МАШИННОЕ ОБУЧЕНИЕ И АНАЛИЗ ДАННЫХ

(Machine Learning and Data Mining)

Н. Ю. Золотых

http://www.uic.unn.ru/~zny/ml

Лекция 8

Линейный и квадратичный дискриминантный анализ

8.1. Дискриминантный анализ

8.1.1. Линейный дискриминантный анализ (LDA)

Линейный дискриминантный анализ (LDA) делает два предположения:

- объекты каждого класса распределены по нормальному закону:

$$p(x \mid y) = \frac{1}{\sqrt{(2\pi)^d \det \boldsymbol{\Sigma}_y}} e^{-\frac{1}{2}(x-\mu_y)^\top \boldsymbol{\Sigma}_y^{-1}(x-\mu_y)}$$

- матрицы ковариации $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_y$ одинаковы для всех классов

В этих предположениях построим оптимальный (байесовский классификатор)
Нужно сравнить две апостериорные вероятности:

$$\Pr(y \mid x) = \frac{p(x \mid y)\ \Pr y}{p(x)} > \Pr(y' \mid x) = \frac{p(x \mid y')\ \Pr y'}{p(x)} \qquad (*)$$

Подставляя выражения для $p(x \mid y)$ и $p(x \mid y')$ в (*) и логарифмируя:

$$p(x \mid y) = \frac{1}{\sqrt{(2\pi)^d \det \boldsymbol{\Sigma}_y}} e^{-\frac{1}{2}(x-\mu_y)^\top \boldsymbol{\Sigma}_y^{-1}(x-\mu_y)}$$

$$\Pr(y \mid x) = \frac{p(x \mid y)\ \Pr y}{p(x)} > \Pr(y' \mid x) = \frac{p(x \mid y')\ \Pr y'}{p(x)} \qquad (*)$$

$$-\frac{1}{2}(x-\mu_y)^\top \boldsymbol{\Sigma}^{-1}(x-\mu_y) + \ln \Pr y > -\frac{1}{2}(x-\mu_{y'})^\top \boldsymbol{\Sigma}^{-1}(x-\mu_{y'}) + \ln \Pr y'$$

Откуда

$$(\mu_y - \mu_{y'})^\top \boldsymbol{\Sigma}^{-1} x > \frac{1}{2}\mu_y^\top \boldsymbol{\Sigma}^{-1}\mu_y - \frac{1}{2}\mu_{y'}^\top \boldsymbol{\Sigma}^{-1}\mu_{y'} - \ln \Pr y + \ln \Pr y'$$

т. е. классы разделяет гиперплоскость $w^\top x = c$: $w^\top x > c$,
где $w = (\mu_y - \mu_{y'})^\top \boldsymbol{\Sigma}^{-1}$, а c – некоторая константа.

Введем *линейную дискриминантную функцию*:

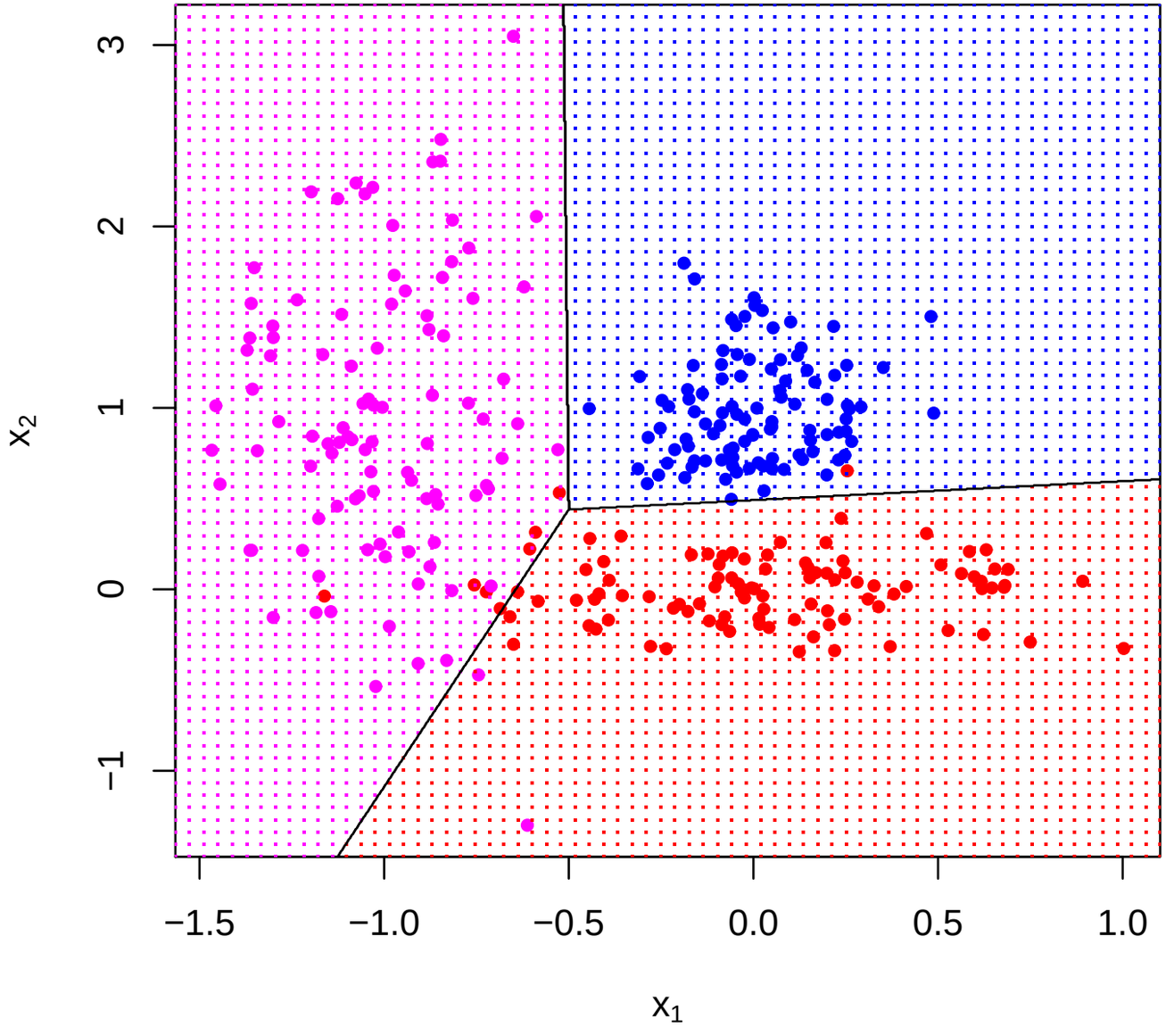
$$\delta_y(x) = -\frac{1}{2}(x-\mu_y)^\top \boldsymbol{\Sigma}^{-1}(x-\mu_y) + \frac{1}{2}x^\top \boldsymbol{\Sigma}^{-1}x + \ln \Pr y = \mu_y^\top \boldsymbol{\Sigma}^{-1}x - \frac{1}{2}\mu_y^\top \boldsymbol{\Sigma}^{-1}\mu_y + \ln \Pr y$$

