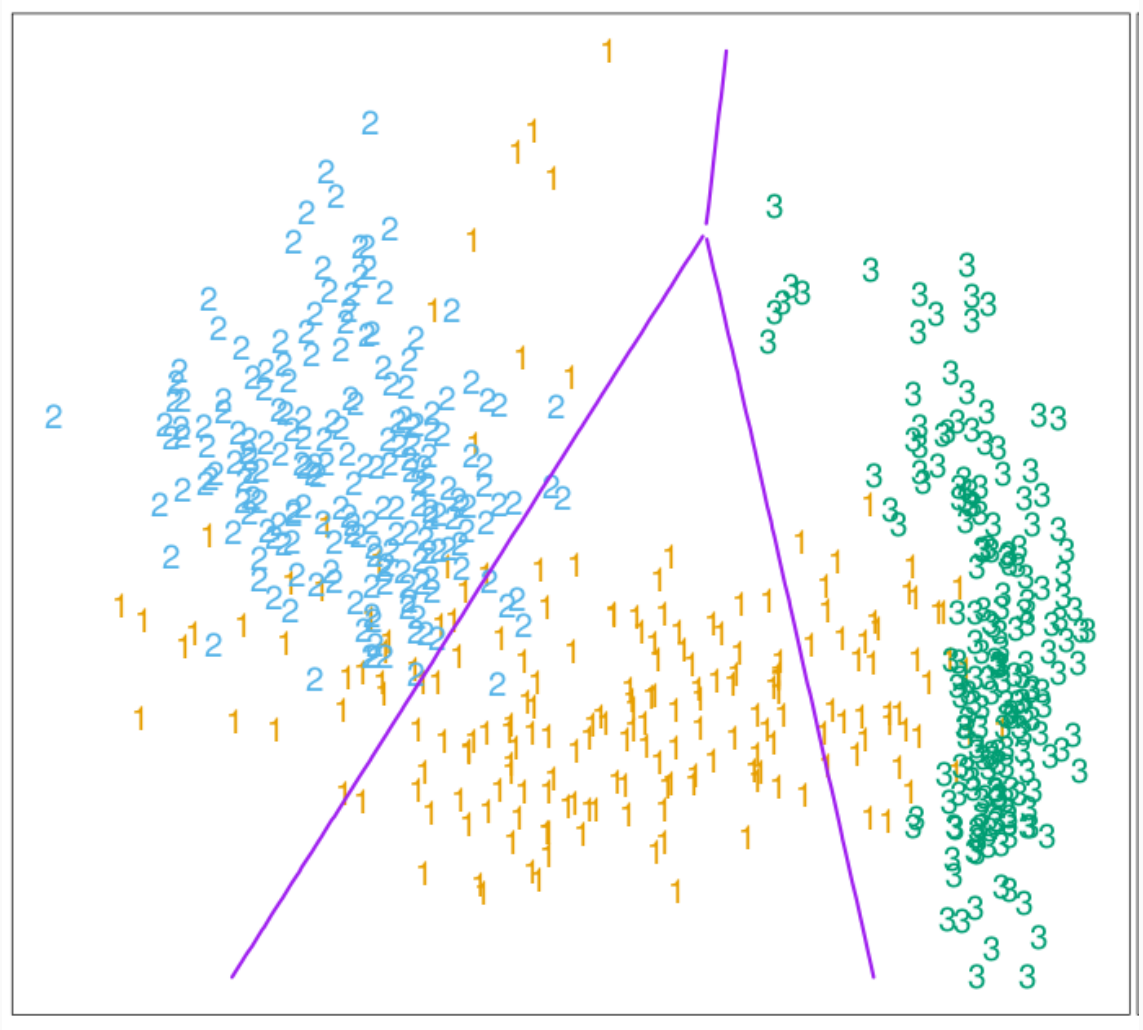
где $\mu_k = \frac{1}{N_k} \sum_{n \in C_k} \mathbf{y}_n$.

