Классификация

Сергей Николенко

НИУ ВШЭ — Санкт-Петербург
29 февраля 2020 г.

Random facts:

- 29 февраля в Шотландии с 1288 г. был объявлен днём, когда женщина может предложить брак мужчине; нужно было, чтобы «всякая дама, которая идёт свататься, надевала нижнюю сорочку из багряной фланели и чтобы кромка её была хорошо видна, иначе мужчине придётся заплатить за неё штраф» в 1 фунт
- 29 февраля 1504 г. Христофор Колумб сумел обмануть аборигенов Ямайки, отказывавшихся поставлять европейцам еду; Колумб знал, что в этот день будет лунное затмение, и объявил, что за такое поведение испанский бог лишит их Луны; во время затмения перепуганные индейцы согласились возобновить поставки, а Колумб милостиво вернул им Луну
- за 29 февраля 1712 г. в Швеции последовало 30 февраля; чтобы вернуться к юлианскому календарю после пропуска високосных дней (собирались постепенно перейти на григорианский, начали в 1700, но потом началась Северная война и стало не до того), Швеция в 1712 г. добавила к февралю два лишних дня вместо одного
- 29 февраля 1880 г. было закончено строительство тоннеля Готард через Альпы

Эквивалентное ядро и сравнение моделей

Эквивалентное ядро

- Вспомним наши байесовские предсказания:

$$p(t \mid \mathbf{t}, \alpha, \beta) = \mathcal{N}(t \mid \boldsymbol{\mu}_N^\top \boldsymbol{\phi}(\mathbf{x}), \sigma_N^2),$$
$$\text{где } \sigma_N^2 = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^\top \boldsymbol{\Sigma}_N \boldsymbol{\phi}(\mathbf{x}).$$

- Давайте перепишем среднее апостериорного распределения в другой форме (вспомним, что $\boldsymbol{\mu}_N = \beta \boldsymbol{\Sigma}_N \boldsymbol{\Phi}^\top \mathbf{t}$):

$$y(\mathbf{x}, \boldsymbol{\mu}_N) = \boldsymbol{\mu}_N^\top \boldsymbol{\phi}(\mathbf{x}) = \beta \boldsymbol{\phi}(\mathbf{x})^\top \boldsymbol{\Sigma}_N \boldsymbol{\Phi}^\top \mathbf{t} =$$
$$= \sum_{n=1}^{N} \beta \boldsymbol{\phi}(\mathbf{x})^\top \boldsymbol{\Sigma}_N \boldsymbol{\phi}(\mathbf{x}_n) t_n.$$

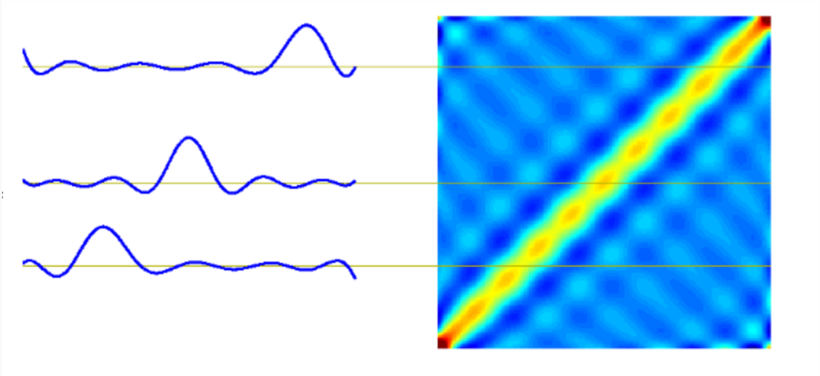
Эквивалентное ядро

- $y(\mathbf{x}, \boldsymbol{\mu}_N) = \sum_{n=1}^{N} \beta\boldsymbol{\phi}(\mathbf{x})^\top \boldsymbol{\Sigma}_N \boldsymbol{\phi}(\mathbf{x}_n) t_n$.
- Это значит, что предсказание можно переписать как