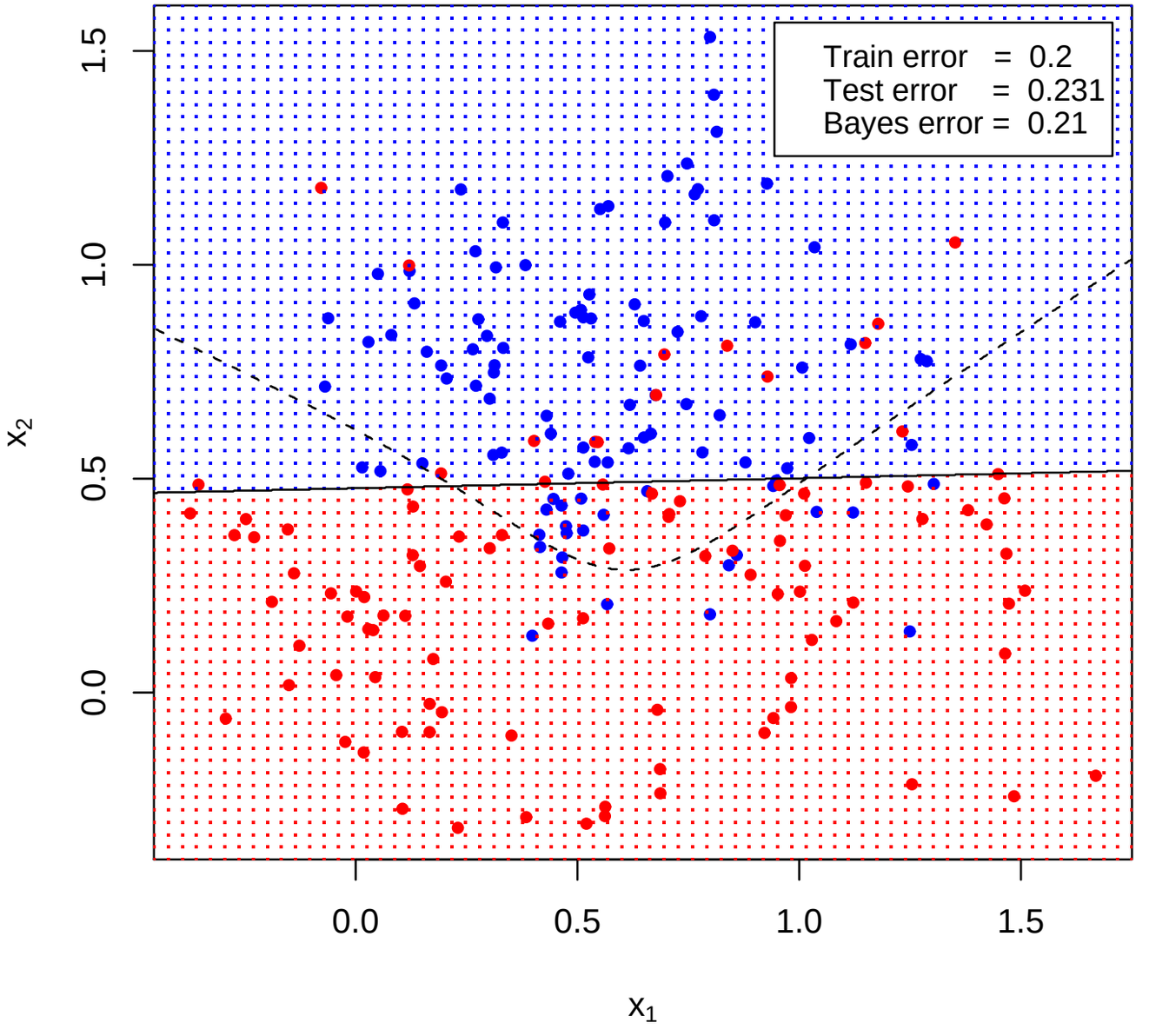
Классификатор: $f(x) = \operatorname{argmax}_y \delta_y(x)$.

где $\rho(x, x') = \sqrt{(x-x')^\top \boldsymbol{\Sigma}^{-1}(x-x')}$ – это *расстояние Махаланобиса*.

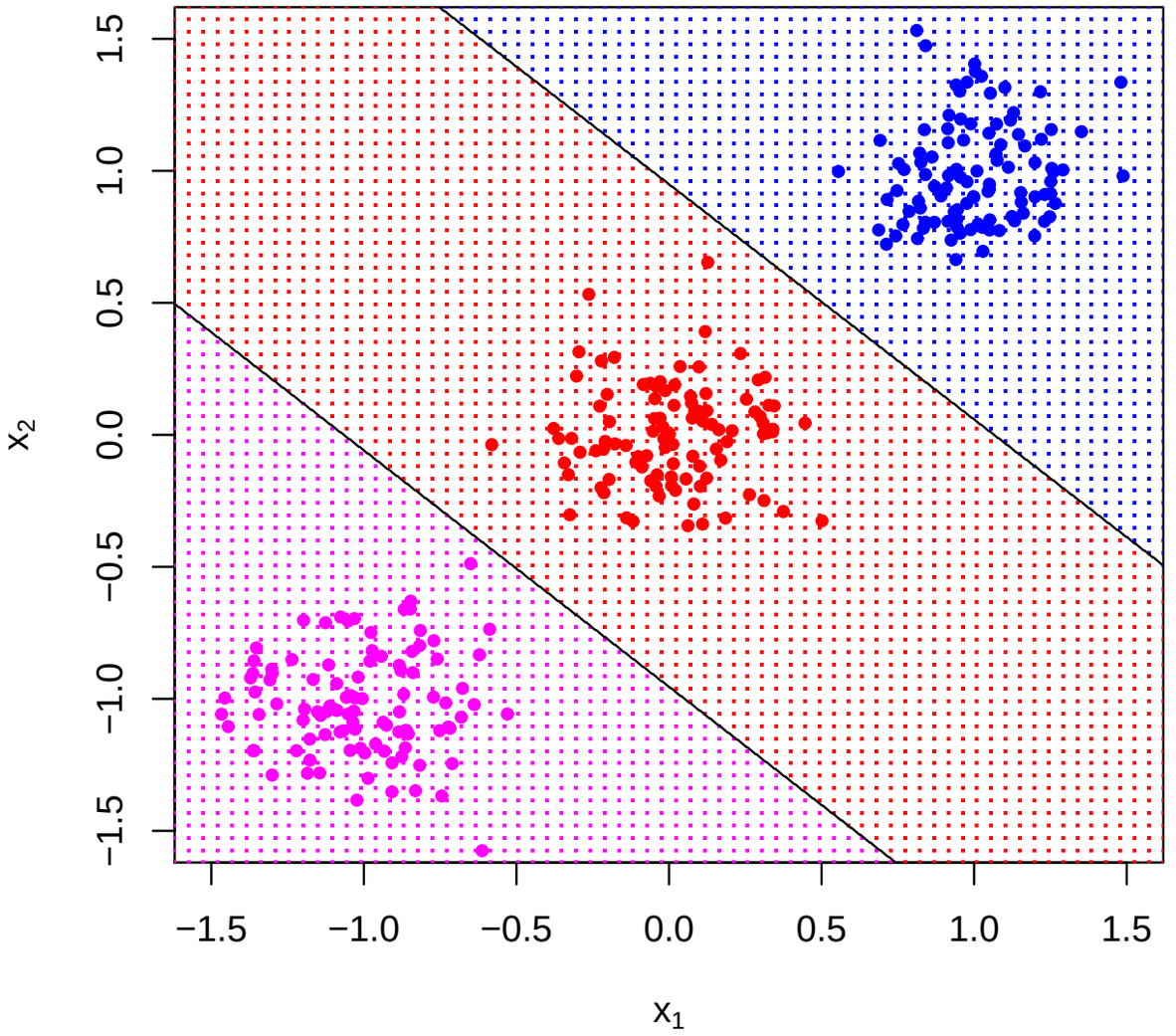
На практике мы не знаем параметров нормального распределения, но можем оценить их по обучающей выборке:

$$\widehat{\Pr}(y) = \frac{N_y}{N}, \qquad \widehat{\mu}_y = \sum_{y^{(i)}=k} \frac{x^{(i)}}{N_y}, \qquad \widehat{\Sigma} = \frac{1}{N} \sum_{k=1}^{K} \sum_{y^{(i)}=k} (x^{(i)} - \mu_k)(x^{(i)} - \mu_k)^\top.$$


3
2
1
0
−1
x_2
−1.5
−1.0
−0.5
0.0
0.5
1.0
x_1


Train error = 0.2
Test error = 0.231
Bayes error = 0.21
0.0
0.5
1.0
1.5
x_2
x_1

LDA



8.1.2. Квадратичный дискриминантный анализ

Рассмотрим теперь

$$p(x\,|\,y) = \frac{1}{(2\pi)^{d/2}\sqrt{\det \boldsymbol{\Sigma}_y}} e^{-\frac{1}{2}(x-\mu_y)^\top \boldsymbol{\Sigma}_y^{-1}(x-\mu_y)},$$

