КЛАССИФИКАЦИЯ

Сергей Николенко

СПбГУ — Санкт-Петербург
29 сентября 2017 г.

Random facts:

- 29 сентября 1916 г. Джон Рокфеллер стал первым в мире миллиардером
- 29 сентября 1984 г. было уложено последнее, «золотое» звено БАМа
- 29 сентября 1998 г. финансовый кризис в Японии привел к банкротству Japan Leasing Corporation, крупнейшему на тот момент банкротству со времен второй мировой войны, а 29 сентября 2008 г. после банкротства Lehman Brothers и Washington Mutual индекс Доу-Джонса упал на 777.68 пунктов, самое большое падение за один день в истории

LDA и QDA

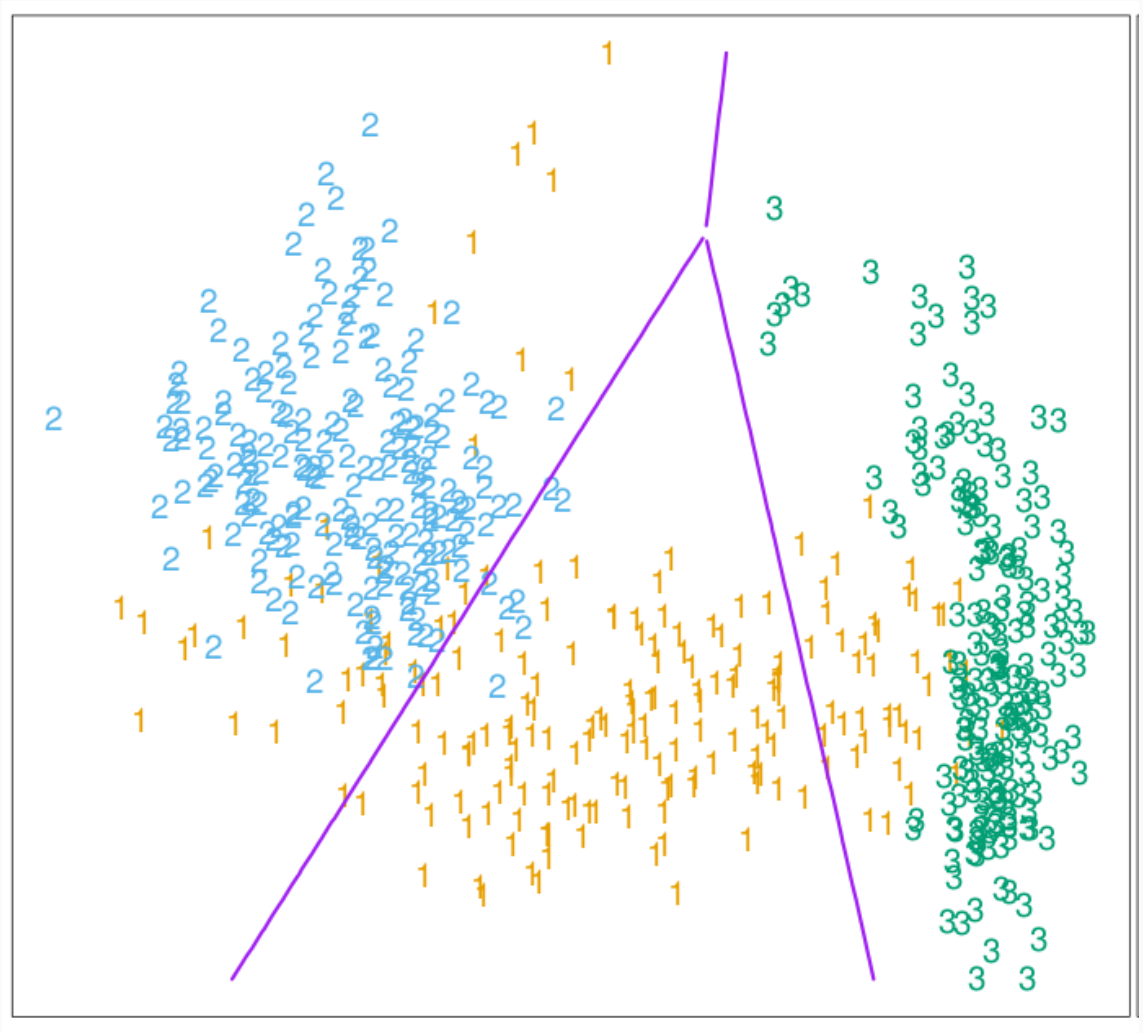
В прошлый раз

- 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.