LDA и QDA

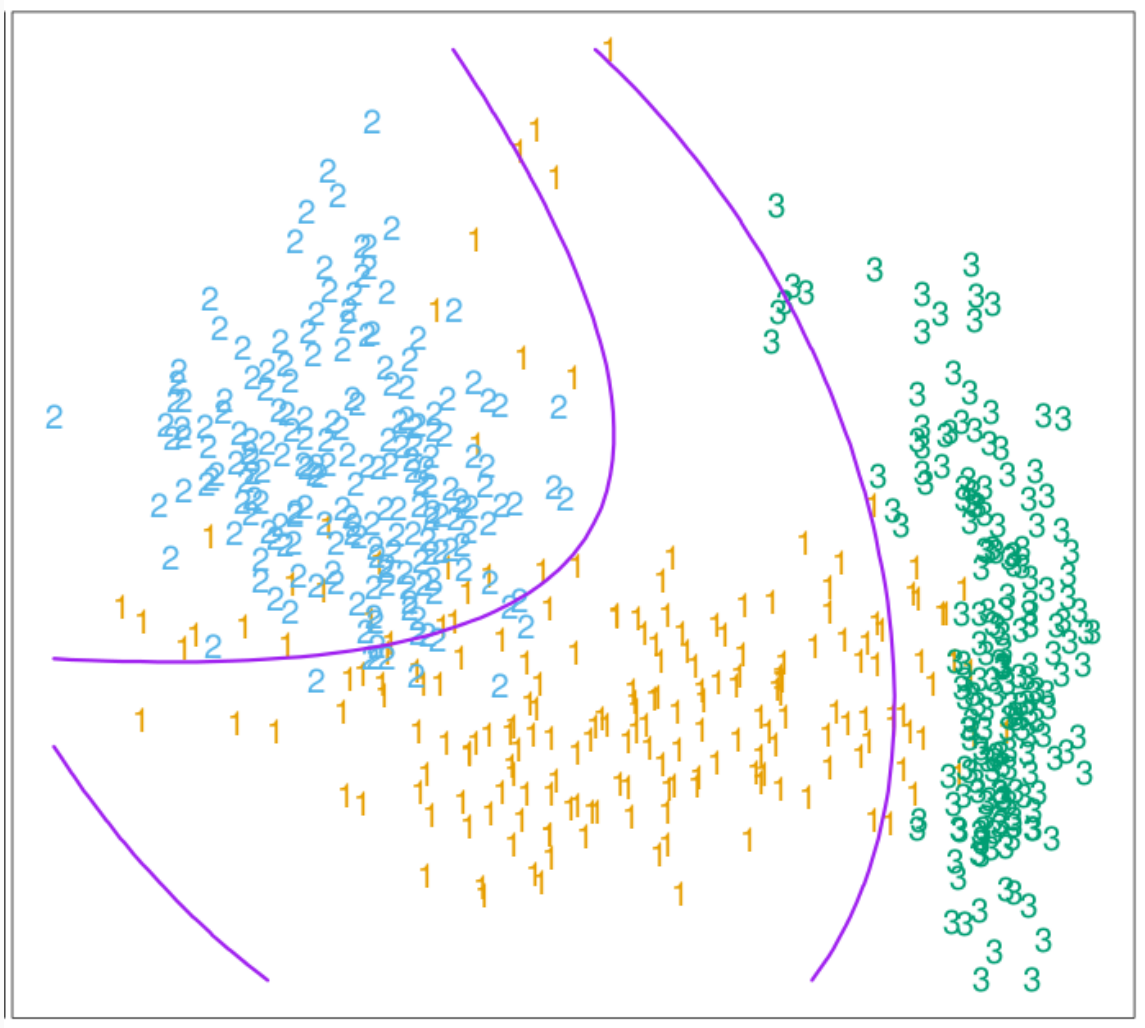
Нелинейные поверхности

- Мы учились проводить разделяющие гиперплоскости.
- Но как же нелинейные поверхности?
- Можно делать нелинейные из линейных, увеличивая размерность.

Нелинейные поверхности



Нелинейные поверхности



Генеративные модели

- Теперь классификация через генеративные модели: давайте каждому классу сопоставим плотность $p(\mathbf{x} \mid C_k)$, найдём априорные распределения $p(C_k)$, будем искать $p(C_k \mid \mathbf{x})$ по теореме Байеса.
- Для двух классов:

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

Генеративные модели

- Перепишем:

$$p(C_1 \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C_1)p(C_1)}{p(\mathbf{x} \mid C_1)p(C_1) + p(\mathbf{x} \mid C_2)p(C_2)} = \frac{1}{1 + e^{-a}} = \sigma(a),$$

где

$$a = \ln \frac{p(\mathbf{x} \mid C_1)p(C_1)}{p(\mathbf{x} \mid C_2)p(C_2)}, \qquad \sigma(a) = \frac{1}{1 + e^{-a}}.$$

- $\sigma(a)$ – *логистический сигмоид*:

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

- $\sigma(-a) = 1 - \sigma(a)$.
- $a = \ln\left(\frac{\sigma}{1-\sigma}\right)$ – *логит-функция*.

Упражнение. Докажите эти свойства.

Несколько классов

- В случае нескольких классов получится

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

- Здесь $a_k = \ln p(\mathbf{x} \mid C_k)p(C_k)$.
- $\frac{e^{a_k}}{\sum_j e^{a_j}}$ – нормализованная экспонента, или softmax-функция (сглаженный максимум).

Пример

- Давайте рассмотрим гауссовы распределения для классов:

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

- Сначала пусть Σ у всех одинаковые, а классов всего два.
- Посчитаем логистический сигмоид...