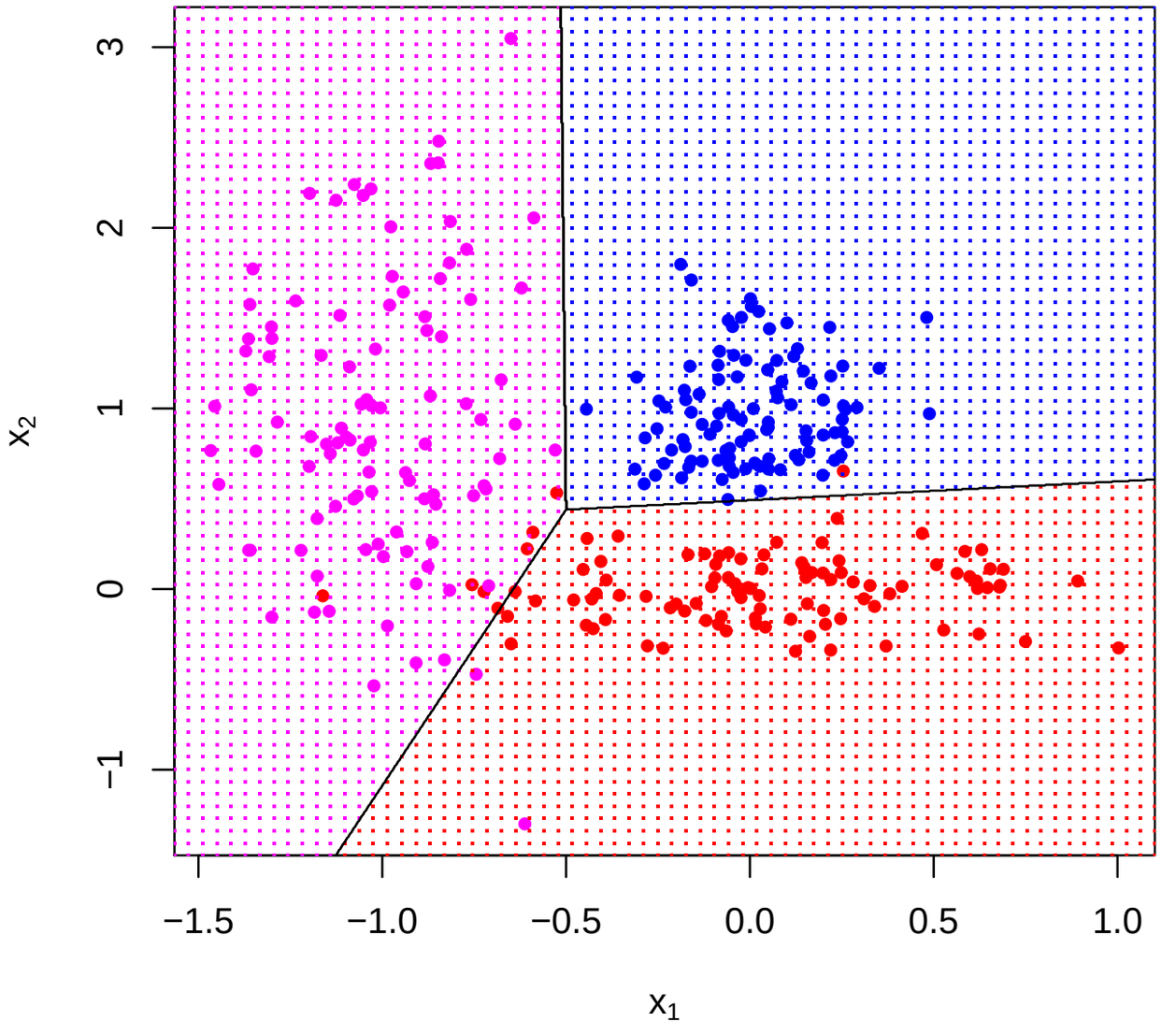
не предполагая, что $\boldsymbol{\Sigma}_y$ равны между собой.

Проводя аналогичные рассуждения, придем к понятию *квадратичной дискриминантной функции*

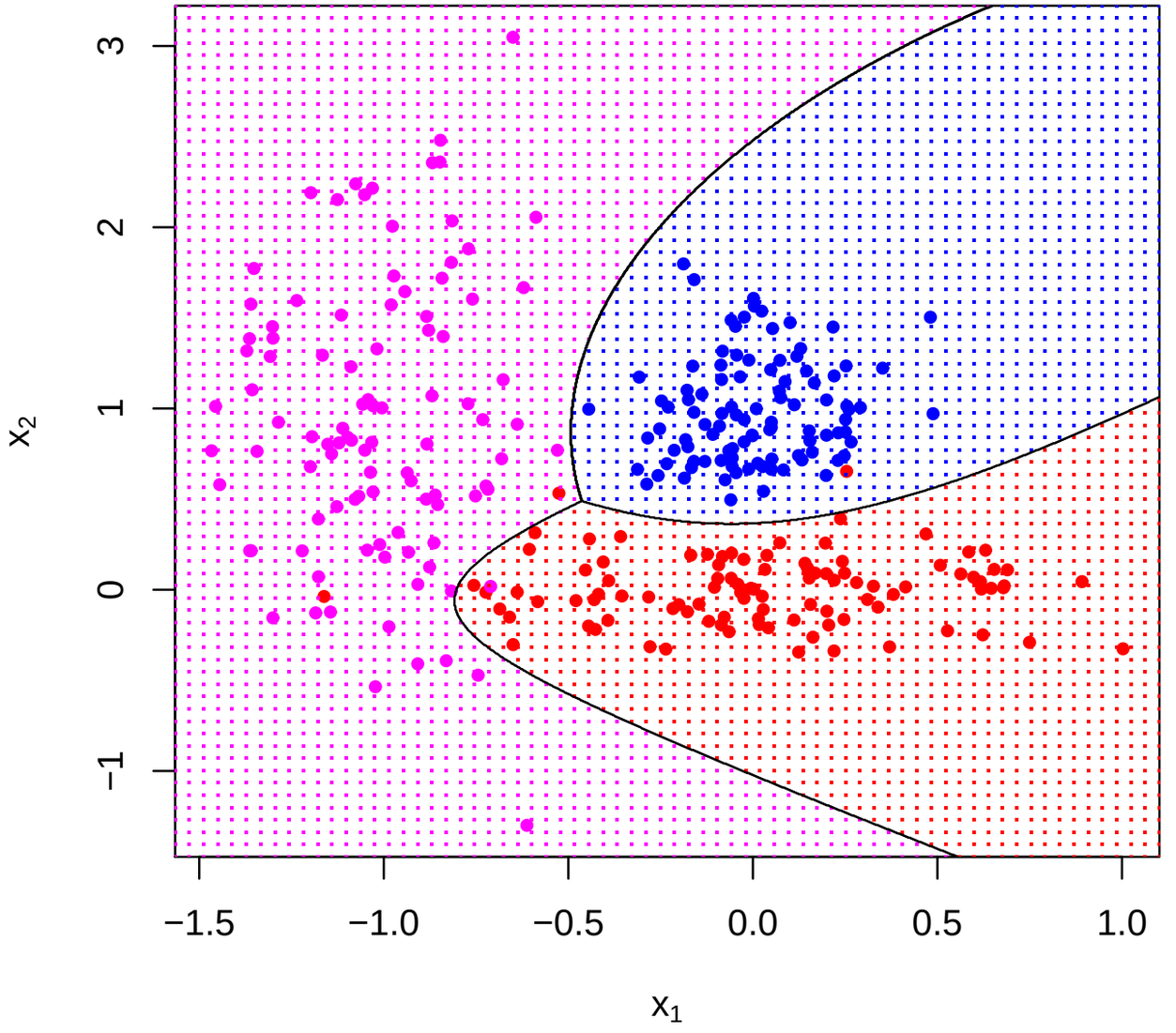
$$\delta_y(x) = -\frac{1}{2}\ln\det\boldsymbol{\Sigma}_y - \frac{1}{2}(x-\mu_y)^\top\boldsymbol{\Sigma}_y^{-1}(x-\mu_y) + \ln \mathsf{Pr}(y)$$

Поверхность, разделяющая два класса y и y' описывается уравнением 2-го порядка $\delta_y(x) = \delta_{y'}(x)$.