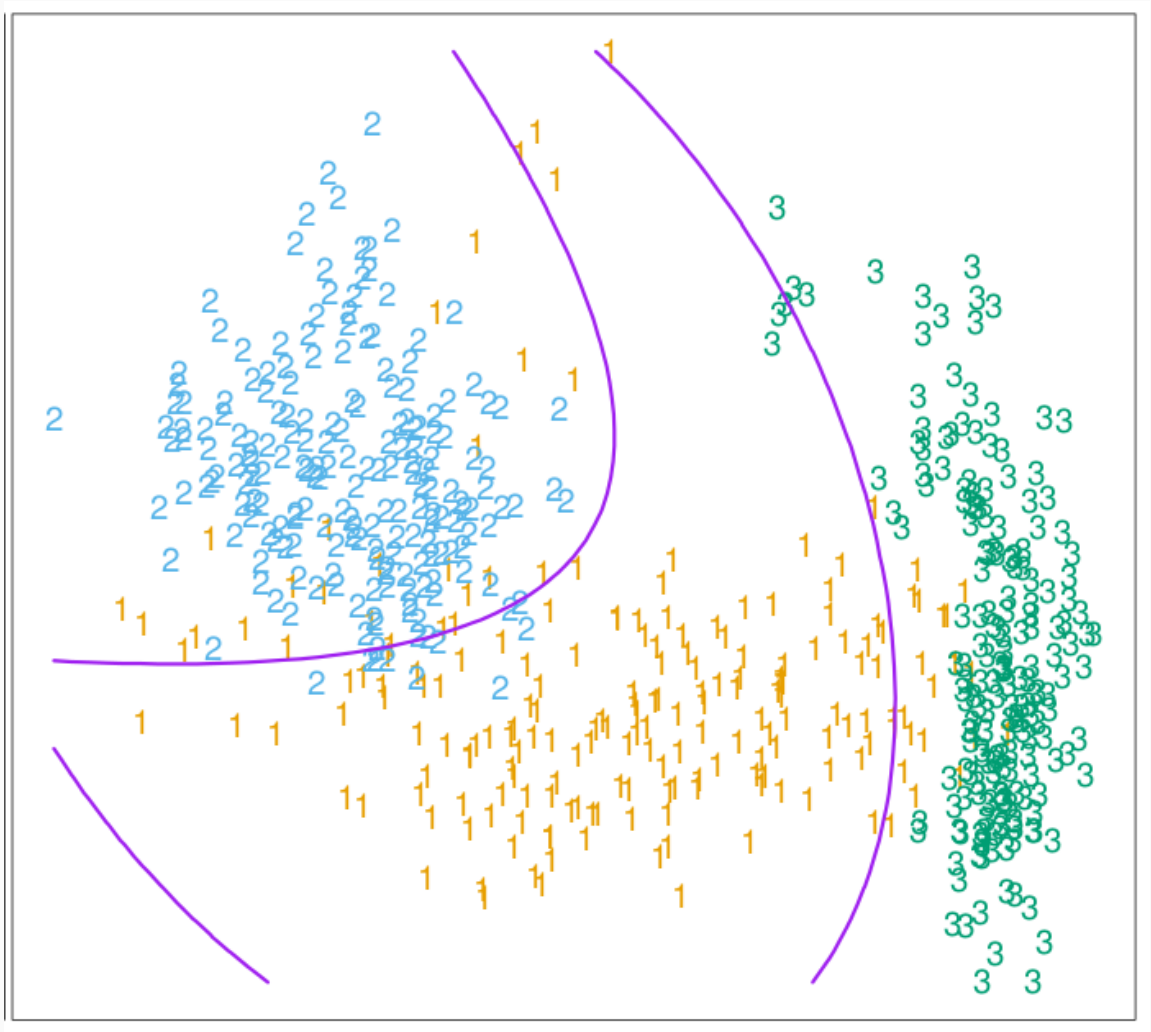
Нелинейные поверхности

- 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:

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

- Перепишем:

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

где

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

- $\sigma(a)$ – *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:

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

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

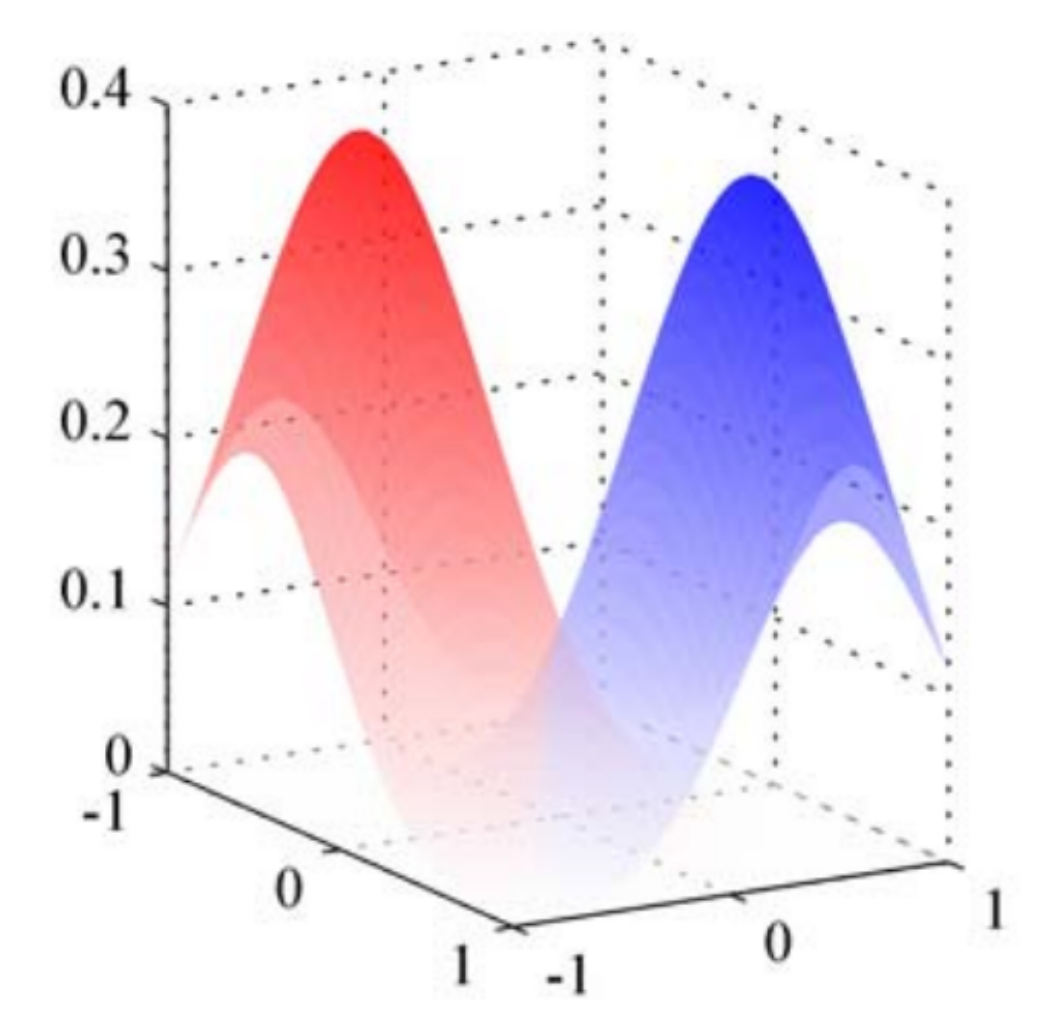
ПРИМЕР

- ...получится

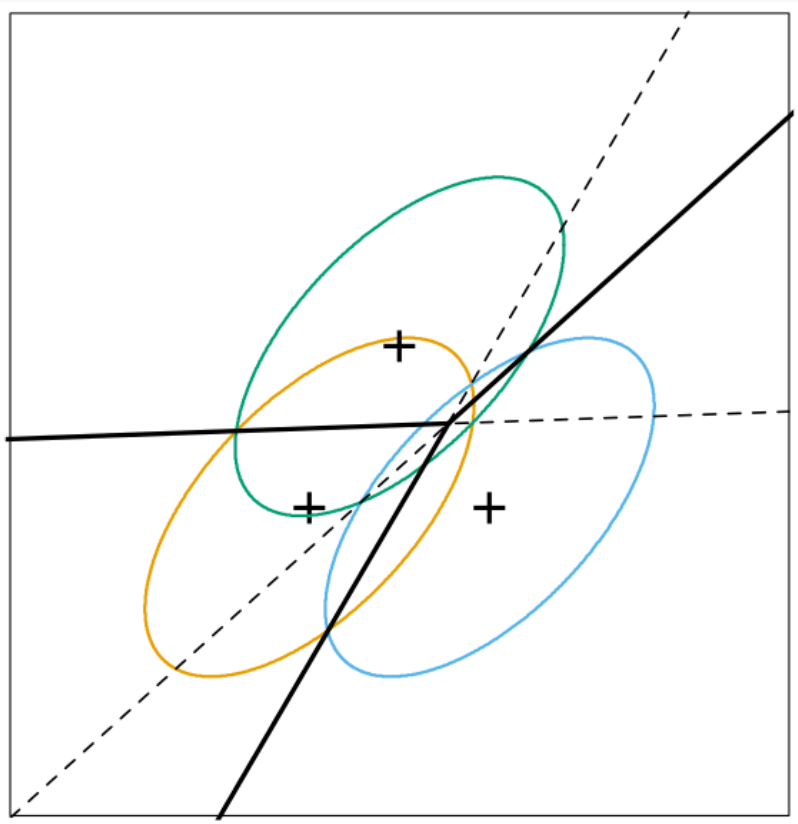
$$p(\mathcal{C}_1 \mid \mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + w_0), \text{ где}$$
$$\mathbf{w} = \Sigma^{-1}\left(\mu_1 - \mu_2\right),$$
$$w_0 = -\frac{1}{2}\mu_1^\top \Sigma^{-1}\mu_1 + \frac{1}{2}\mu_2^\top \Sigma^{-1}\mu_2 + \ln\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}.$$

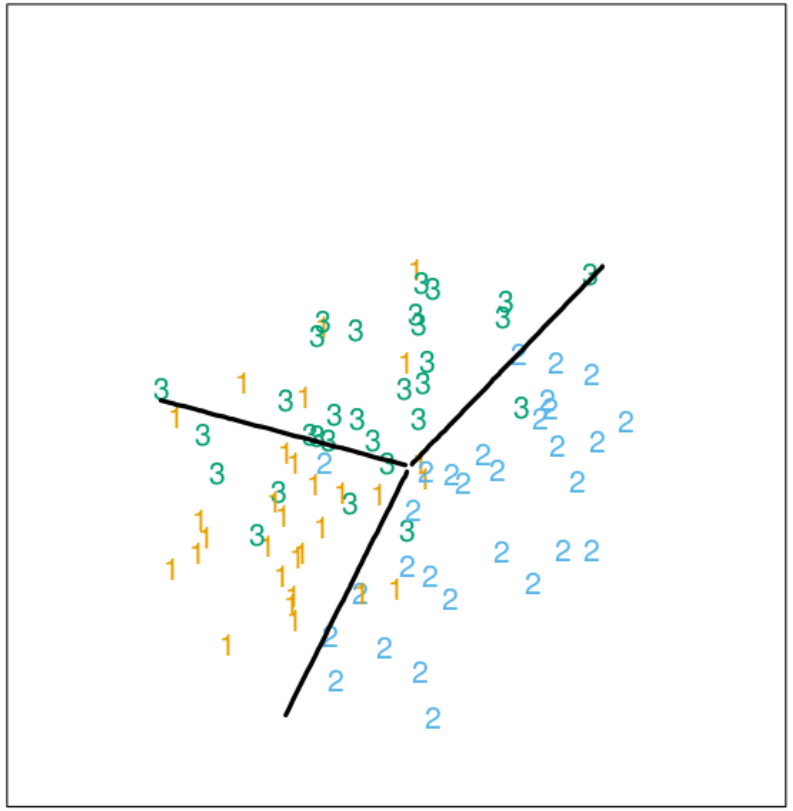
- Т.е. в аргументе сигмоида получается линейная функция от $\mathbf{x}$. Поверхности уровня – это когда $p(\mathcal{C}_1 \mid \mathbf{x})$ постоянно, т.е. гиперплоскости в пространстве $\mathbf{x}$. Априорные вероятности $p(\mathcal{C}_k)$ просто сдвигают эти гиперплоскости.