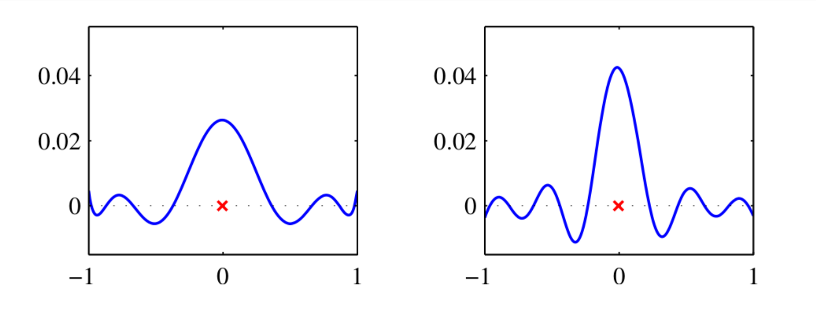
$$y(\mathbf{x}, \boldsymbol{\mu}_N) = \sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) t_n.$$

- Т.е. мы предсказываем следующую точку как линейную комбинацию значений в известных точках.
- Функция $k(\mathbf{x}, \mathbf{x}') = \beta\boldsymbol{\phi}(\mathbf{x})^\top \boldsymbol{\Sigma}_N \boldsymbol{\phi}(\mathbf{x}')$ называется *эквивалентным ядром* (equivalent kernel).

Эквивалентное ядро



Эквивалентное ядро



Выводы про эквивалентное ядро

- Эквивалентное ядро $k(\mathbf{x}, \mathbf{x}')$ локализовано вокруг $\mathbf{x}$ как функция $\mathbf{x}'$, т.е. каждая точка оказывает наибольшее влияние около себя и затухает потом.
- Можно было бы с самого начала просто определить ядро и предсказывать через него, безо всяких базисных функций ϕ – такой подход мы ещё будем рассматривать.

Упражнение. Докажите, что $\sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) = 1$.

Байесовское сравнение моделей

- Мы говорили о том, что при увеличении числа параметров модели возникает оверфиттинг.
- Как этого избежать? Как сравнить модели с разным числом параметров?
- Теория байесовского вывода предлагает такой выход: давайте будем не точечные оценки параметров модели рассматривать, а тоже интегрировать по параметрам модели.

Байесовское сравнение моделей

- Пусть мы хотим сравнить модели из множества $\{\mathcal{M}_i\}_{i=1}^{L}$.
- Модель – это распределение вероятностей над данными D.
- По тестовому набору D можно оценить апостериорное распределение

$$p(\mathcal{M}_i \mid D) \propto p(\mathcal{M}_i)p(D \mid \mathcal{M}_i).$$

Байесовское сравнение моделей

- Если знать апостериорное распределение, то можно сделать предсказание:

$$p(t \mid \mathrm{x}, D) = \sum_{i=1}^{L} p(t \mid \mathrm{x}, \mathcal{M}_i, \mathcal{D}) p(\mathcal{M}_i \mid D).$$

- *Model selection* (выбор модели) – это когда мы приближаем предсказание, выбирая просто самую (апостериорно) вероятную модель.

Байесовское сравнение моделей

- Если модель определена параметрически, через $\mathbf{w}$, то

$$p(D \mid \mathcal{M}_i) = \int p(D \mid \mathbf{w}, \mathcal{M}_i) p(\mathbf{w} \mid \mathcal{M}_i) d\mathbf{w}.$$

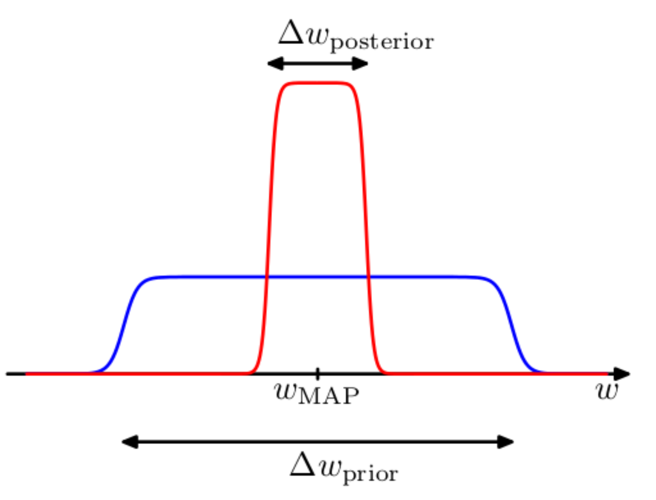
- Т.е. это вероятность сгенерировать D, если выбирать параметры модели по её априорному распределению, а потом накидывать данные.
- Это, кстати, в точности знаменатель из теоремы Байеса:

$$p(\mathbf{w} \mid \mathcal{M}_i, D) = \frac{p(D \mid \mathbf{w}, \mathcal{M}_i) p(\mathbf{w} \mid \mathcal{M}_i)}{p(D \mid \mathcal{M}_i)}.$$