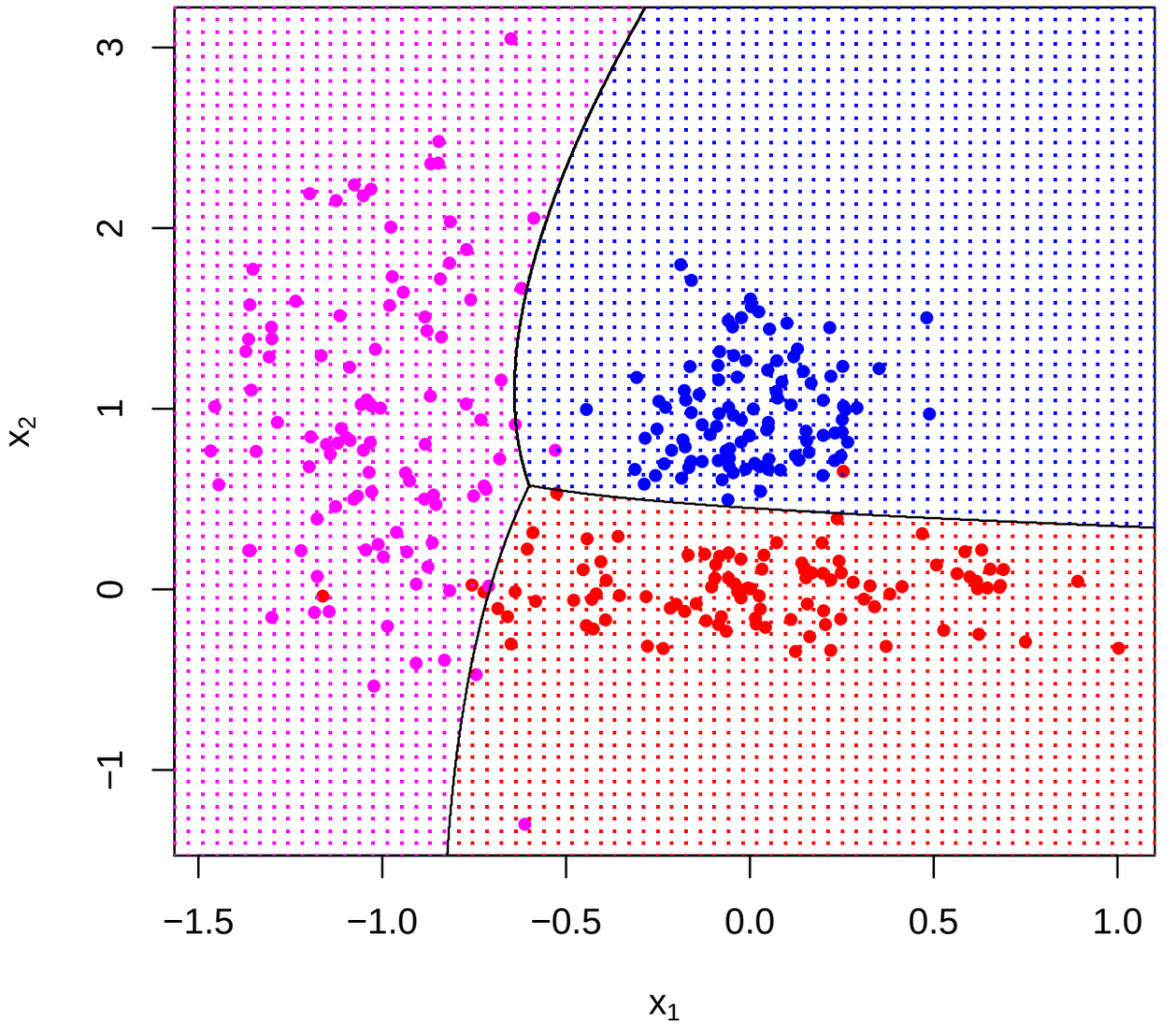
LDA



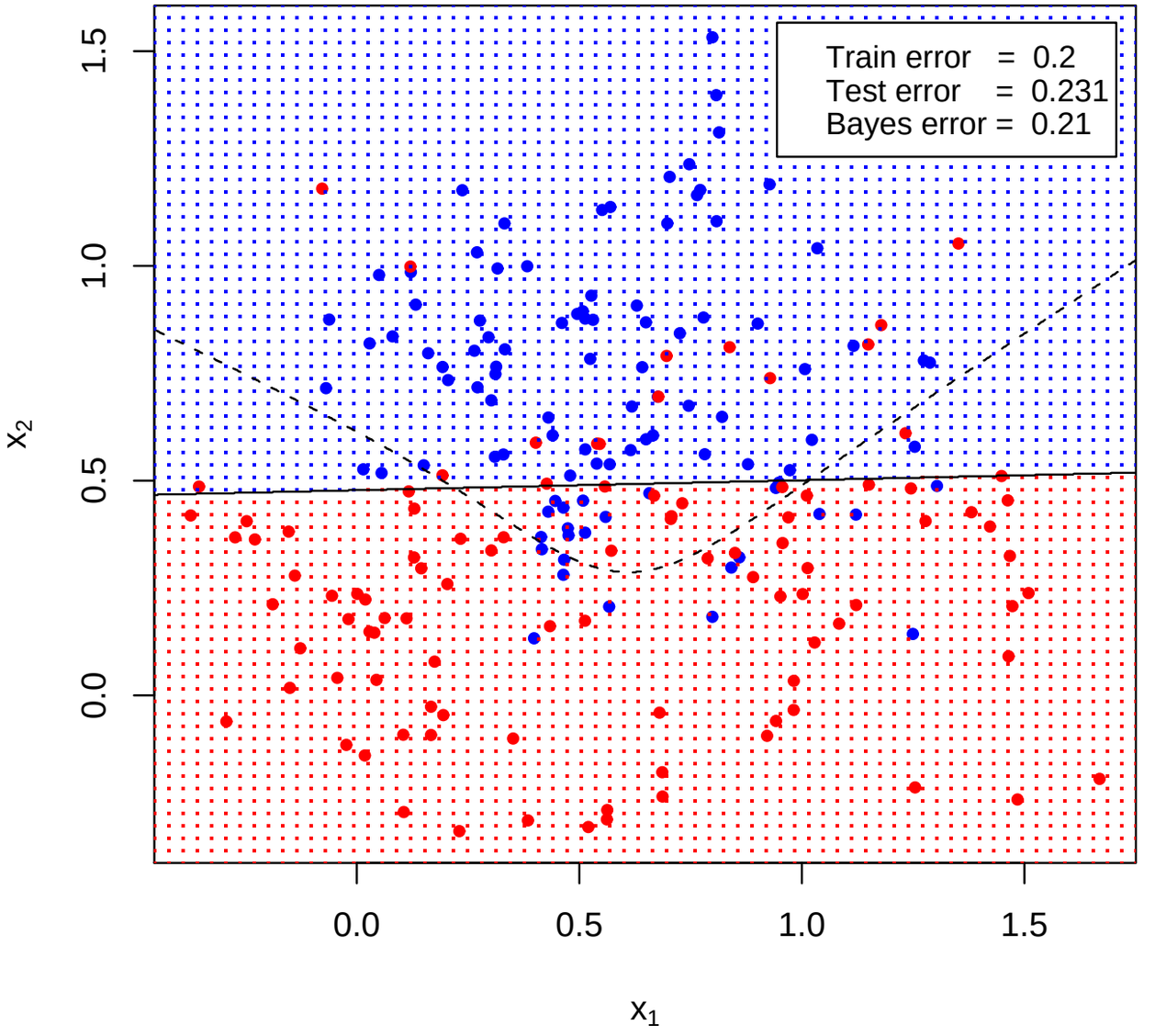
QDA



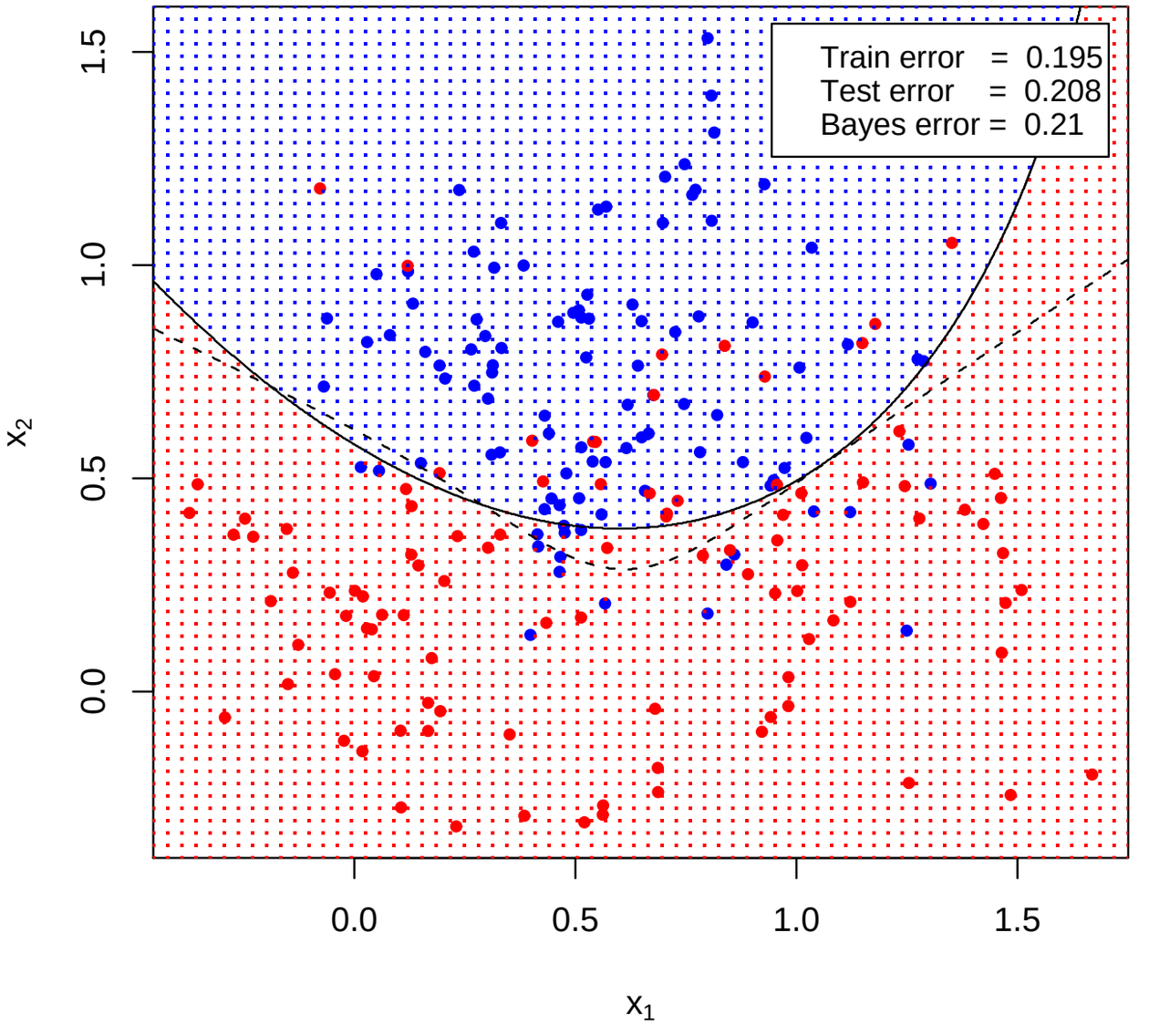
LDA $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2$



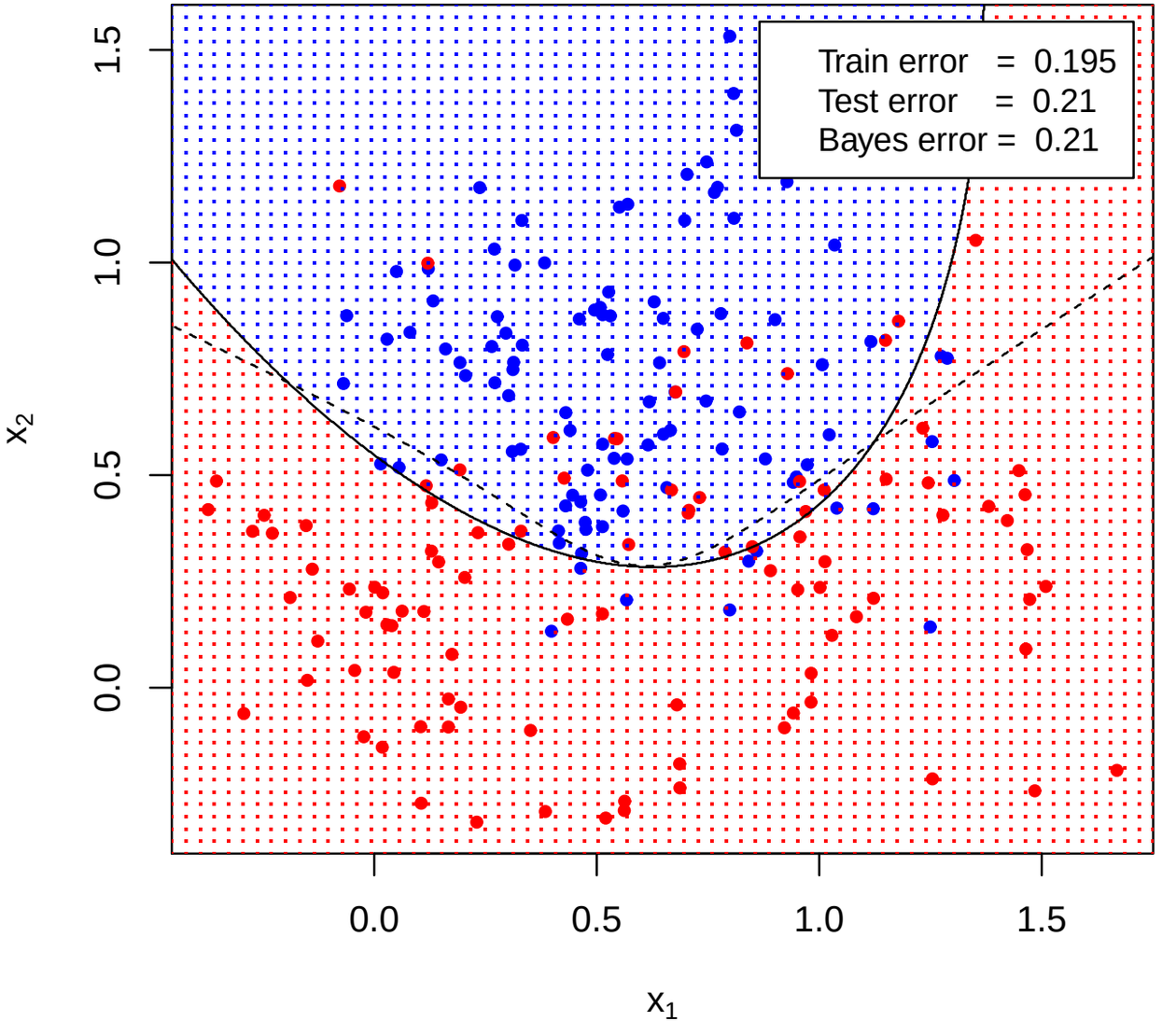
LDA


Train error = 0.2
Test error = 0.231
Bayes error = 0.21
x_2
x_1
0.0
0.5
1.0
1.5

QDA


Train error = 0.195
Test error = 0.208
Bayes error = 0.21
0.0
0.5
1.0
1.5
x_1
x_2

LDA $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2$



8.1.3. Другой взгляд на LDA (дискриминант Фишера)

Максимизация отношения межклассовой дисперсии к внутриклассовой эквивалентна LDA:

$$\max_w \frac{w^\top \mathbf{B} w}{w^\top \mathbf{W} w}$$

или

$$\max_w w^\top \mathbf{B} w \qquad \text{при ограничении } w^\top \mathbf{W} w = 1,$$