РАЗДЕЛЯЮЩАЯ ГИПЕРПЛОСКОСТЬ



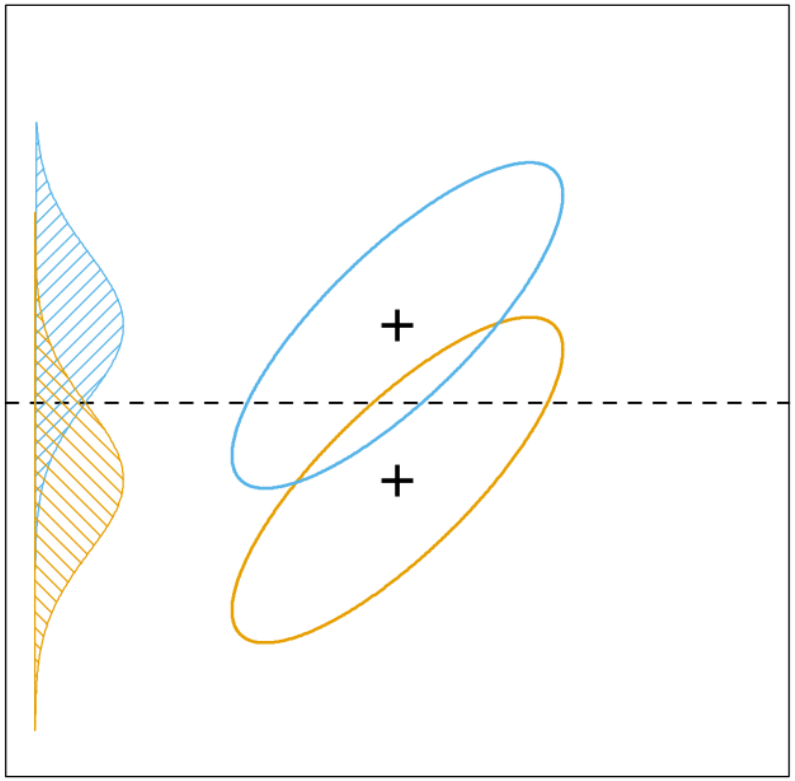
РАЗДЕЛЯЮЩАЯ ГИПЕРПЛОСКОСТЬ

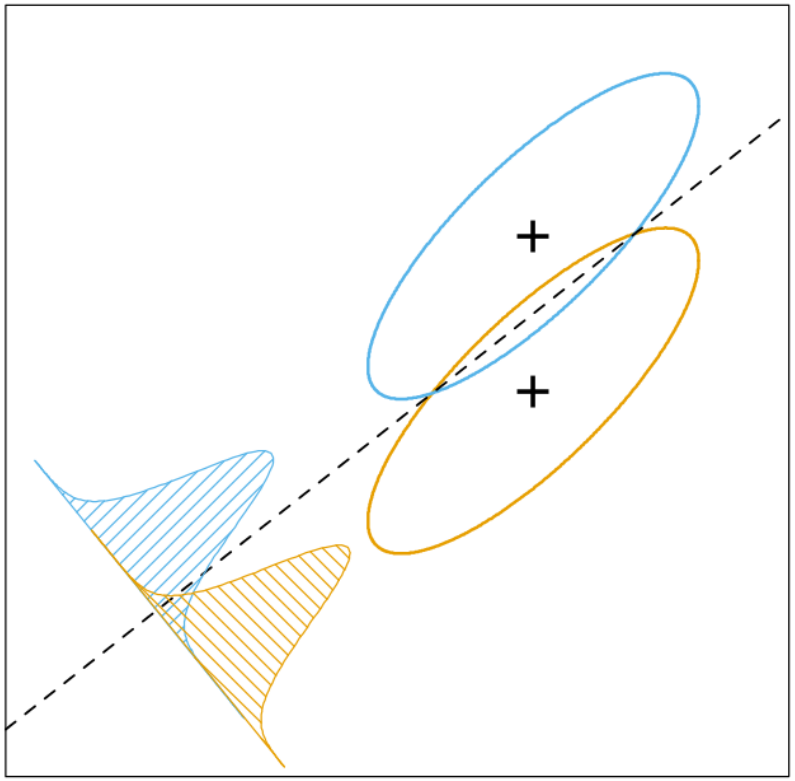




ДИСКРИМИНАНТ ФИШЕРА

Кстати, с дискриминантом Фишера эта разделяющая поверхность отлично сходится.





НЕСКОЛЬКО КЛАССОВ

- С несколькими классами получится тоже примерно так же:

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

  где $\pi_k = p(\mathcal{C}_k)$.
- 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 $p(\mathbf{x} \mid \mathcal{C}_k)$, 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:

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

- 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

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

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

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

- 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.

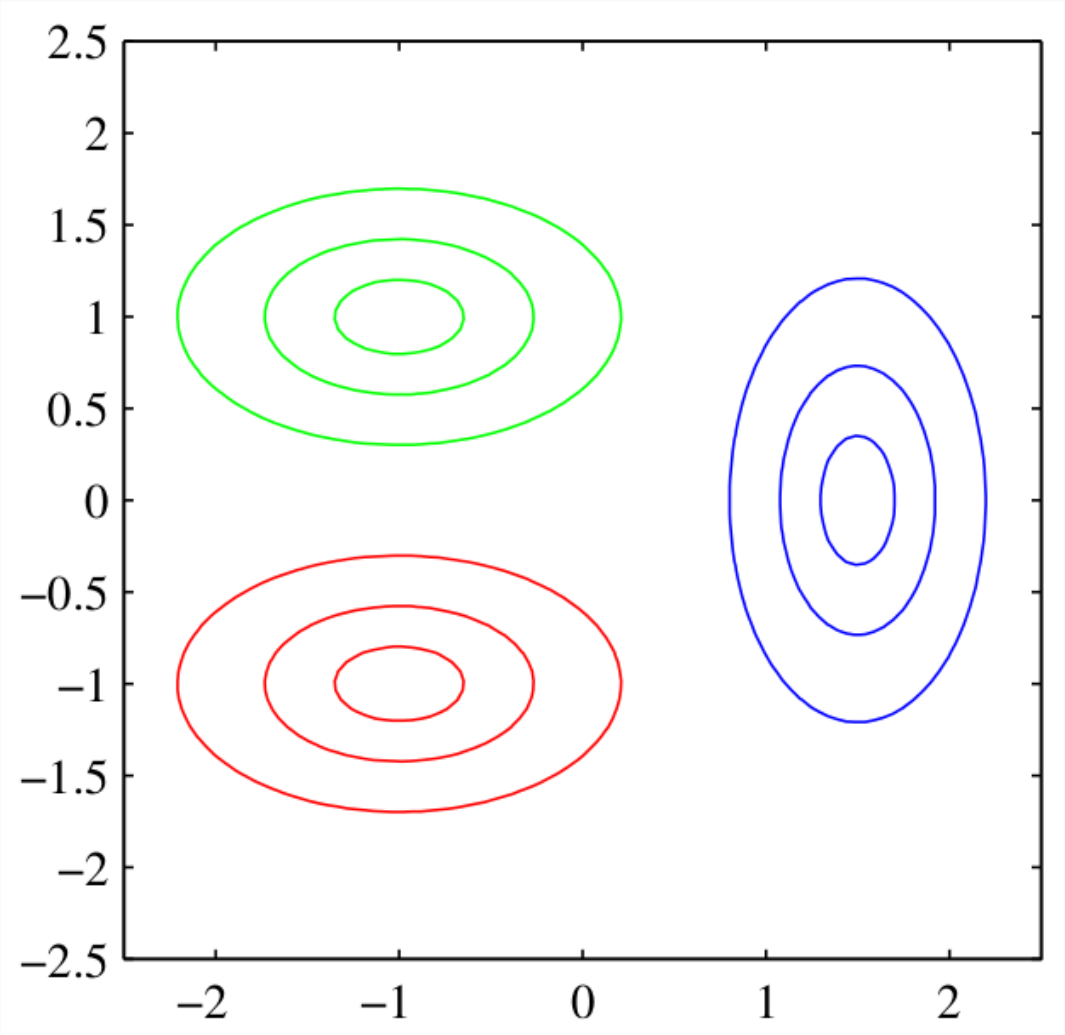
РАЗНЫЕ МАТРИЦЫ КОВАРИАЦИЙ

- В QDA получится

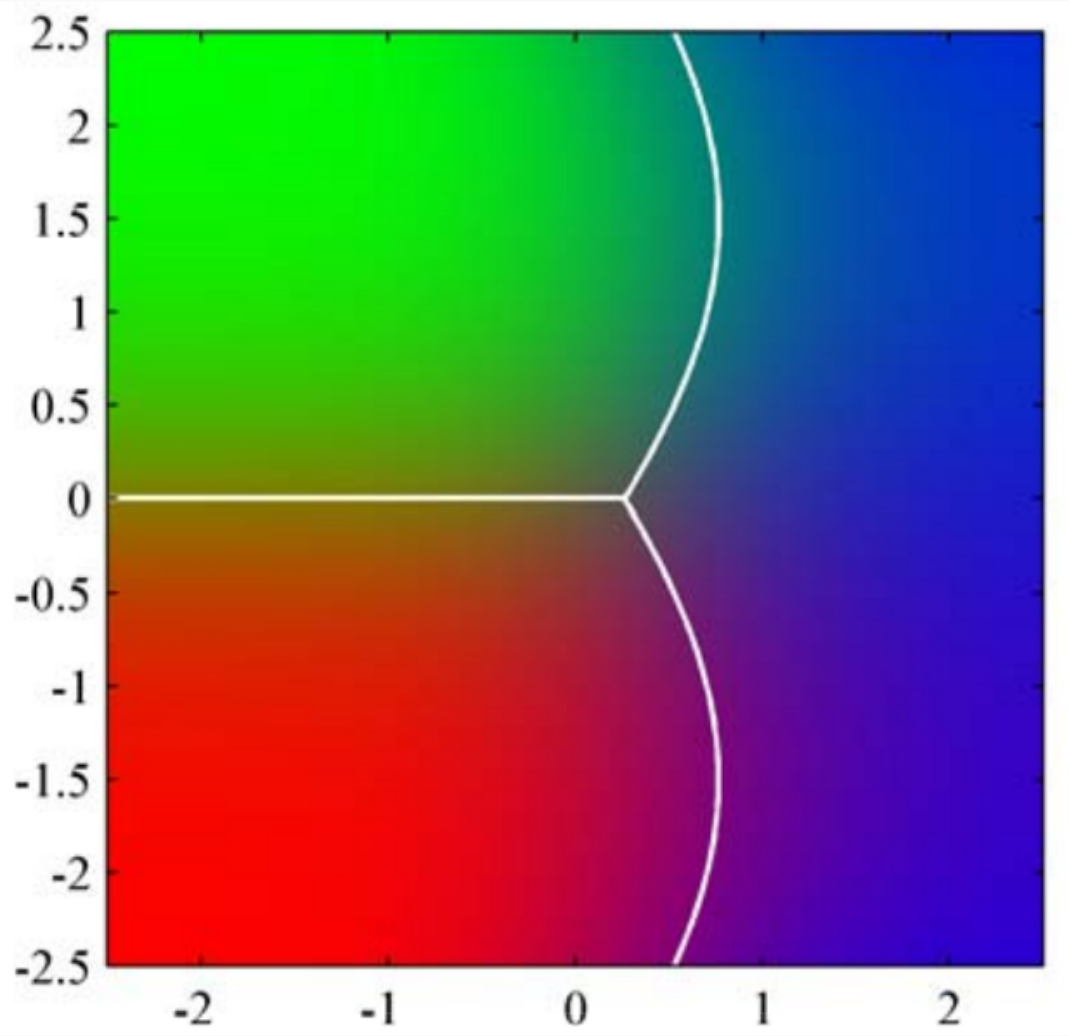
$$\delta_k(\mathbf{x}) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}\left(\mathbf{x}-\mu_k\right)^{-1}\Sigma_k^{-1}\left(\mathbf{x}-\mu_k\right) + \log\pi_k.$$

- Разделяющая поверхность между $\mathcal{C}_i$ и $\mathcal{C}_j$ – 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

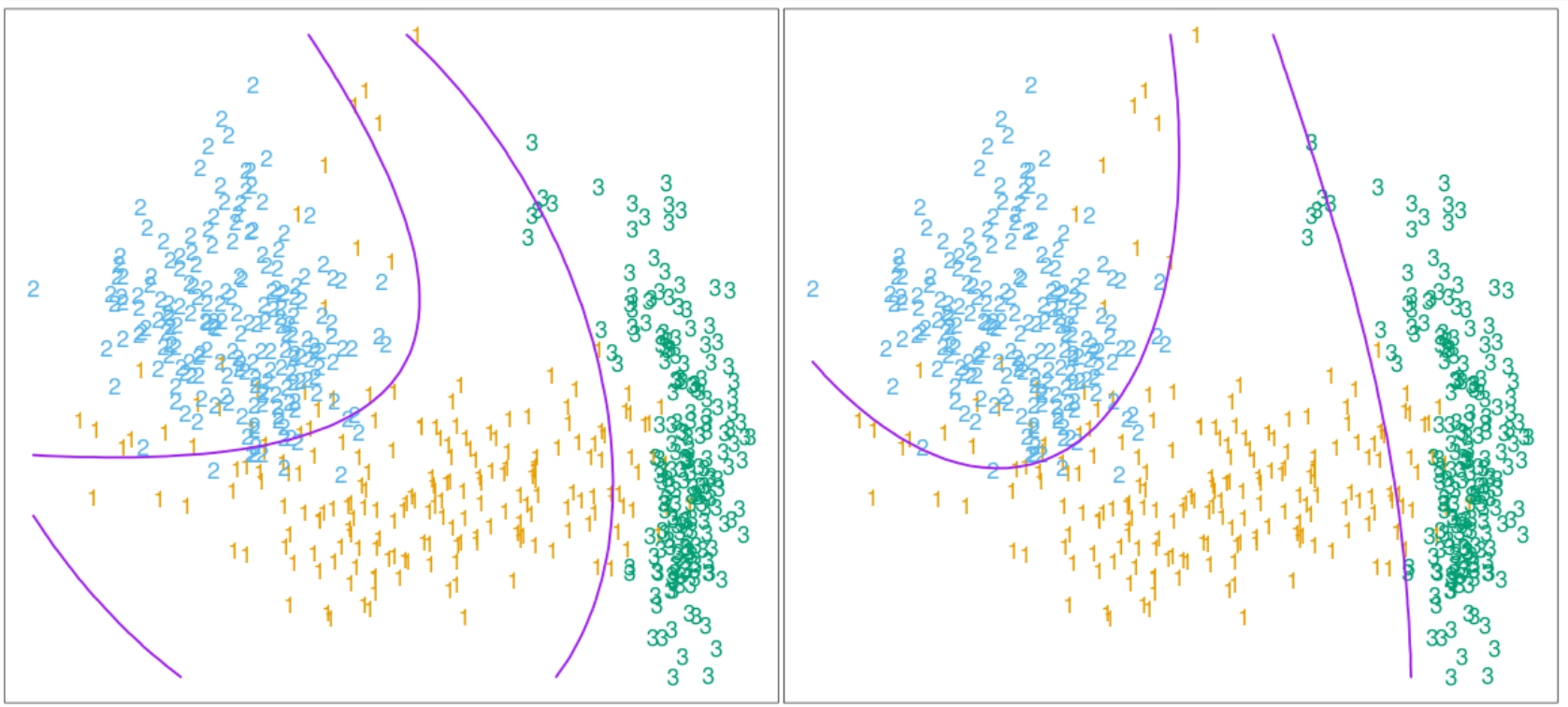


РАЗНЫЕ МАТРИЦЫ КОВАРИАЦИЙ



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.

ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

В ПРОШЛЫЙ РАЗ

- Итак, мы рассмотрели логистический сигмоид:

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

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

- Вывели из него LDA и QDA, обучили их методом максимального правдоподобия, а потом отвлеклись на naive Bayes.

ДВА КЛАССА

- Возвращаемся к задаче классификации.
- Два класса, и апостериорное распределение – логистический сигмоид на линейной функции:

$$p(\mathcal{C}_1 \mid \phi) = y(\phi) = \sigma(\mathbf{w}^\top \phi), \quad p(\mathcal{C}_2 \mid \phi) = 1 - p(\mathcal{C}_1 \mid \phi).$$

- *Логистическая регрессия* – это когда мы напрямую оптимизируем $\mathbf{w}$.

ДВА КЛАССА

- Для датасета $\{\phi_n, t_n\}$, $t_n \in \{0, 1\}$, $\phi_n = \phi(\mathbf{x}_n)$:

$$p(\mathbf{t} \mid \mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n} (1 - y_n)^{1-t_n}, \quad y_n = p(\mathcal{C}_1 \mid \phi_n).$$

- Ищем параметры максимального правдоподобия, минимизируя $-\ln p(\mathbf{t} \mid \mathbf{w})$:

$$E(\mathbf{w}) = -\ln p(\mathbf{t} \mid \mathbf{w}) = -\sum_{n=1}^{N} \left[t_n \ln y_n + (1 - t_n) \ln(1 - y_n) \right].$$

ДВА КЛАССА

- Пользуясь тем, что $\sigma' = \sigma(1-\sigma)$, берём градиент (похоже на перцептрон):

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n)\phi_n.$$

- Если теперь сделать градиентный спуск, получим как раз разделяющую поверхность.
- Заметим, правда, что если данные действительно разделимы, то может получиться жуткий оверфиттинг: $\|\mathbf{w}\| \to \infty$, и сигмоид превращается в функцию Хевисайда. Надо регуляризовать.

IRLS

- В логистической регрессии не получается замкнутого решения из-за сигмоида.
- Но функция $E(\mathbf{w})$ всё равно выпуклая, и можно воспользоваться методом Ньютона-Рапсона – на каждом шаге использовать локальную квадратичную аппроксимацию к функции ошибки:

$$\mathbf{w}^{\text{new}} = \mathbf{w}^{\text{old}} - \mathbf{H}^{-1}\nabla E(\mathbf{w}),$$

где $\mathbf{H}$ (Hessian) – матрица вторых производных $E(\mathbf{w})$.

IRLS

- Замечание: давайте применим Ньютона-Рапсона к обычной линейной регрессии с квадратической ошибкой:

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} \left(\mathbf{w}^\top \phi_n - t_n\right) \phi_n = \Phi^\top \Phi \mathbf{w} - \Phi^\top \mathbf{t},$$

$$\nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} \phi_n \phi_n^\top = \Phi^\top \Phi,$$

и шаг оптимизации будет

$$\mathbf{w}^{\text{new}} = \mathbf{w}^{\text{old}} - \left(\Phi^\top \Phi\right)^{-1} \left[\Phi^\top \Phi \mathbf{w}^{\text{old}} - \Phi^\top \mathbf{t}\right] = \\ = \left(\Phi^\top \Phi\right)^{-1} \Phi^\top \mathbf{t},$$

т.е. мы за один шаг придём к решению.

IRLS

- Для логистической регрессии:

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n)\,\phi_n = \Phi^\top (\mathbf{y} - \mathbf{t}),$$

$$\mathbf{H} = \nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} y_n(1 - y_n)\phi_n\phi_n^\top = \Phi^\top R\Phi$$

для диагональной матрицы R с $R_{nn} = y_n(1 - y_n)$.

IRLS

- Формула шага оптимизации:

$$\mathbf{w}^{\text{new}} = \mathbf{w}^{\text{old}} - \left(\Phi^\top R \Phi\right)^{-1} \Phi^\top (\mathbf{y} - \mathbf{t}) = \left(\Phi^\top R \Phi\right)^{-1} \Phi^\top R \mathbf{z},$$

  где $\mathbf{z} = \Phi \mathbf{w}^{\text{old}} - R^{-1} (\mathbf{y} - \mathbf{t})$.
- Получилось как бы решение взвешенной задачи минимизации квадратического отклонения с матрицей весов R.
- Отсюда название: iterative reweighted least squares (IRLS).

СПАСИБО!

Спасибо за внимание!