Байесовское сравнение моделей

- Предположим, что у модели один параметр w, а апостериорное распределение – это острый пик вокруг w_{MAP} шириной $\Delta w_{\mathrm{posterior}}$.
- Тогда можно приблизить $p(D) = \int p(D \mid w)p(w)dw$ как значение в максимуме, умноженное на ширину.
- Предположим ещё, что априорное распределение тоже плоское, $p(w) = \frac{1}{\Delta w_{\mathrm{prior}}}$.

Приближение $p(D)$



Приближение $p(D)$

- Тогда получится

$$p(D) = \int p(D \mid w) p(w) dw \approx p\left(D \mid w_{\mathrm{MAP}}\right) \frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}},$$

$$\ln p(D) \approx \ln p\left(D \mid w_{\mathrm{MAP}}\right) + \ln\left(\frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}}\right).$$

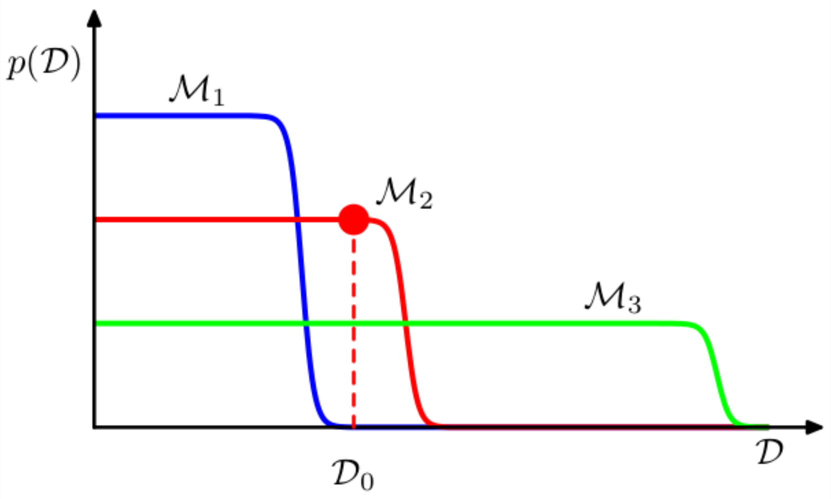
- Это значит, что мы добавляем штраф за «слишком узкое» апостериорное распределение – то есть в точности штраф за оверфиттинг!
- Для модели из M параметров, если предположить, что у них одинаковые $\Delta w_{\mathrm{posterior}}$, получим

$$\ln p(D) \approx \ln p\left(D \mid w_{\mathrm{MAP}}\right) + M \ln\left(\frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}}\right).$$

Другой взгляд

- Другими словами: давайте посмотрим, какие датасеты может генерировать та или иная модель.
- Простая модель (e.g., линейная) генерирует похожие датасеты, «мало» разных датасетов, у неё высокая $p(D \mid \mathcal{M})$.
- Сложная модель (e.g., многочлен девятой степени) генерирует «много» разных датасетов, у неё низкая $p(D \mid \mathcal{M})$.
- Но сложная может хорошо выразить датасеты, которые не может выразить простая; поэтому в сумме надо выбирать «среднюю».

Приближение $p(D)$



Правильный ответ лучше

- Sanity check: тут какие-то штрафы мы навводили; будет ли истинный правильный ответ $p(D \mid \mathcal{M}_{\mathrm{true}})$ всегда оптимальным в этом смысле?
- Конечно, для конкретного датасета может так повезти, что не будет.
- Но если усреднить по всем датасетам, выбранным по $p(D \mid \mathcal{M}_{\mathrm{true}})$...

Правильный ответ лучше

- …то получится

$$\mathrm{E}\left[\ln \frac{p(D \mid \mathcal{M}_{\mathrm{true}})}{p(D \mid \mathcal{M})}\right] = \int p(D \mid \mathcal{M}_{\mathrm{true}}) \ln \frac{p(D \mid \mathcal{M}_{\mathrm{true}})}{p(D \mid \mathcal{M})} dD.$$

- Это называется *расстоянием Кульбака-Лейблера* (Kullback-Leibler divergence) между распределениями $p(D \mid \mathcal{M}_{\mathrm{true}})$ и $p(D \mid \mathcal{M})$.