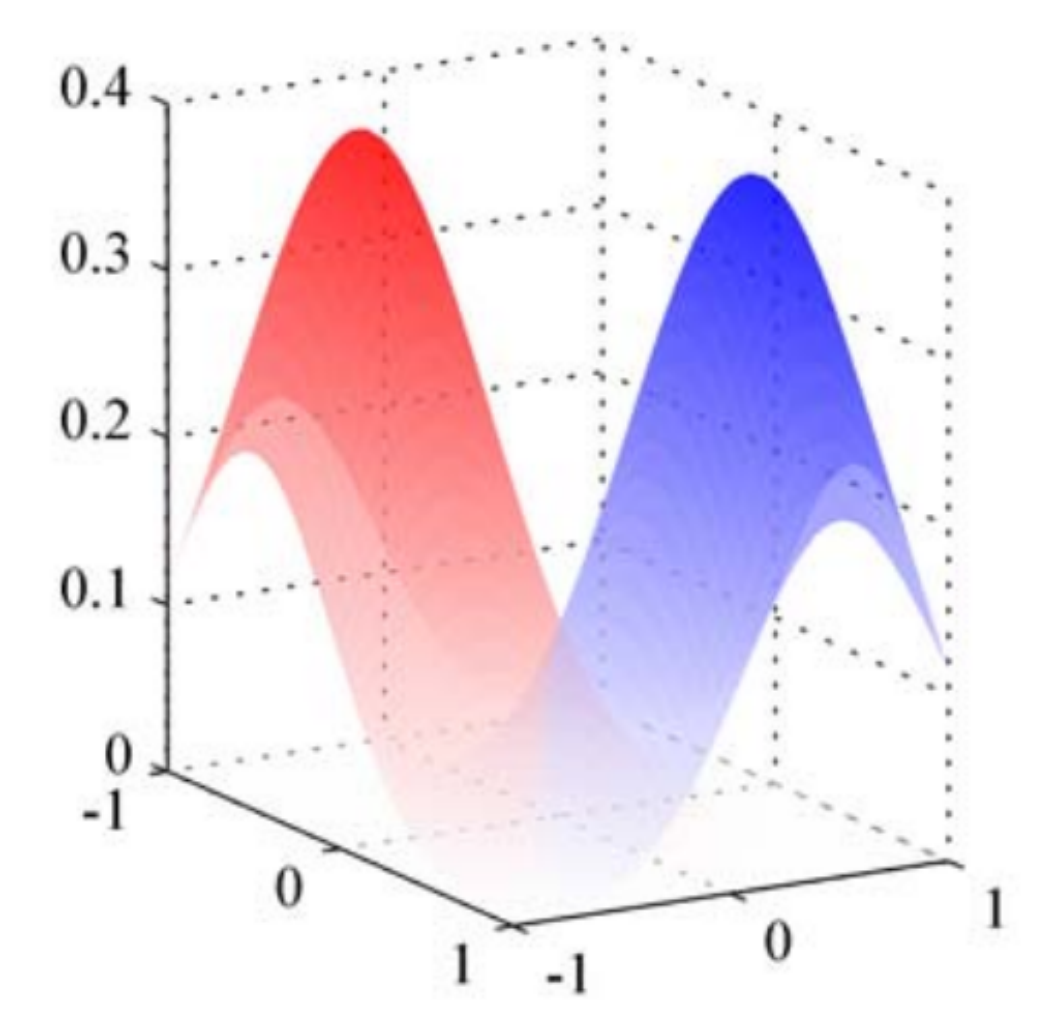
Пример

- ...получится

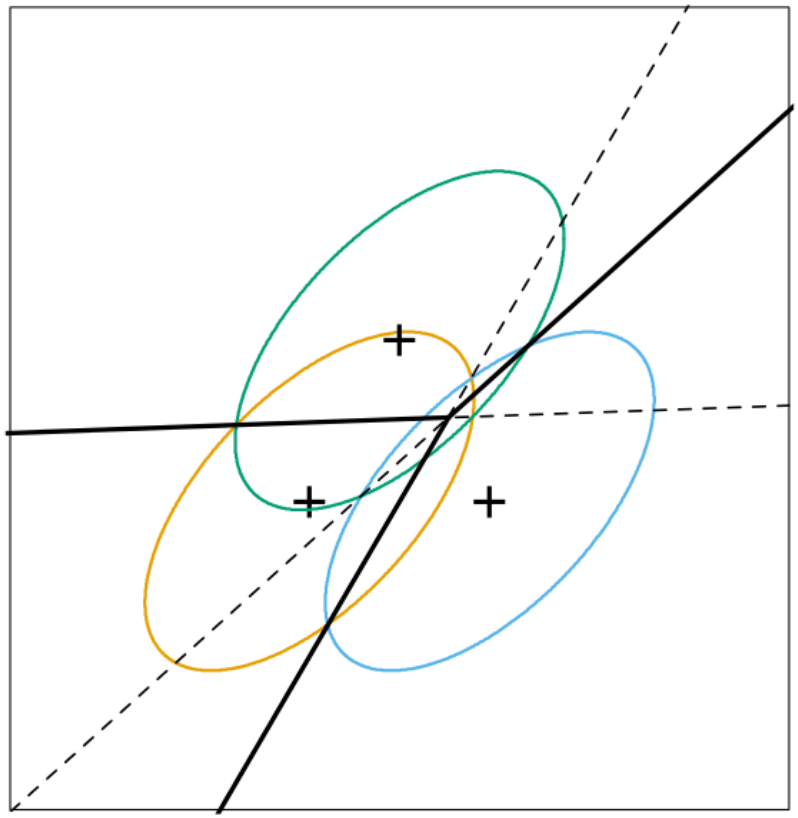
$$p(C_1 \mid \mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + w_0), \text{ где}$$
$$\mathbf{w} = \Sigma^{-1} \left(\mu_1 - \mu_2\right),$$
$$w_0 = -\frac{1}{2}\mu_1^\top \Sigma^{-1} \mu_1 + \frac{1}{2}\mu_2^\top \Sigma^{-1} \mu_2 + \ln \frac{p(C_1)}{p(C_2)}.$$