где

$$\mathbf{B} = \sum_{k=1}^{K} N_k (\mu_k - \mu)(\mu_k - \mu)^\top, \qquad \mathbf{W} = \sum_{k=1}^{K} \sum_{y^{(i)}=k} (x^{(i)} - \mu_k)(x^{(i)} - \mu_k)^\top$$

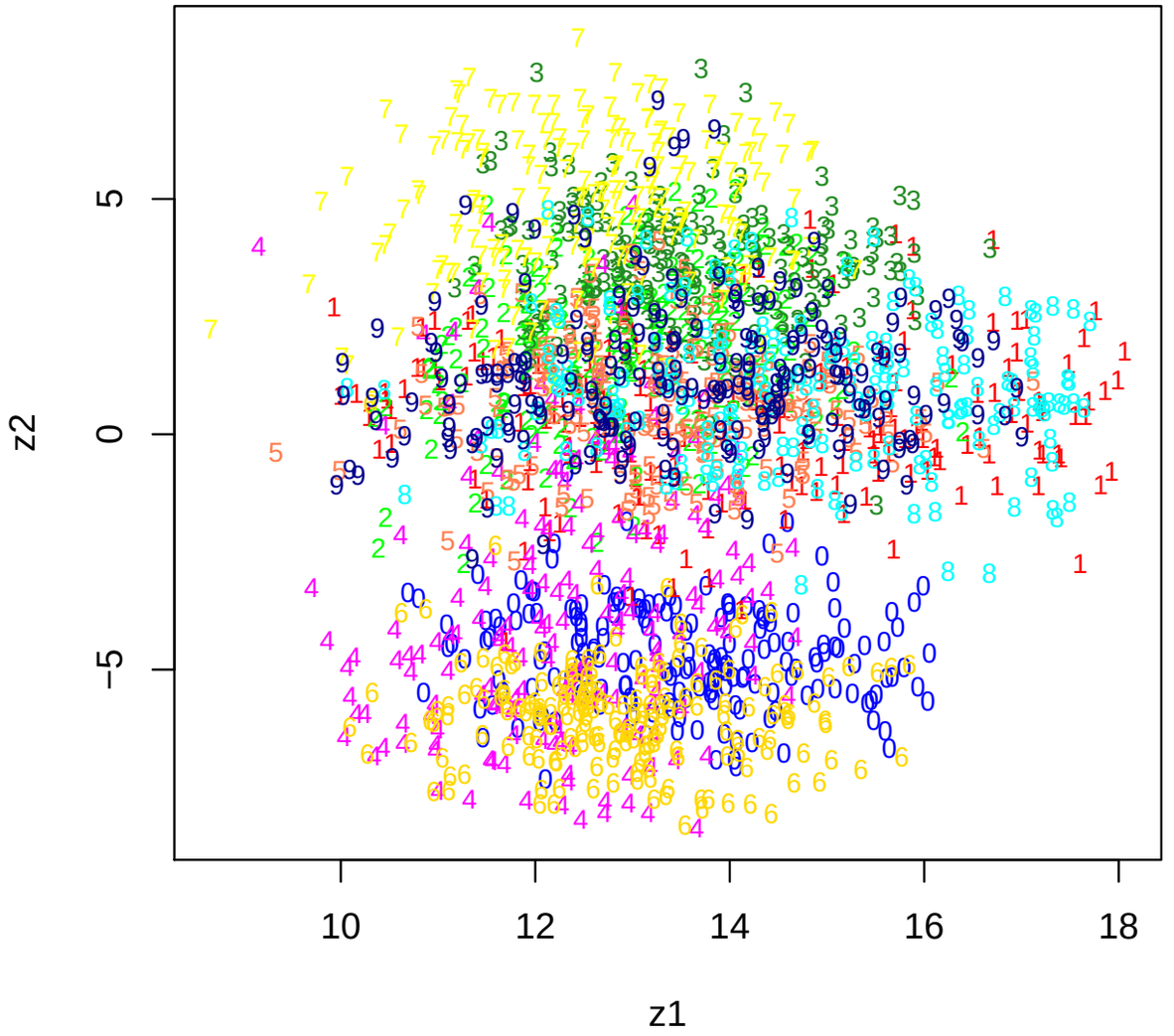
Это эквивалентно *обобщенной задаче на собственные значения*:

$$\mathbf{B} w = \lambda \mathbf{W} w \qquad \Leftrightarrow \qquad \mathbf{W}^{-1} \mathbf{B} w = \lambda w.$$

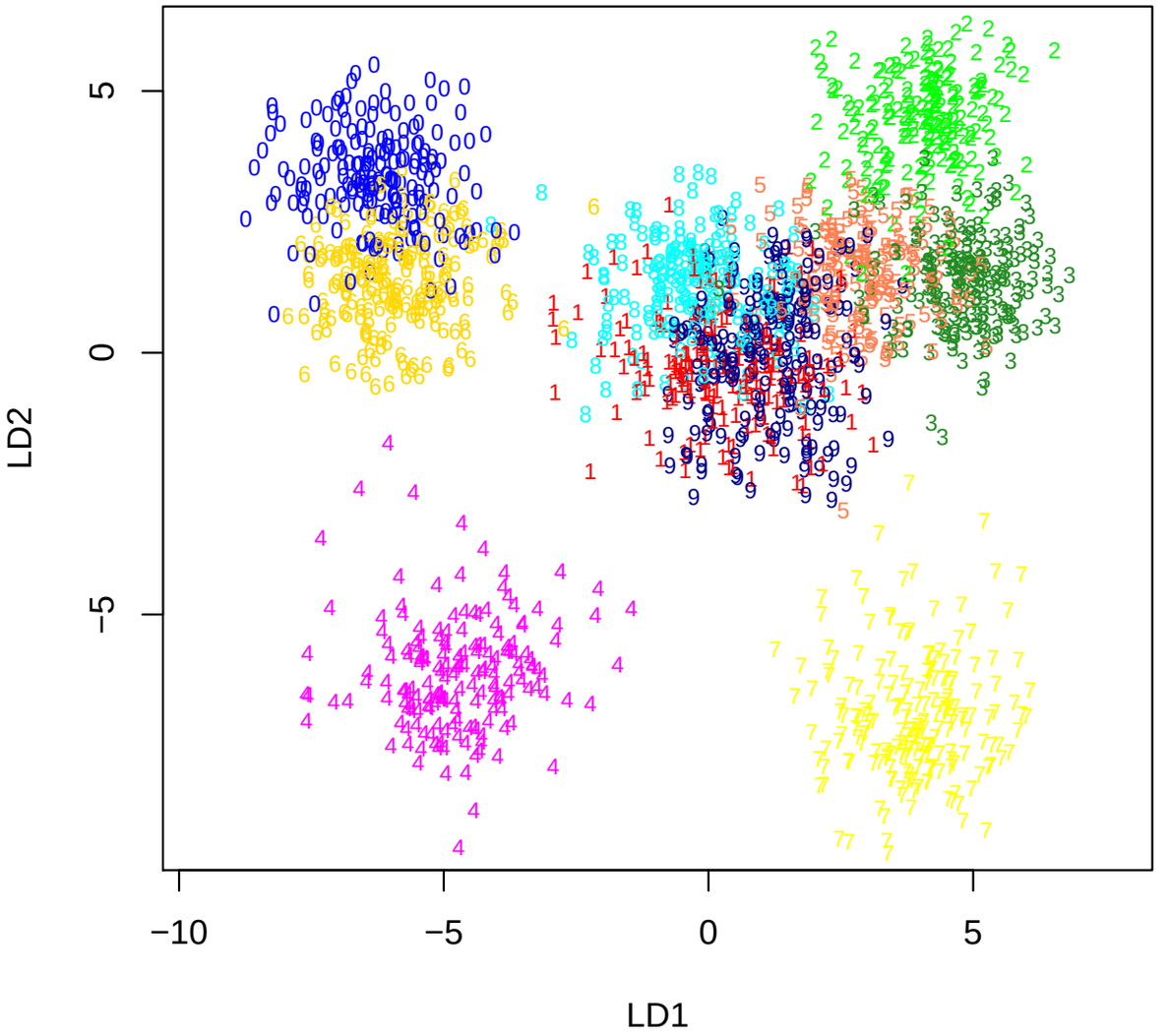
Решением является вектор w_1, соответствующий максимальному значению λ_1.