Упражнение. Докажите, что расстояние Кульбака-Лейблера всегда неотрицательно, т.е. что $p(D \mid \mathcal{M}_{\mathrm{true}}) \geq p(D \mid \mathcal{M})$ для любой $\mathcal{M}$.

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:

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

- Тоже можно интерпретировать как вероятности – или пропорционально им.

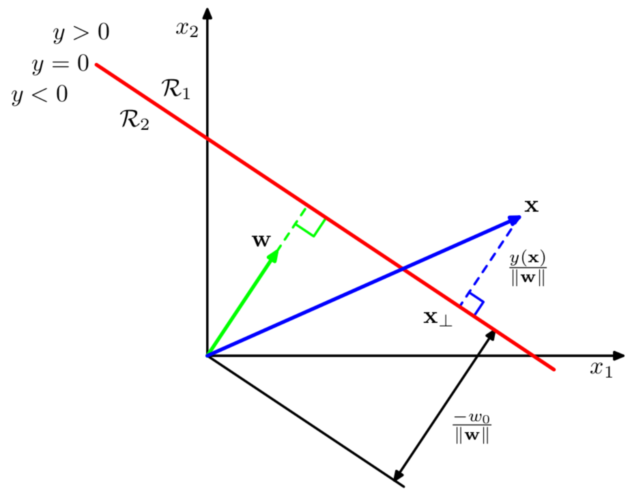
Разделяющая гиперплоскость

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

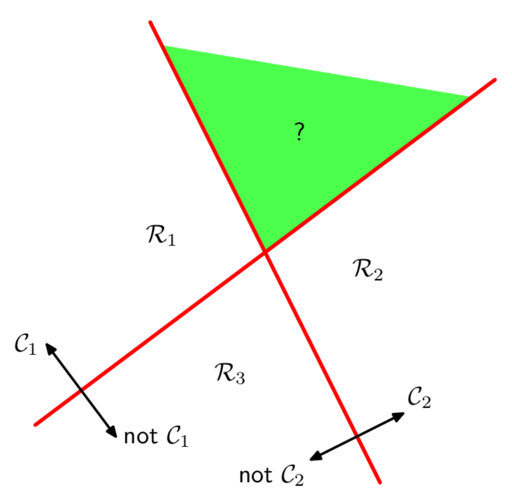
Разделяющая гиперплоскость



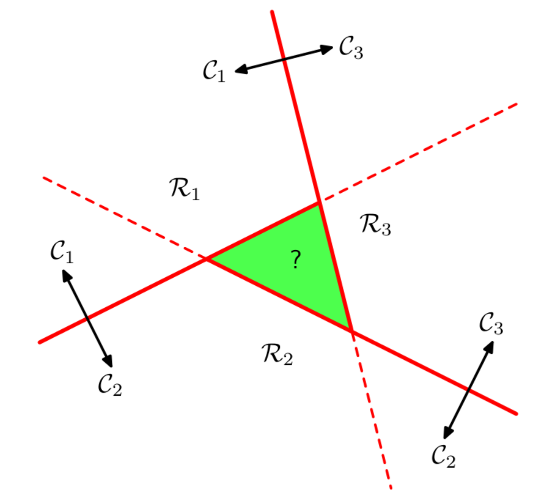
Несколько классов

- 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

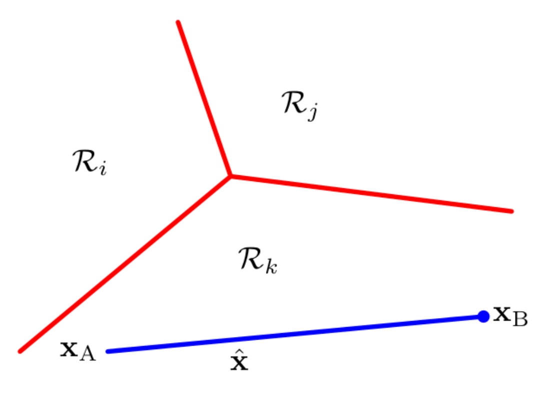
- Лучше рассмотреть единый дискриминант из K линейных функций:

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

- 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

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

- 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:

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

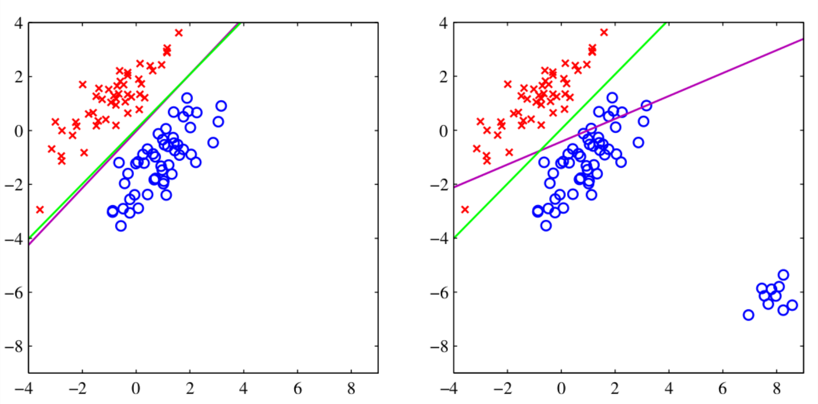
Метод наименьших квадратов

- Это решение сохраняет линейность.

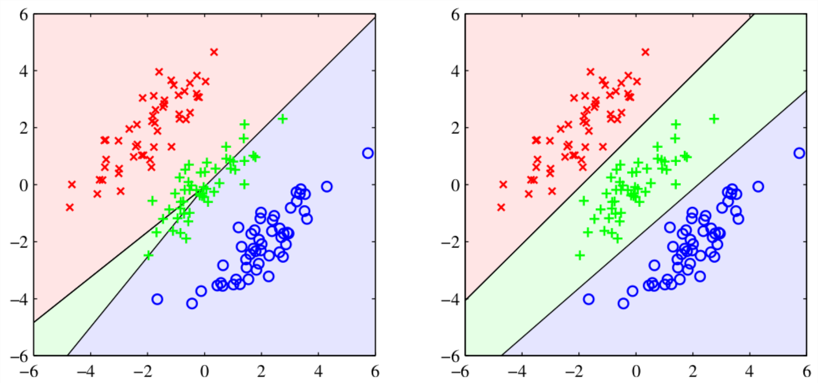
Упражнение. Докажите, что в схеме кодирования 1-of-K предсказания $y_k(\mathbf{x})$ для разных классов при любом $\mathbf{x}$ будут давать в сумме 1. Почему они всё-таки не будут разумными оценками вероятностей?

- 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

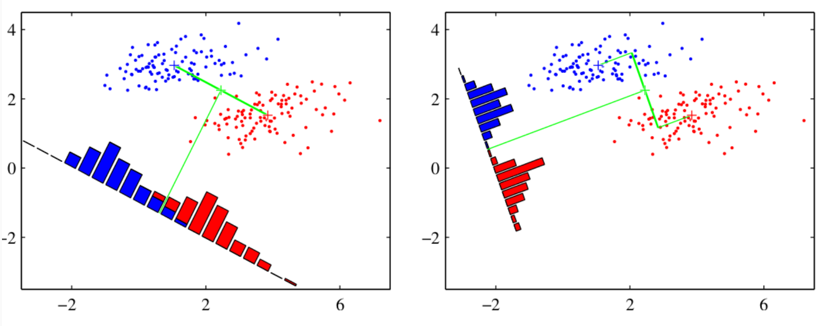
- Рассмотрим два класса $\mathcal{C}_1$ и $\mathcal{C}_2$ с N_1 и N_2 точками.
- Первая идея – надо найти серединный перпендикуляр между центрами кластеров

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

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

- 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 $\mathbf{y}$:

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

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

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

LDA и QDA

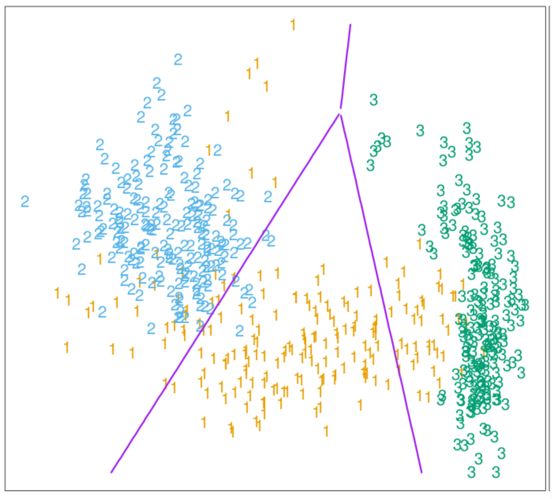
В прошлый раз

- В прошлый раз мы рассмотрели задачу классификации.
- Построили разделяющую гиперплоскость методом наименьших квадратов.
- И методом линейного дискриминанта Фишера.
- А потом научились обучать перцептрон и доказали сходимость метода.

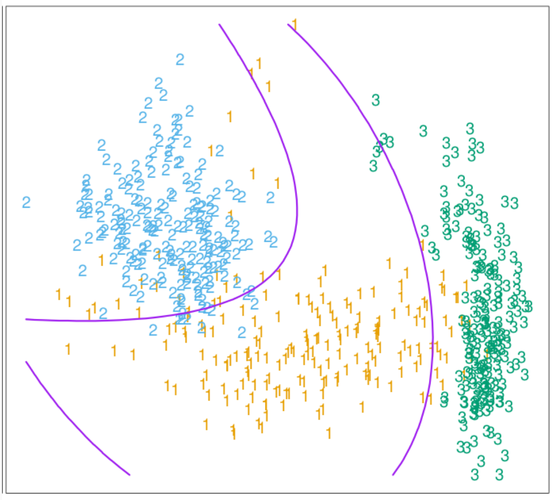
Нелинейные поверхности

- 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 $p(\mathbf{x} \mid \mathcal{C}_k)$, найдём априорные распределения $p(\mathcal{C}_k)$, будем искать $p(\mathcal{C}_k \mid \mathbf{x})$ по теореме Байеса.
- Для двух классов:

$$p(\mathcal{C}_1 \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x} \mid \mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x} \mid \mathcal{C}_2)p(\mathcal{C}_2)}.$$

Генеративные модели

- Перепишем:

$$p(\mathcal{C}_1 \mid \mathrm{x}) = \frac{p(\mathrm{x} \mid \mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathrm{x} \mid \mathcal{C}_1)p(\mathcal{C}_1) + p(\mathrm{x} \mid \mathcal{C}_2)p(\mathcal{C}_2)} = \frac{1}{1 + e^{-a}} = \sigma(a),$$

  где

$$a = \ln \frac{p(\mathrm{x} \mid \mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathrm{x} \mid \mathcal{C}_2)p(\mathcal{C}_2)}, \qquad \sigma(a) = \frac{1}{1 + e^{-a}}.$$

- $\sigma(a)$ – *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

$$p(\mathcal{C}_k \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \mathcal{C}_k)p(\mathcal{C}_k)}{\sum_j p(\mathbf{x} \mid \mathcal{C}_j)p(\mathcal{C}_j)} = \frac{e^{a_k}}{\sum_j e^{a_j}}.$$

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

Пример

- Давайте рассмотрим гауссовы распределения для классов:

$$p(\mathbf{x} \mid \mathcal{C}_k) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \Sigma).$$

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

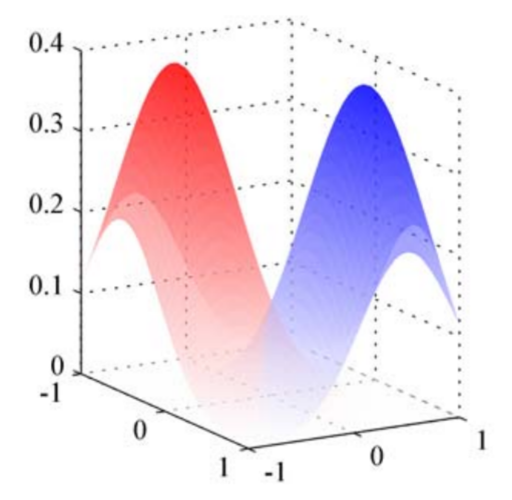
Пример

- …получится

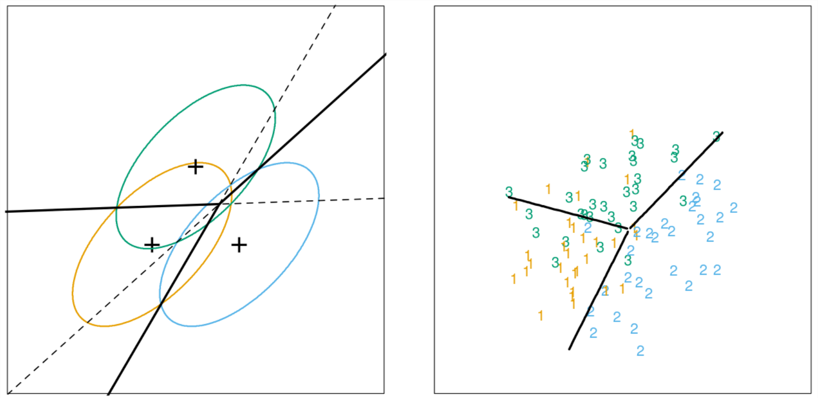
$$
\begin{aligned}
p(\mathcal{C}_1 \mid \mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + w_0), &\ \text{где} \\
\mathbf{w} &= \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \\
w_0 &= -\frac{1}{2}\boldsymbol{\mu}_1^\top \Sigma^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^\top \Sigma^{-1}\boldsymbol{\mu}_2 + \ln \frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}.
\end{aligned}
$$

- Т.е. в аргументе сигмоида получается линейная функция от $\mathbf{x}$. Поверхности уровня – это когда $p(\mathcal{C}_1 \mid \mathbf{x})$ постоянно, т.е. гиперплоскости в пространстве $\mathbf{x}$. Априорные вероятности $p(\mathcal{C}_k)$ просто сдвигают эти гиперплоскости.

Разделяющая гиперплоскость

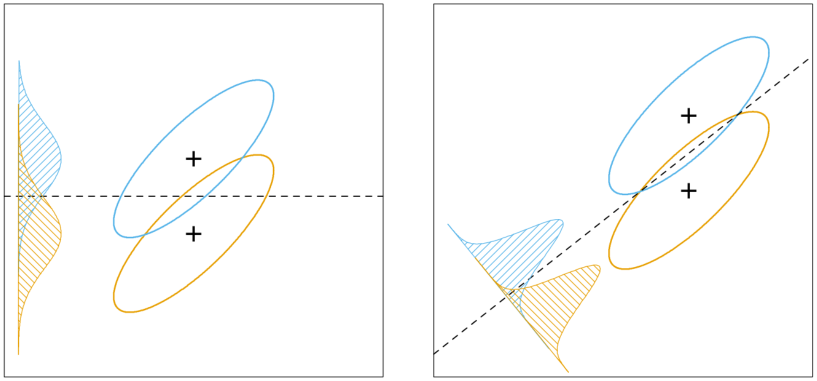


Разделяющая гиперплоскость



Дискриминант Фишера

Кстати, с дискриминантом Фишера эта разделяющая поверхность отлично сходится.



Несколько классов

- С несколькими классами получится тоже примерно так же:

$$\delta_k(\mathrm{x}) = \mathrm{x}^\top \Sigma^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^\top \Sigma^{-1} \boldsymbol{\mu}_k + \ln \pi_k,$$

  где $\pi_k = p(\mathcal{C}_k)$.
- Получились линейные $\delta_k(\mathrm{x})$, 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 $p(\mathbf{x} \mid \mathcal{C}_k)$, 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 $\mathcal{C}_1$, $t_n = 0$ значит $\mathcal{C}_2$.
- Обозначим $p(\mathcal{C}_1) = \pi$, $p(\mathcal{C}_2) = 1 - \pi$.

Метод максимального правдоподобия

- Для одной точки в классе $\mathcal{C}_1$:

$$p(\mathbf{x}_n, \mathcal{C}_1) = p(\mathcal{C}_1)p(\mathbf{x}_n \mid \mathcal{C}_1) = \pi\mathcal{N}(\mathbf{x}_n \mid \mu_1, \Sigma).$$

- В классе $\mathcal{C}_2$:

$$p(\mathbf{x}_n, \mathcal{C}_2) = p(\mathcal{C}_2)p(\mathbf{x}_n \mid \mathcal{C}_2) = (1-\pi)\mathcal{N}(\mathbf{x}_n \mid \mu_2, \Sigma).$$

- Функция правдоподобия:

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

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:

$$\sum_n t_n \ln \mathcal{N}(\mathbf{x}_n \mid \mu_1, \Sigma) = -\frac{1}{2}\sum_n t_n \left(\mathbf{x}_n - \boldsymbol{\mu}_1\right)^\top \Sigma^{-1} \left(\mathbf{x}_n - \boldsymbol{\mu}_1\right) + C.$$

- Берём производную, и получается, опять внезапно,

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

- Аналогично,

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

- Для матрицы ковариаций придётся постараться; в результате получится

$$\hat{\Sigma} = \frac{N_1}{N_1+N_2}\mathrm{S}_1 + \frac{N_2}{N_1+N_2}\mathrm{S}_2, \text{ где}$$
$$\mathrm{S}_1 = \frac{1}{N_1}\sum_{n\in\mathcal{C}_1}\left(\mathrm{x}_n - \boldsymbol{\mu}_1\right)\left(\mathrm{x}_n - \boldsymbol{\mu}_1\right)^{\top},$$
$$\mathrm{S}_2 = \frac{1}{N_2}\sum_{n\in\mathcal{C}_2}\left(\mathrm{x}_n - \boldsymbol{\mu}_2\right)\left(\mathrm{x}_n - \boldsymbol{\mu}_2\right)^{\top}.$$

- 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.

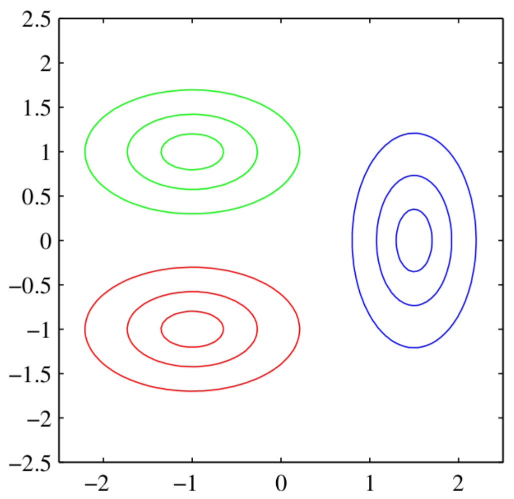
Разные матрицы ковариаций

- В QDA получится

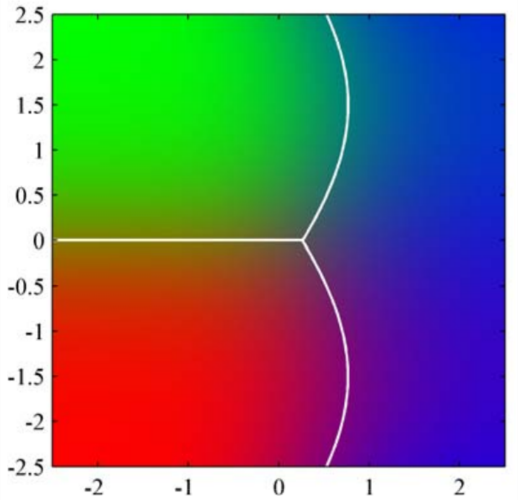
$$\delta_k(\mathrm{x}) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(\mathrm{x}-\boldsymbol{\mu}_k)^{-1}\Sigma_k^{-1}(\mathrm{x}-\boldsymbol{\mu}_k) + \log\pi_k.$$

- Разделяющая поверхность между $\mathcal{C}_i$ и $\mathcal{C}_j$ – это $\{\mathrm{x} \mid \delta_i(\mathrm{x}) = \delta_j(\mathrm{x})\}$.
- Оценки максимального правдоподобия такие же, только надо отдельно матрицы ковариаций оценивать.

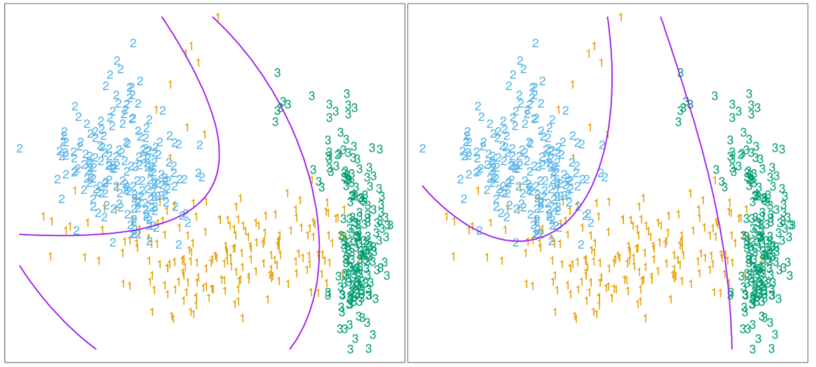
Разные матрицы ковариаций



Разные матрицы ковариаций



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 $\delta_k(\mathrm{x}) - \delta_K(\mathrm{x})$;
  - 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!