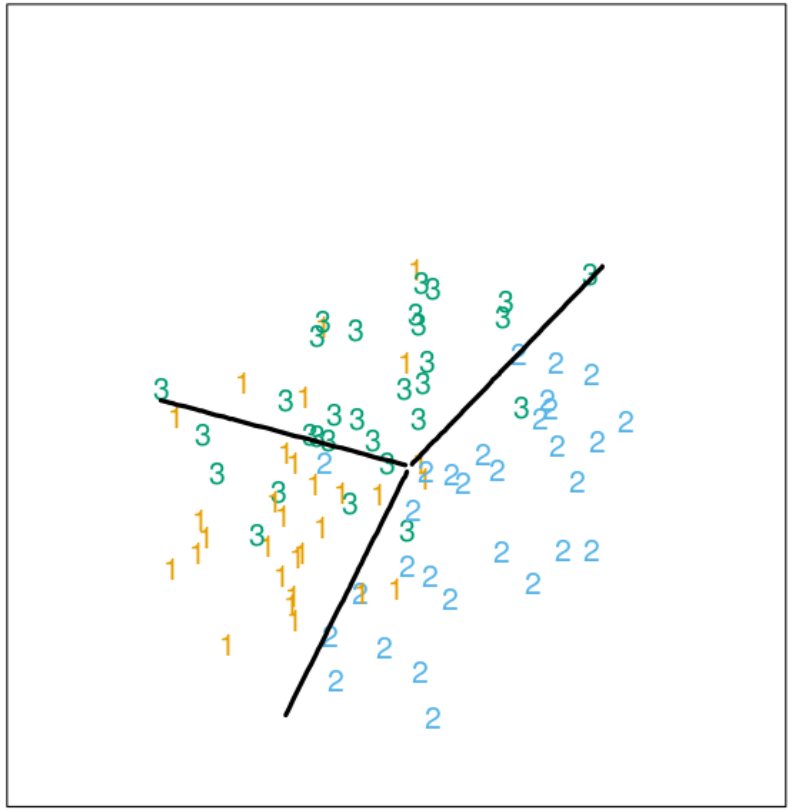
- Т.е. в аргументе сигмоида получается линейная функция от $\mathbf{x}$. Поверхности уровня – это когда $p(C_1 \mid \mathbf{x})$ постоянно, т.е. гиперплоскости в пространстве $\mathbf{x}$. Априорные вероятности $p(C_k)$ просто сдвигают эти гиперплоскости.