Можно найти вектор w_2, ортогональный w_1, на котором отношение Рэлея максимально и т. д.

Рукописные цифры. PCA $d = 1024$, $k = 2$



Рукописные цифры. LDA

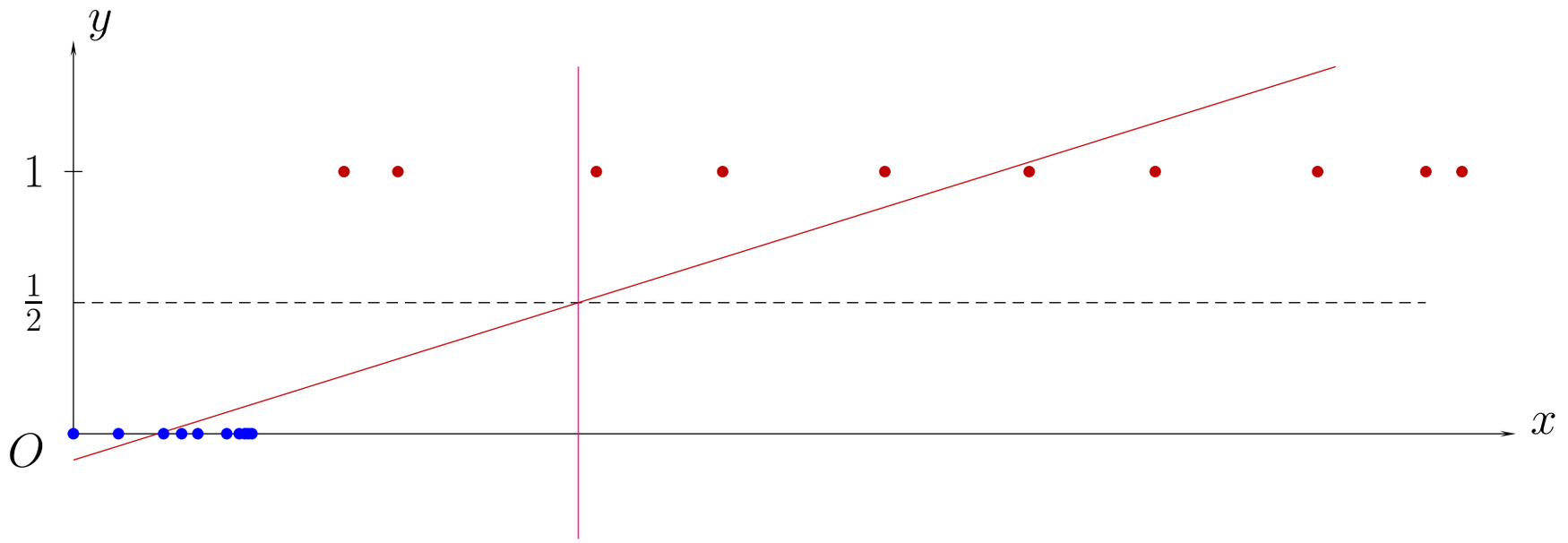


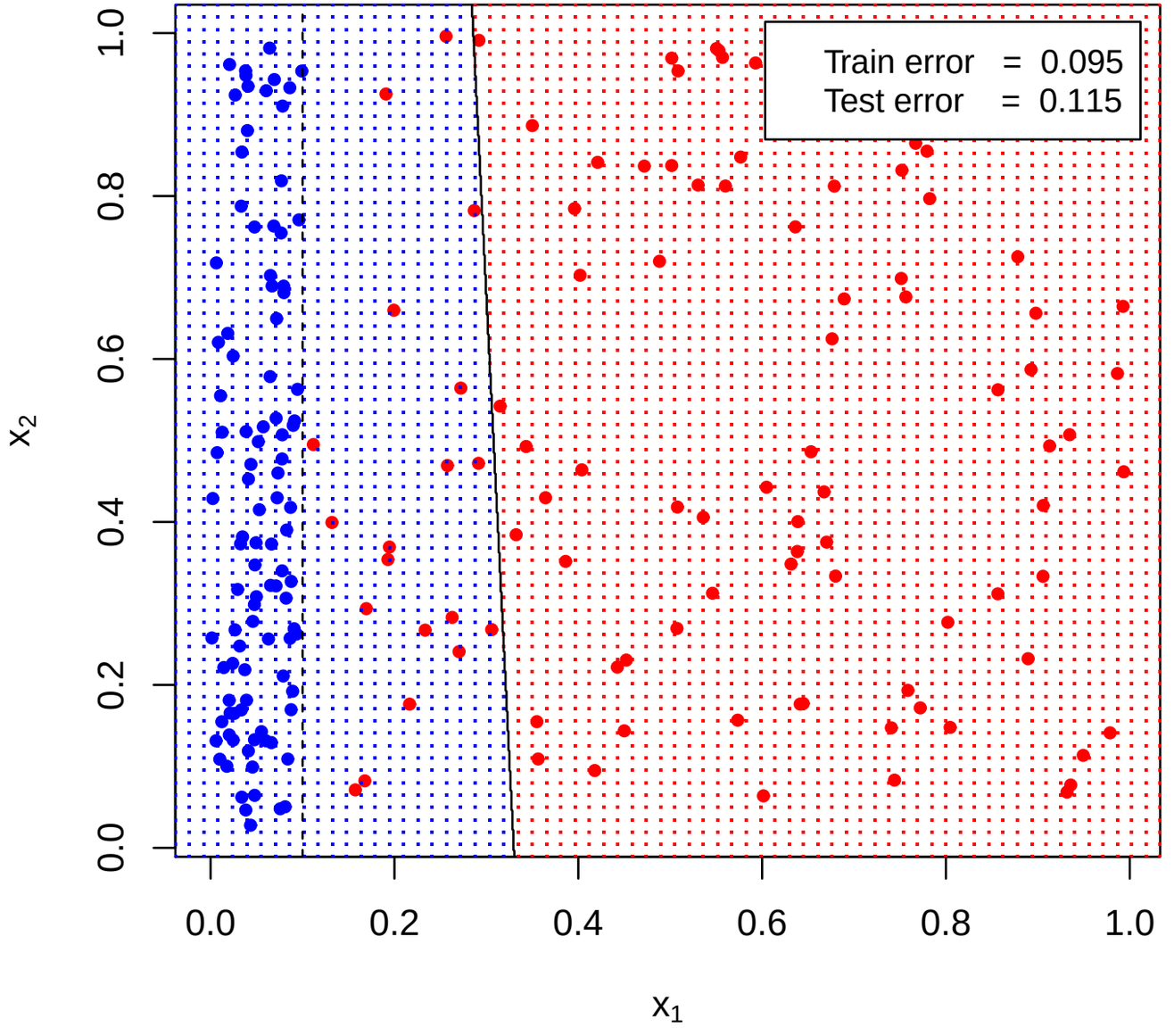
8.2. МНК для задачи классификации

(дискриминантный метод)

Нельзя ли использовать линейную регрессию (МНК) для решения задачи классификации?

$K = 2$




Train error = 0.095
Test error = 0.115
0.0
0.2
0.4
0.6
0.8
1.0
x_2
x_1

$K > 2$

Сопоставим каждому классу k *характеристический*, или *индикаторный*, *вектор* $(y_1, y_2, \ldots, y_K)$, в котором $y_k = 1$, а $y_i = 0$ при $i \neq k$.

Собрав вместе индикаторные векторы объектов обучающей выборки, получим матрицу $\mathbf{Y}$ размера $N \times K$, называемую *индикаторной*.

Таким образом, $\mathbf{Y}$ состоит только из нулей и единиц и каждая строка имеет ровно одну единицу.

Как обычно, $\mathbf{X}$ — матрица размера $N \times (d+1)$, первый столбец которой состоит из единиц, а последующие представляют собой векторы из обучающей выборки.

Применяя метод наименьших квадратов одновременно к каждому столбцу матрицы $\mathbf{Y}$, получаем значения

$$\widehat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}.$$

Для каждого столбца $\mathbf{y}_k$ матрицы $\mathbf{Y}$ получим свой столбец коэффициентов $\widehat{\beta}_k$. Соберем их в матрицу $\widehat{\mathbf{B}}$ размера $(d+1) \times K$.