Разделяющая гиперплоскость



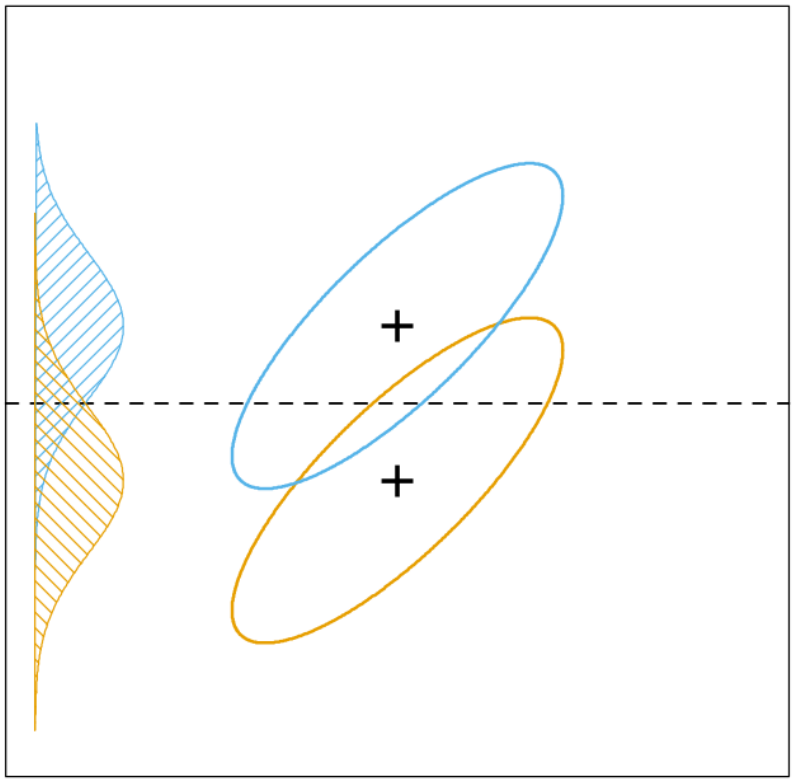
Разделяющая гиперплоскость

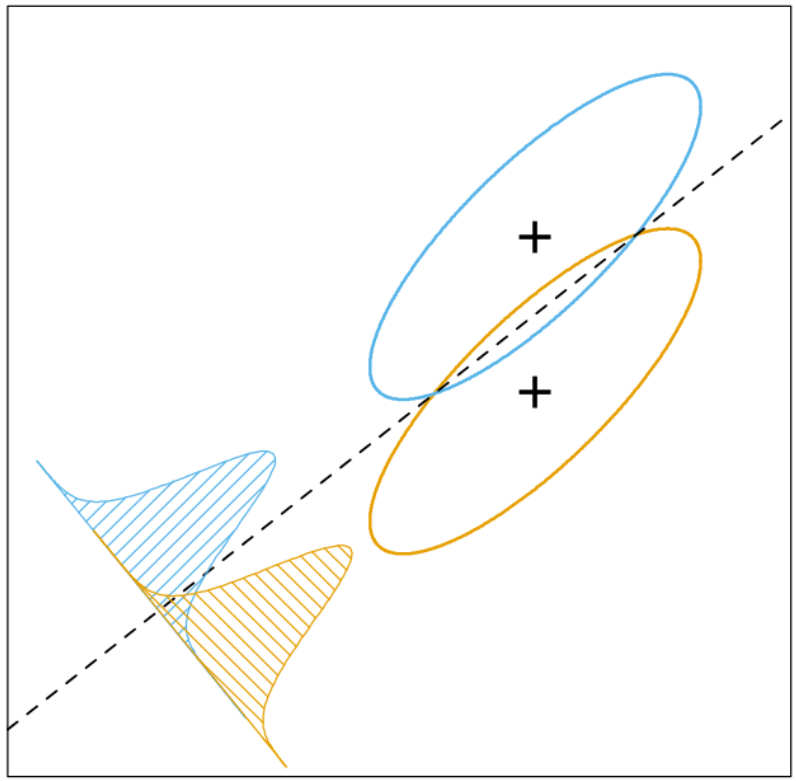




Дискриминант Фишера

Кстати, с дискриминантом Фишера эта разделяющая поверхность отлично сходится.





Несколько классов

- С несколькими классами получится тоже примерно так же:

$$\delta_k(\mathbf{x}) = \mathbf{x}^\top \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^\top \Sigma^{-1} \mu_k + \ln \pi_k,$$

  где $\pi_k = p(C_k)$.
- Получились линейные $\delta_k(\mathbf{x})$, и опять разделяющие поверхности линейные (тут разделяющие поверхности – когда две максимальных вероятности равны).
- Этот метод называется LDA – linear discriminant analysis.

Метод максимального правдоподобия

- Как оценить распределения $p(\mathbf{x} \mid C_k)$, если даны только данные?
- Можно по методу максимального правдоподобия.
- Опять рассмотрим тот же пример: два класса, гауссианы с одинаковой матрицей ковариаций, и есть $D = \{\mathbf{x}_n, t_n\}_{n=1}^N$, где $t_n = 1$ значит C_1, $t_n = 0$ значит C_2.
- Обозначим $p(C_1) = \pi$, $p(C_2) = 1 - \pi$.

Метод максимального правдоподобия

- Для одной точки в классе C_1:

$$p(\mathbf{x}_n, C_1) = p(C_1)p(\mathbf{x}_n \mid C_1) = \pi N(\mathbf{x}_n \mid \mu_1, \Sigma).$$

- В классе C_2:

$$p(\mathbf{x}_n, C_2) = p(C_2)p(\mathbf{x}_n \mid C_2) = (1-\pi) N(\mathbf{x}_n \mid \mu_2, \Sigma).$$

- Функция правдоподобия:

$$p(\mathbf{t} \mid \pi, \mu_1, \mu_2, \Sigma) = \\ = \prod_{n=1}^{N} \left[\pi N(\mathbf{x}_n \mid \mu_1, \Sigma)\right]^{t_n} \left[(1-\pi) N(\mathbf{x}_n \mid \mu_2, \Sigma)\right]^{1-t_n}.$$

Метод максимального правдоподобия

- Максимизируем логарифм правдоподобия. Сначала по π, там останется только

$$\sum_{n=1}^{N} \left[t_n \ln \pi + (1 - t_n) \ln(1 - \pi) \right],$$

и, взяв производную, получим, совершенно неожиданно,

$$\hat{\pi} = \frac{N_1}{N_1 + N_2}.$$

Метод максимального правдоподобия

- Теперь по μ_1; всё, что зависит от μ_1:

$$\sum_n t_n \ln N(\mathbf{x}_n \mid \mu_1, \Sigma) = -\frac{1}{2} \sum_n t_n \left(\mathbf{x}_n - \mu_1\right)^\top \Sigma^{-1} \left(\mathbf{x}_n - \mu_1\right) + C.$$

- Берём производную, и получается, опять внезапно,

$$\hat{\mu}_1 = \frac{1}{N_1} \sum_{n=1}^{N} t_n \mathbf{x}_n.$$

- Аналогично,

$$\hat{\mu}_2 = \frac{1}{N_2} \sum_{n=1}^{N} (1 - t_n) \mathbf{x}_n.$$

Метод максимального правдоподобия

- Для матрицы ковариаций придётся постараться; в результате получится

$$\hat{\Sigma} = \frac{N_1}{N_1 + N_2}\mathbf{S}_1 + \frac{N_2}{N_1 + N_2}\mathbf{S}_2, \text{ где}$$

$$\mathbf{S}_1 = \frac{1}{N_1}\sum_{n \in C_1}(\mathbf{x}_n - \mu_1)(\mathbf{x}_n - \mu_1)^\top,$$

$$\mathbf{S}_2 = \frac{1}{N_2}\sum_{n \in C_2}(\mathbf{x}_n - \mu_2)(\mathbf{x}_n - \mu_2)^\top.$$

- Тоже совершенно неожиданно: взвешенное среднее оценок для двух матриц ковариаций.

Метод максимального правдоподобия

- Это самым прямым образом обобщается на случай нескольких классов.

Упражнение. Сделайте это.

Разные матрицы ковариаций

- А вот с разными матрицами ковариаций уже будет по-другому.
- Квадратичные члены не сократятся.
- Разделяющие поверхности станут квадратичными; QDA – quadratic discriminant analysis.

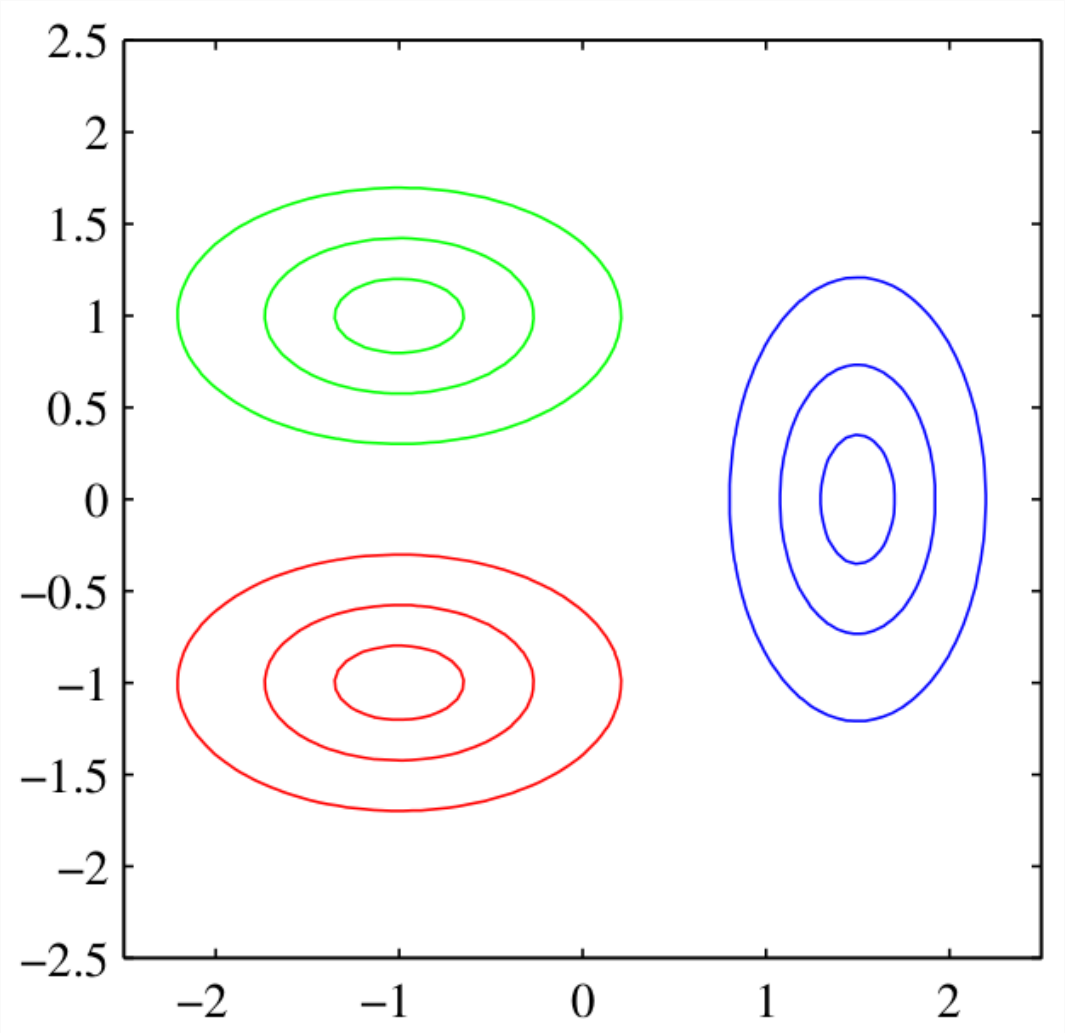
Разные матрицы ковариаций

- В QDA получится

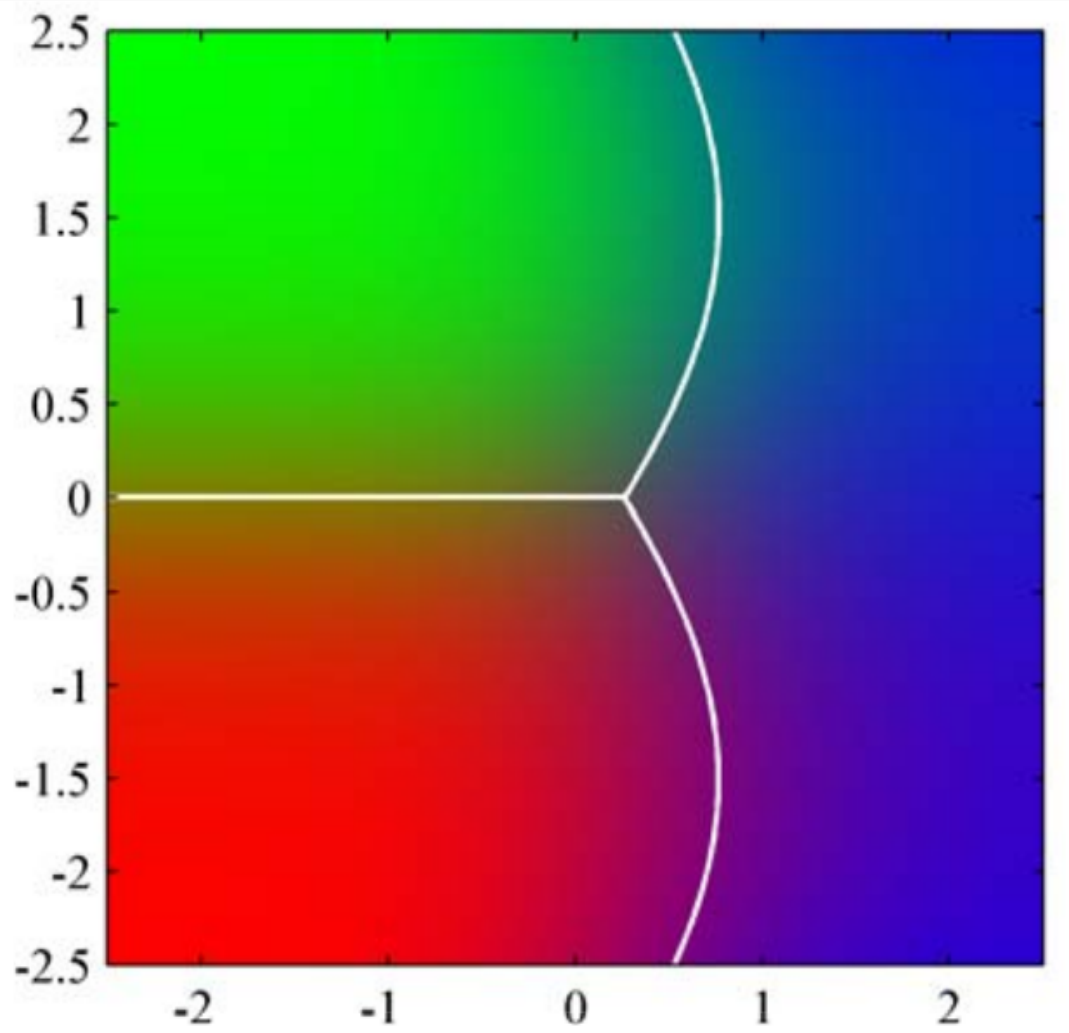
$$\delta_k(\mathbf{x}) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}\left(\mathbf{x}-\mu_k\right)^{-1}\Sigma_k^{-1}\left(\mathbf{x}-\mu_k\right) + \log\pi_k.$$

- Разделяющая поверхность между C_i и C_j – это $\{\mathbf{x} \mid \delta_i(\mathbf{x}) = \delta_j(\mathbf{x})\}$.
- Оценки максимального правдоподобия такие же, только надо отдельно матрицы ковариаций оценивать.

Разные матрицы ковариаций

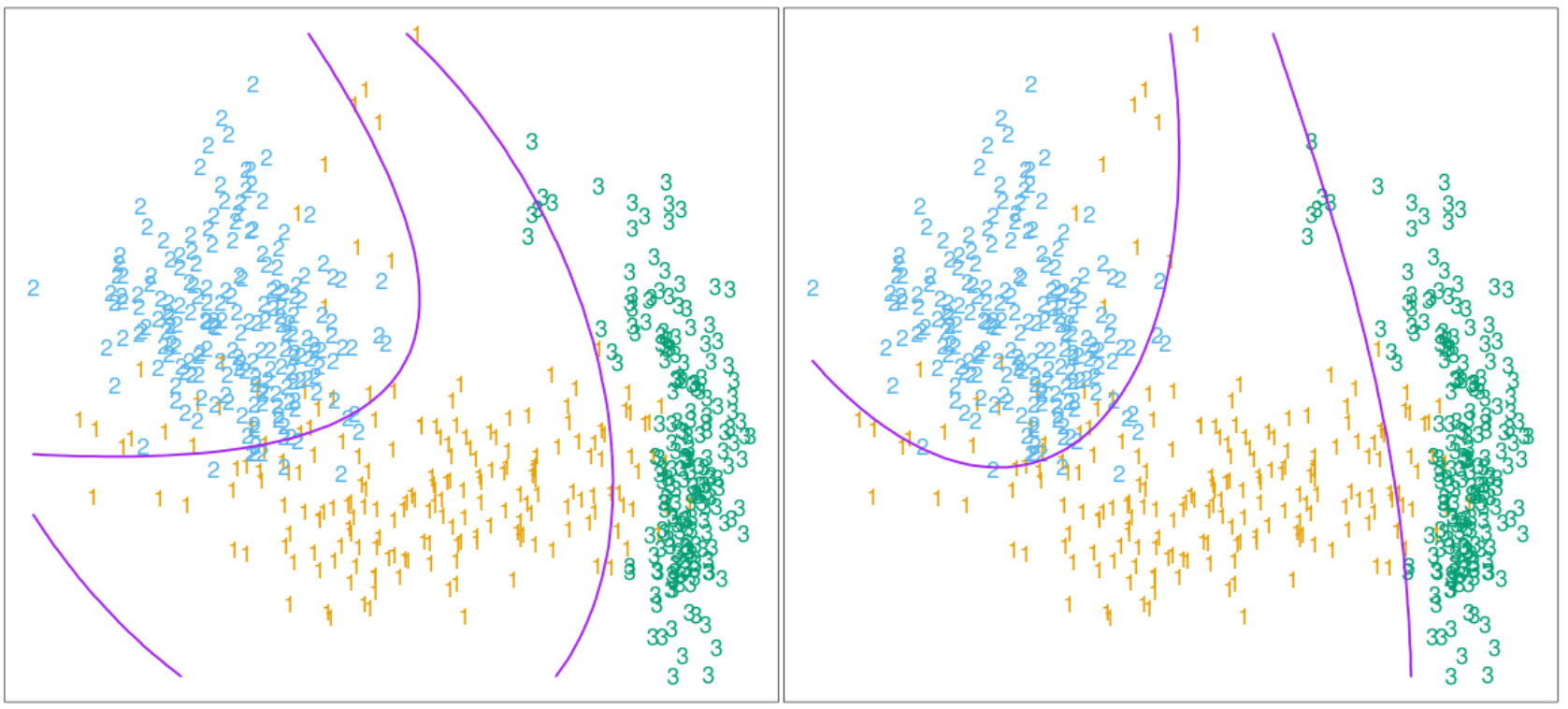


Разные матрицы ковариаций



LDA vs. QDA

Разница между LDA с квадратичными членами и QDA обычно невелика.



LDA vs. QDA

- LDA и QDA неплохо работают на практике. Часто это первая идея в классификации.
- Число параметров:
  - у LDA $(K-1)(d+1)$ параметр: по $d+1$ на каждую разницу вида $\delta_k(\mathbf{x}) - \delta_K(\mathbf{x})$;
  - у QDA $(K-1)(d(d+3)/2+1)$ параметр, но он выглядит гораздо лучше своих лет.

LDA vs. QDA

- Почему хорошо работают?
- Скорее всего, потому, что линейные и квадратичные оценки достаточно стабильны: даже если bias относительно большой (как будет, если данные всё-таки не гауссианами порождены), variance будет маленькой.

RDA

- Компромисс между LDA и QDA – регуляризованный дискриминантный анализ, RDA.
- Стянем ковариации каждого класса к общей матрице ковариаций:

$$\hat{\Sigma}_k(\alpha) = \alpha\hat{\Sigma}_k + (1-\alpha)\hat{\Sigma},$$

  где $\hat{\Sigma}_k$ – оценка из QDA, $\hat{\Sigma}$ – оценка из LDA.
- Или стянем к единичной матрице:

$$\hat{\Sigma}_k(\gamma) = \gamma\hat{\Sigma}_k + (1-\gamma)\hat{\sigma}^2\mathbf{I}.$$

Снижение ранга в LDA

- Предположим, что размерность d больше, чем число классов K.
- Тогда центроиды классов $\hat{\mu}_k$ лежат в подпространстве размерности $\leq K - 1$.
- И когда мы определяем ближайший центроид, нам достаточно считать расстояния только в этом подпространстве.
- Таким образом, можно сократить ранг задачи.

Снижение ранга в LDA

- Куда именно проецировать? Не обязательно само подпространство, порождённое центроидами, будет оптимальным.
- Это мы уже проходили: для размерности 1 это линейный дискриминант Фишера.
- Это он и есть: оптимальное подпространство будет там, где межклассовая дисперсия максимальна по отношению к внутриклассовой.

Спасибо!

Спасибо за внимание!