Имеем

$$\widehat{\mathbf{B}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

Объект x будем классифицировать согласно следующему правилу:

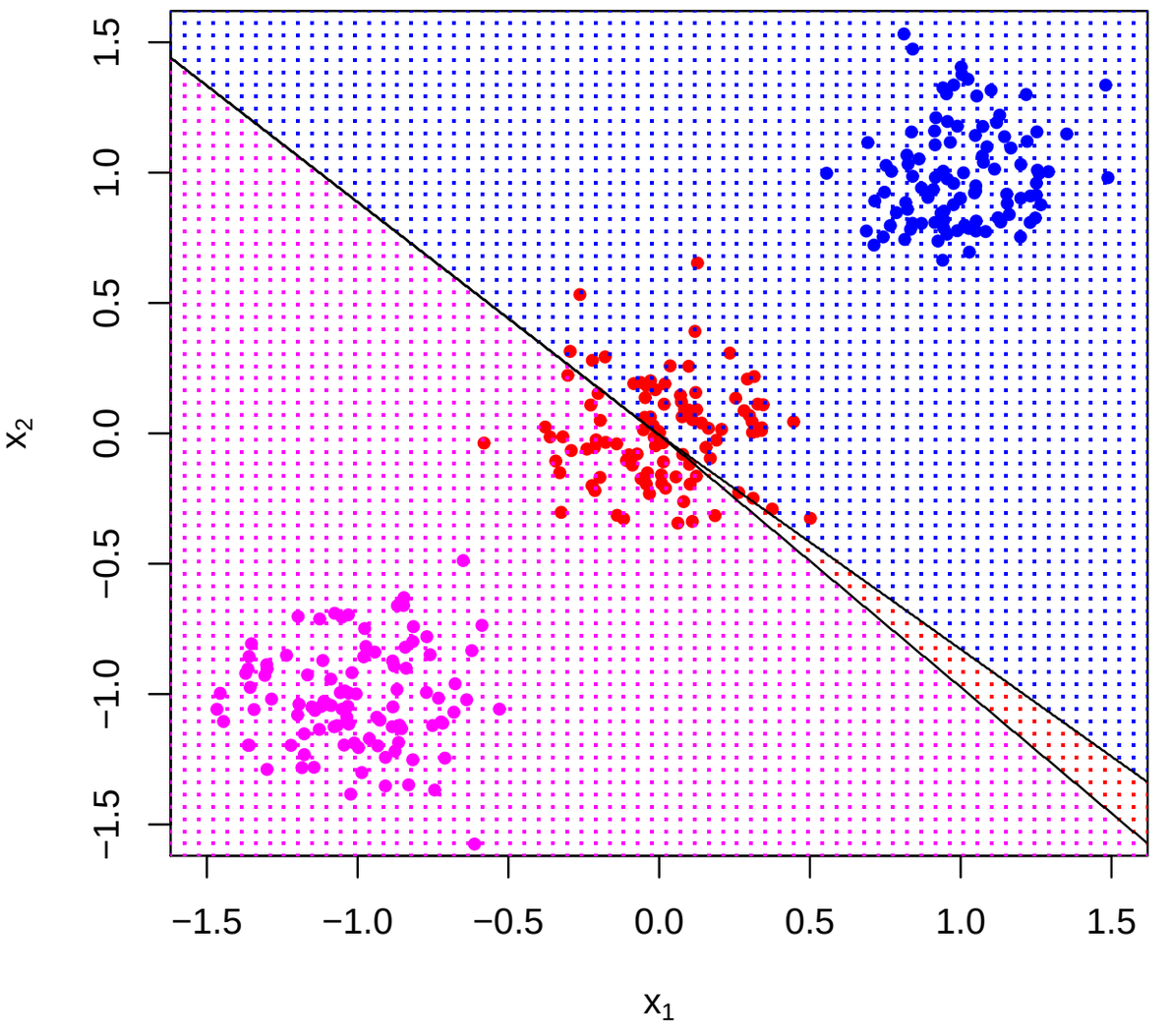
Вычислим вектор-строку длины K

$$g(x) = (1, x)\, \widehat{\mathbf{B}}.$$

Отнесем x к классу

$$f(x) = \operatorname*{argmax}_k g_k(x).$$

При $K \geq 3$ линейная регрессия может «не замечать» некоторых хорошо отделимых классов.



Замечание 8.1 Пусть также $\widehat{\beta}_0$, $\widehat{\beta}$ — параметры, полученные с помощью линейной регрессии с критерием (см. ниже)

$$\sum_{i=1}^{N}(y^{(i)} - \beta_0 - \beta^\top x^{(i)})^2 \to \min\,.$$

Можно показать, что вектор $\widehat{\beta}$ коллинеарен вектору $\widehat{\mathbf{\Sigma}}^{-1}(\widehat{\mu}_2 - \widehat{\mu}_1)$.

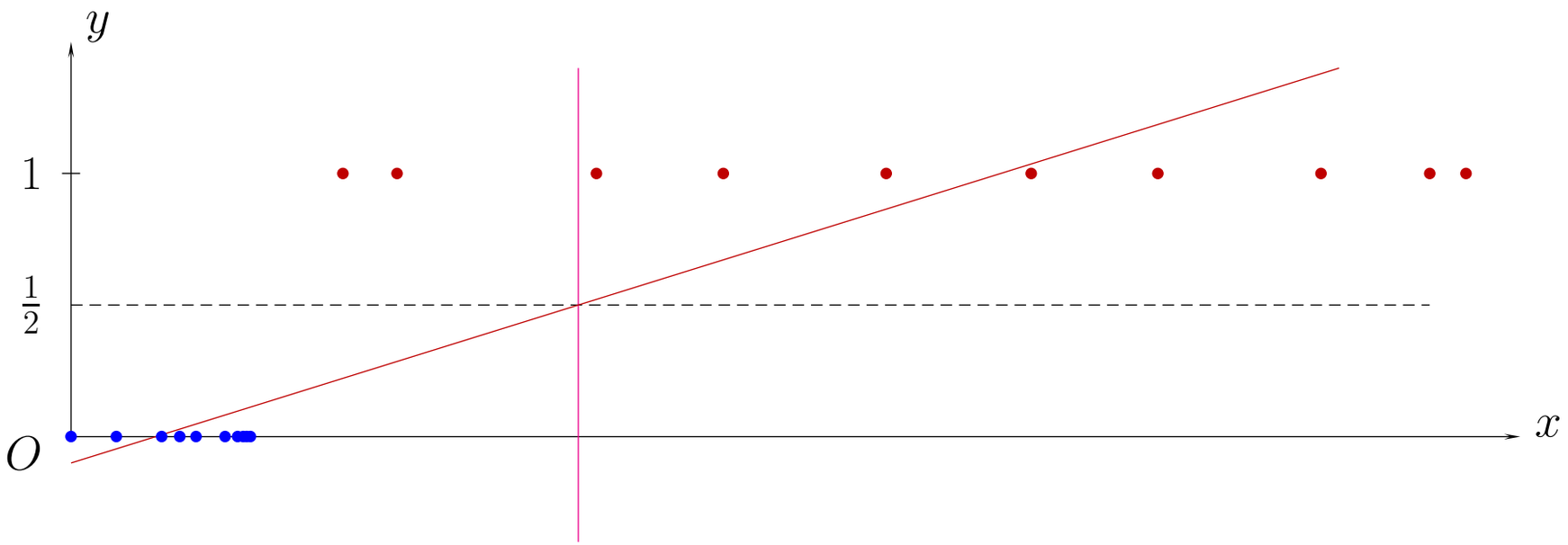
Таким образом, разделяющие гиперплоскости, полученные с помощью линейной регрессии и методом *LDA*, параллельны (но при $N_1 \neq N_2$ не совпадают друг с другом).

Это уже не так при числе классов, большем 2.

Если имеется более 2-х классов, то, напомним, линейная регрессия при классификации может «не замечать» некоторых из них.

LDA-метод свободен от этого недостатка.

МНК:



Хочется: