Machine Learning Mechods
P160B124

LDA and QDA

assoc. prof. dr. Tomas Iešmantas

tomas.iesmantas@ktu.lt

Room 319

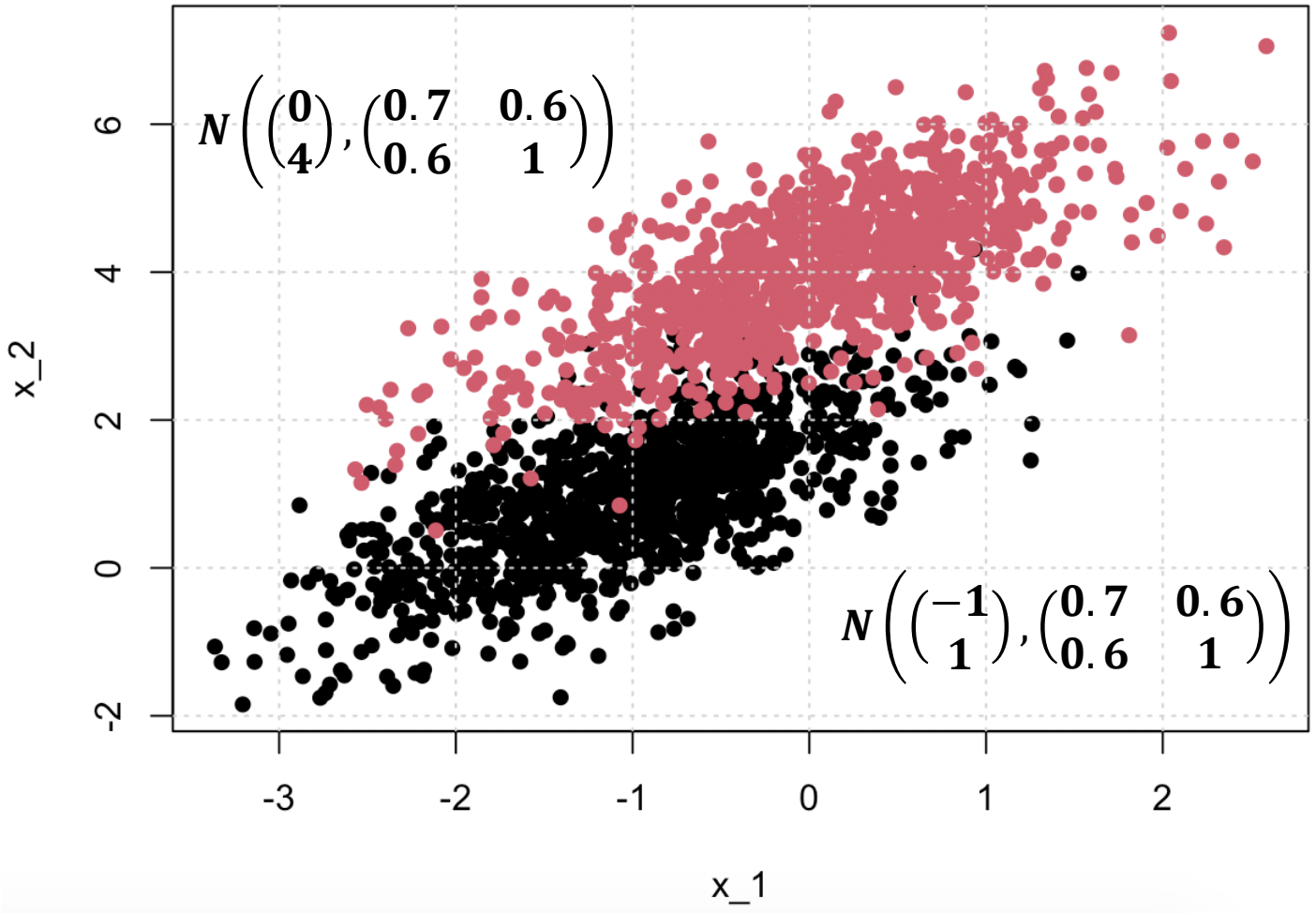
Linear discriminant analysis

- Logistic regression – directly modelling the probability $P(y = 1|x)$.

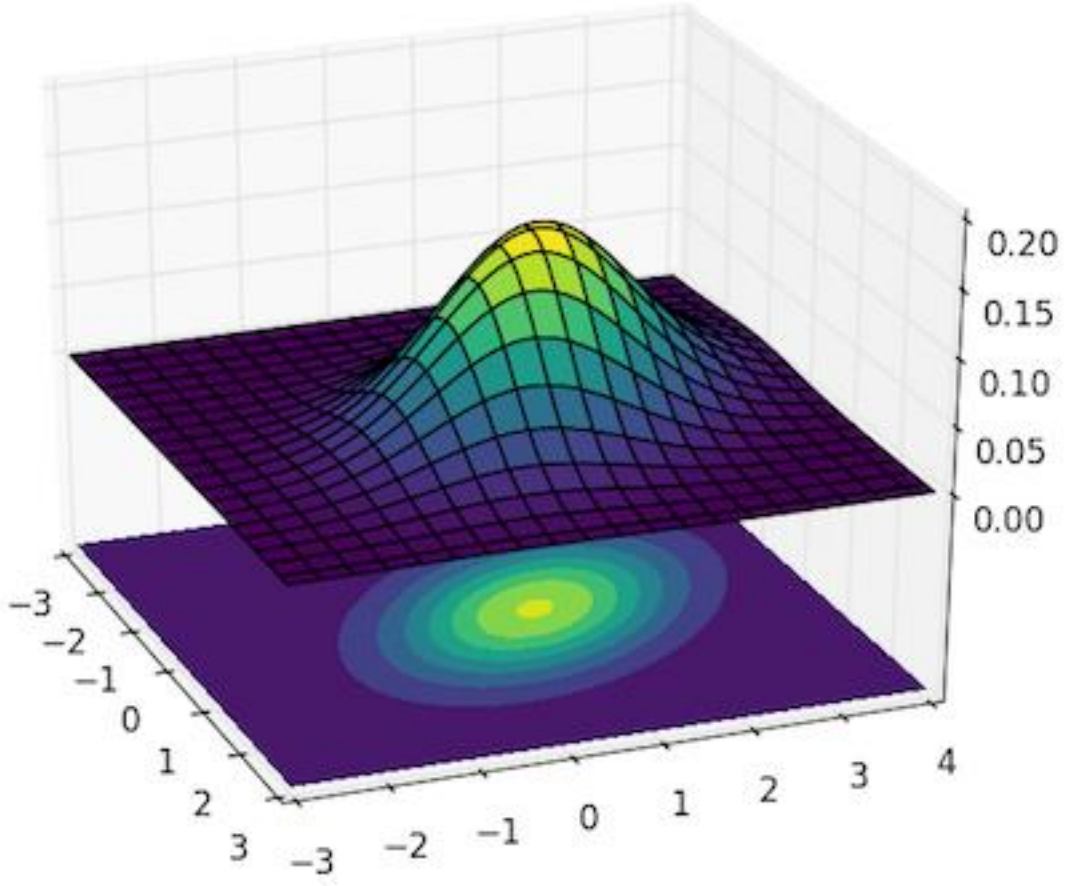
- **Idea**: instead of modelling directly $p_k(x) = P(y = k|x)$ (as in logistic regression), we could try ***modelling the data itself and then derive*** the model for probability.

Linear discriminant analysis

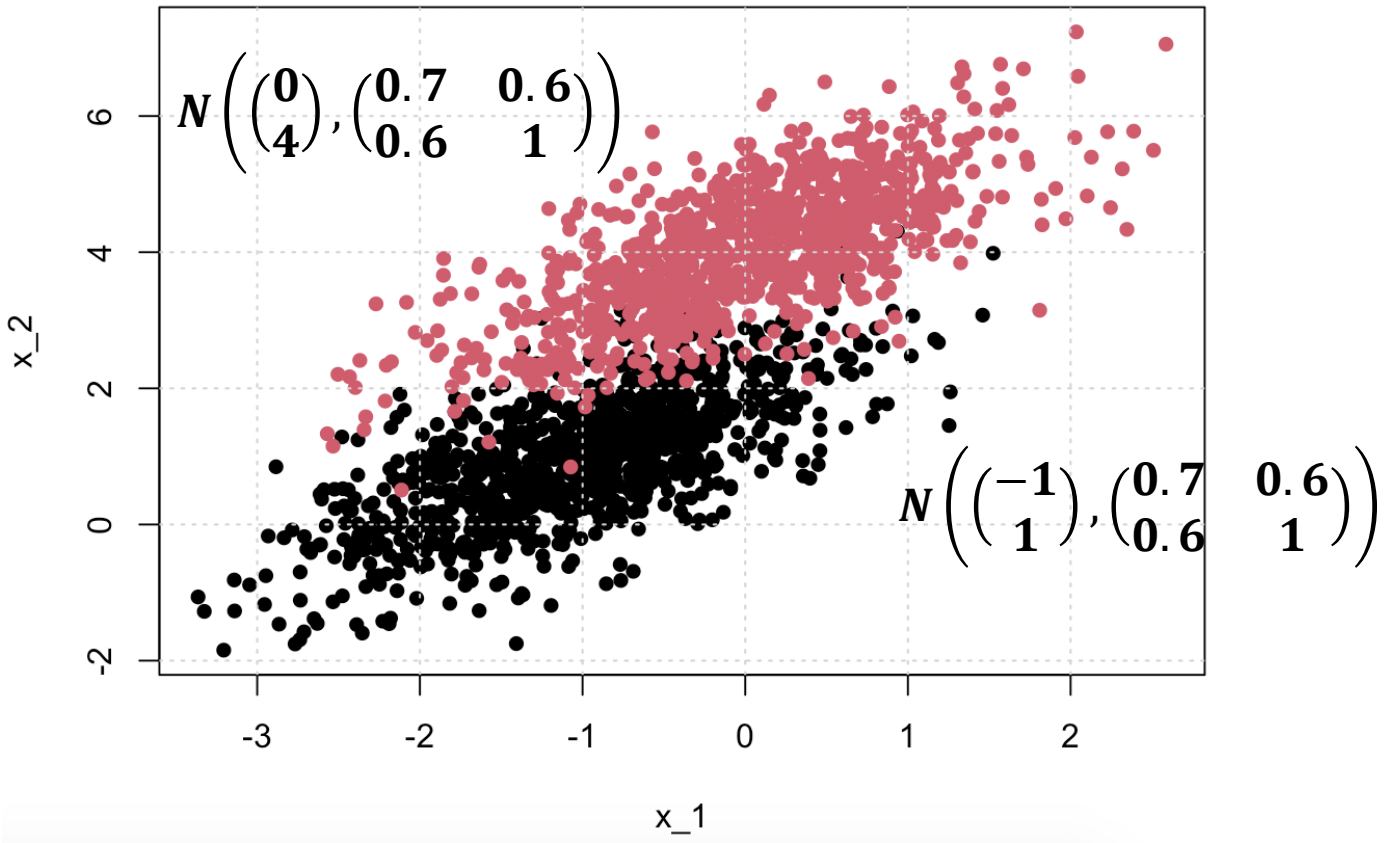
- Consider 2 features as in figure below and a binary classification problem.



Quick reminder about Gaussian data

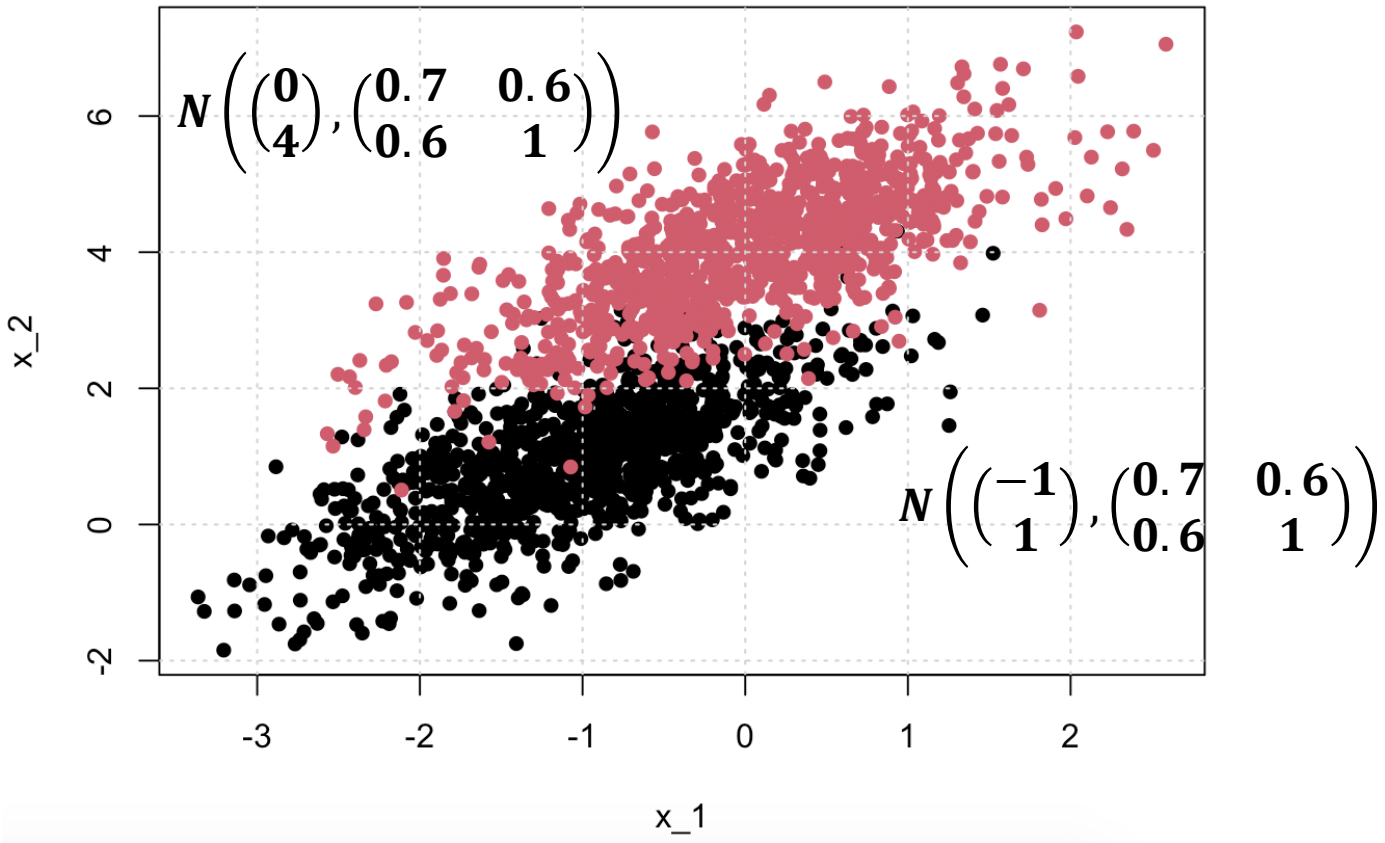


Linear discriminant analysis



- (We assume that) we know the stochastic nature of the data for each of the classes;
- In the above data case, features are generated from the normal distribution with different means but equal covariances.

Linear discriminant analysis



- Thus, distributions are dependent on the class, that the features belongs to.
- We will mark this fact with class-conditional distributions

$$f_k(x) = f(x|y = k).$$

Linear discriminant analysis

- We have two classes, thus class conditional distributions are:

$$f_1(x) = N\left(\begin{pmatrix}0\\4\end{pmatrix}, \begin{pmatrix}0.7 & 0.6\\0.6 & 1\end{pmatrix}\right)$$

$$f_2(x) = N\left(\begin{pmatrix}-1\\1\end{pmatrix}, \begin{pmatrix}0.7 & 0.6\\0.6 & 1\end{pmatrix}\right)$$

- ***A side note***: whenever we try to model the data itself, we have a class of **generative models** (Deep generative adversarial networks are an example of this class).

Linear discriminant analysis

- OK, we assume to know the nature of the data. So what?
- The probability can be obtained by the Bayes theorem:

$$P(y = k|x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

where $\pi_k = P(y = k)$ - ***a priori*** **class probability**.

- By assuming the nature of the data, we no longer need to assume the functional form of the probability function. It is given by the Bayes theorem and therefore always true … given that we truly know the nature of the data.
- ***a priori*** **class probability** is a probability of a class when we know nothing about the features. Most often it is equal simply to the frequency of a class in population.

Linear discriminant analysis

- **Bayes theorem**:

$$P(y = k|x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

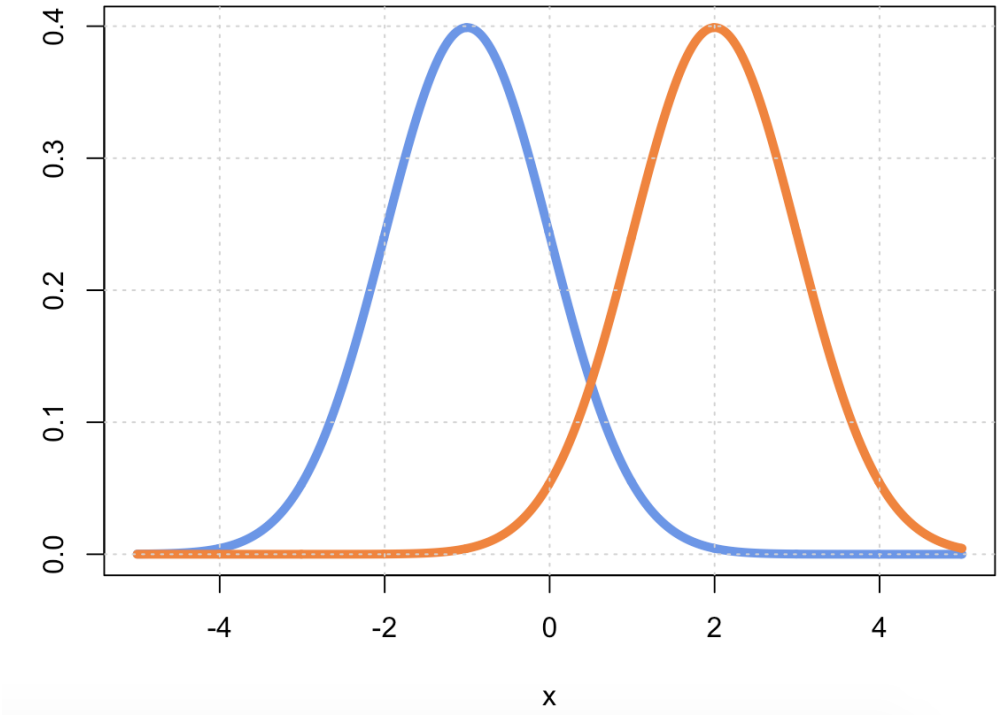
- Class probabilities are easy – just use fraction of each class in the training dataset.
- Densities are much more challenging unless we assume some simple forms (like Gaussian, exponential, etc.).

Linear discriminant analysis

- Assume univariate features with Gaussian distributions

$$x|y = k \sim \mathcal{N}\left(\mu_k, \sigma_k^2\right), \qquad k = 1, 2.$$

- First, assume equal variances: $\sigma_1 = \sigma_2 = \sigma$. But $\mu_1 \neq \mu_2$.



Linear discriminant analysis

- We will use a classification rule: pick the most probable class. This is equivalent to: pick class $k = 1$ if $P(k = 1|x) \geq P(k = 2|x)$
- The classification change at those points x, where

$$P(k = 1|x) = P(k = 2|x).$$

In other words, the decision boundary is made of those points x, which satisfy the above equation. Let's find this decision boundary.

Linear discriminant analysis

- Our **assumptions**:

$$f_1(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu_1)^2}, f_2(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu_2)^2}$$

- **Bayes theorem**:

$$P(y = k|x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

- **Decision boundary**: $P(k = 1|x) = P(k = 2|x)$ becomes:

$$\frac{\pi_1 f_1(x)}{\sum_{l=1}^{K} \pi_l f_l(x)} = \frac{\pi_2 f_2(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

Or:

$$\pi_1 f_1(x) = \pi_2 f_2(x)$$

Take logarithms:

$$x \cdot \frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} + \ln(\pi_1) = x \cdot \frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2} + \ln(\pi_2)$$

Question 1

It is known that one dimensional features come from two Gaussian distribution with equal variances ($\sigma = 1$) but different means ($\mu_1 = -1,\ \mu_2 = 2$). Find the decision boundary between two classes (assume, that in population, classes are equally likely, i.e. $\pi_1 = \pi_2 = 0.5$).

a) x=0.5;

b) x=0;

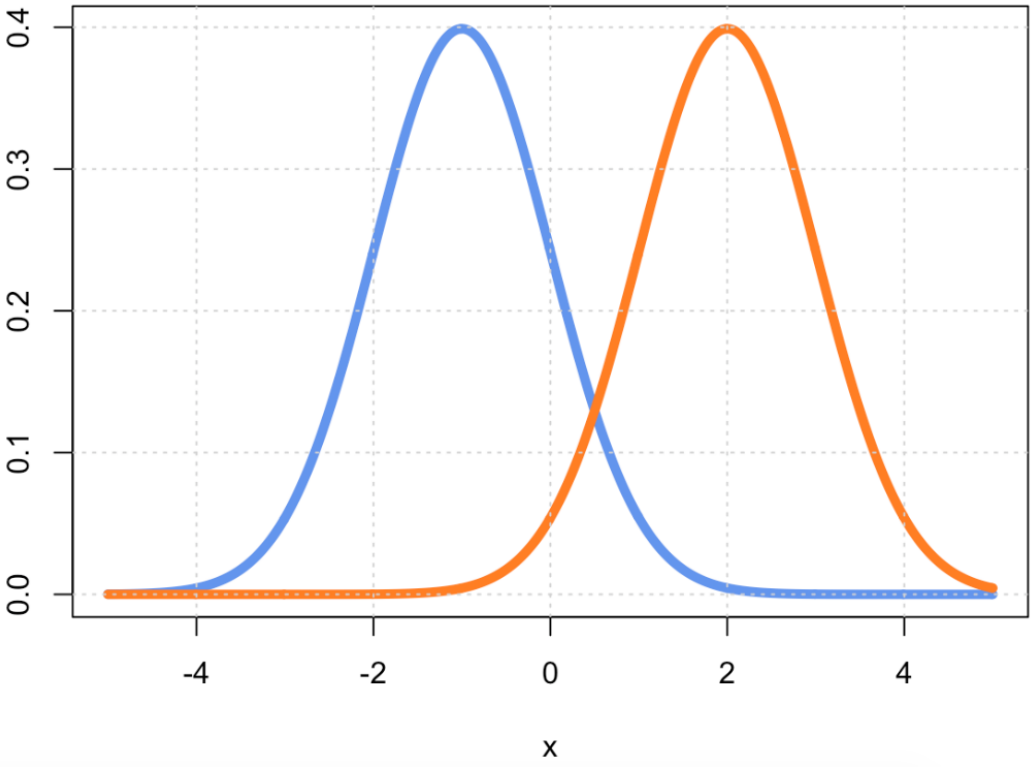
c) x=1;

d) Correct answer is not given

$$x \cdot \frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} + \ln(\pi_1) = x \cdot \frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2} + \ln(\pi_2)$$

Question 1

It is known that one dimensional features come from two Gaussian distribution with equal variances ($\sigma = 1$) but different means ($\mu_1 = -1,\ \mu_2 = 2$). Find the decision boundary between two classes (assume, that in population, classes are equally likely, i.e. $\pi_1 = \pi_2 = 0.5$).



Linear discriminant analysis

- Assume univariate features with Gaussian distributions

$$x|y = k \sim \mathcal{N}\left(\mu_k, \sigma_k^2\right), \qquad k = 1, \cdots, K.$$

- First, assume equal variances: $\sigma_1 = \sigma_2 = \cdots = \sigma_K = \sigma$. Then the decision rule "pick the most likely class" in this case assigns an observation x to the class for which

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \ln(\pi_k)$$

is largest.

- We do not know the exact values of parameters of Gaussians $(\mu_1, \mu_2, \cdots, \mu_K, \sigma)$. K+1 unknown parameters.

Linear discriminant analysis

- **Linear discriminant analysis classifier** is a classifier, which assigns to the feature x the most likely class, with x assumed to be Gaussian and unknown parameters are estimated by the following expressions (ML method):

$$\hat{\mu}_k = \frac{1}{N_k} \sum_{i:y_i=k} x_i \, , \hat{\sigma}^2 = \frac{1}{N-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

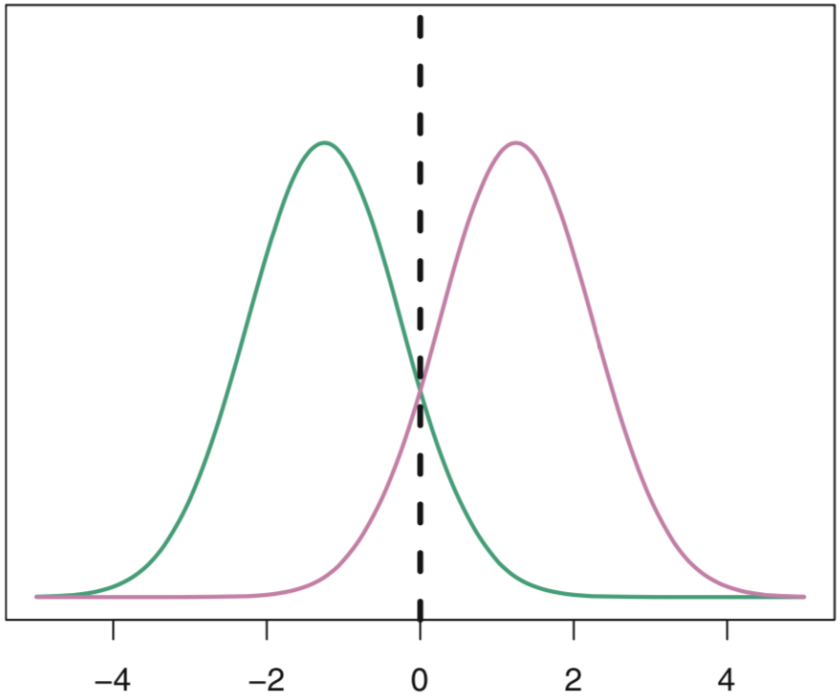
- If no other prior information about class distribution in the population is available, then we simply use frequencies:

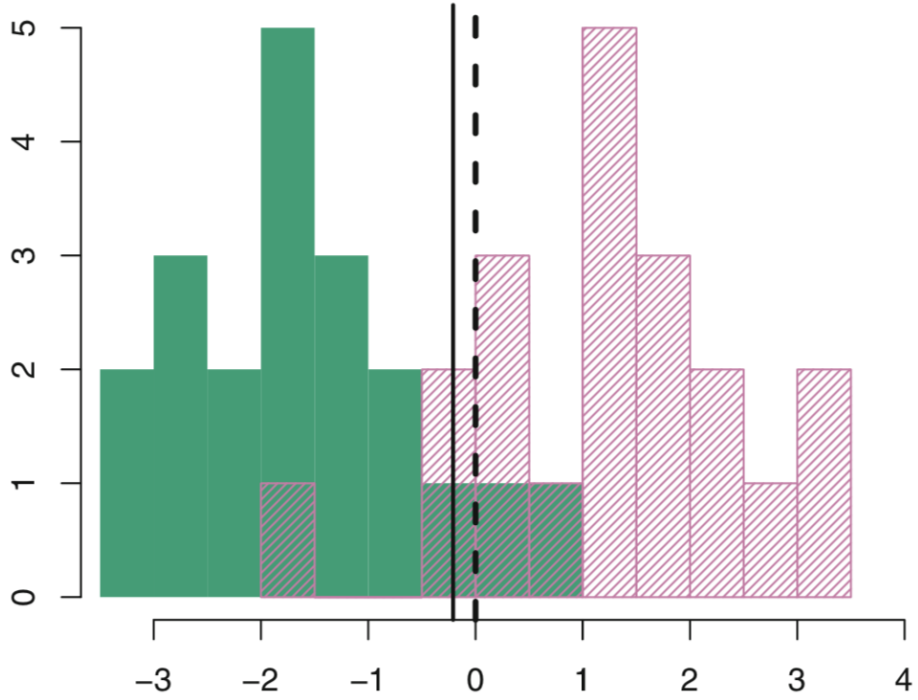
$$\hat{\pi}_k = \frac{N_k}{N}.$$

- Linear because $\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \ln(\pi_k)$ is linear in terms of x.

Linear discriminant analysis

- Example: $\mu_1 = -1.25, \mu_2 = 1.25, \sigma_1^2 = \sigma_2^2 = 1$
- True decision boundary is $x = 0$

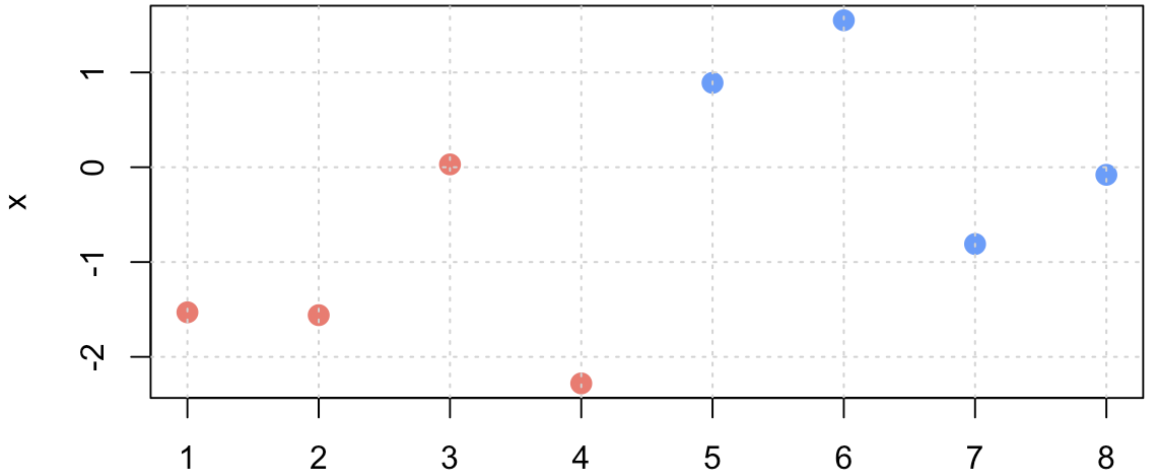




Question 2

- Assume a training dataset:

x	y
-1.53	0
-1.56	0
0.03	0
-2.28	0
0.89	1
1.55	1
-0.81	1
-0.08	1



What is the decision boundary for LDA assuming $\hat{\sigma} = 1$ and classes are equally likely?

a) -0.67375;

b) -0.47375;

c) 0.67375;

d) 0.47375

$$x \cdot \frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} + \ln(\pi_1) = x \cdot \frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2} + \ln(\pi_2)$$

Question 3

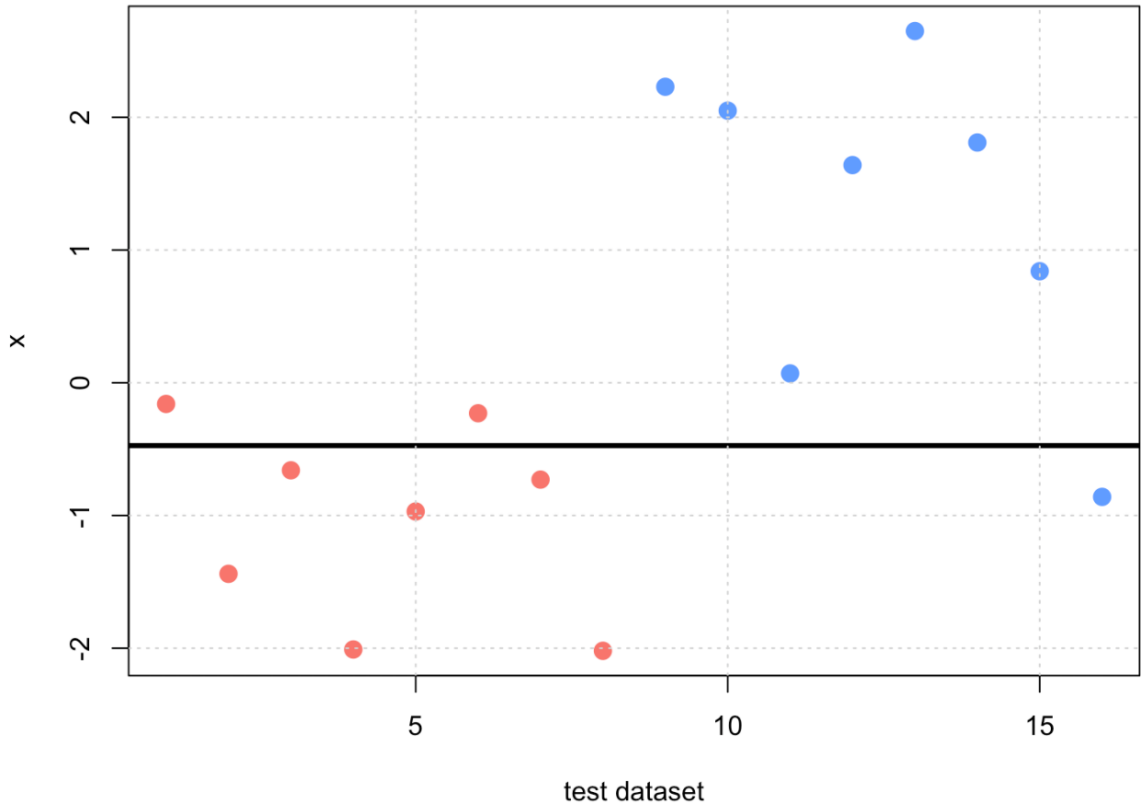
- What is the overall accuracy of LDA classifier (decision boundary is plotted in black) and accuracies for each class?

a) Overall: 56.25 %; Class1: 75%; Class2: 87.5%

b) Overall: 81.25%; Class1: 75%; Class2: 87.5%

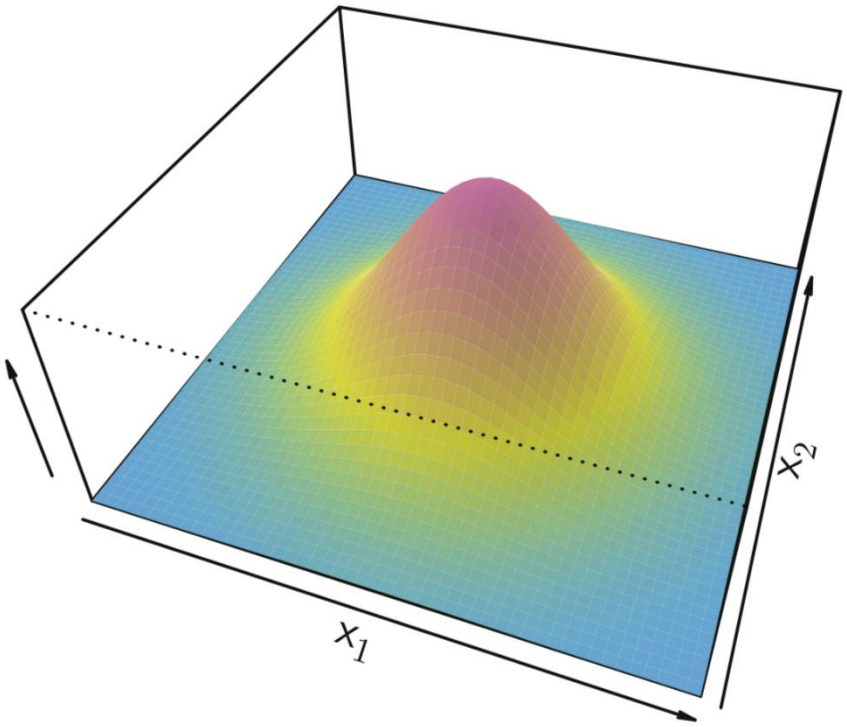
c) Overall: 81.25%; Class1: 25%; Class2: 12.5%

d) Overall: 43.75%; Class1: 25%; Class2: 12.5%

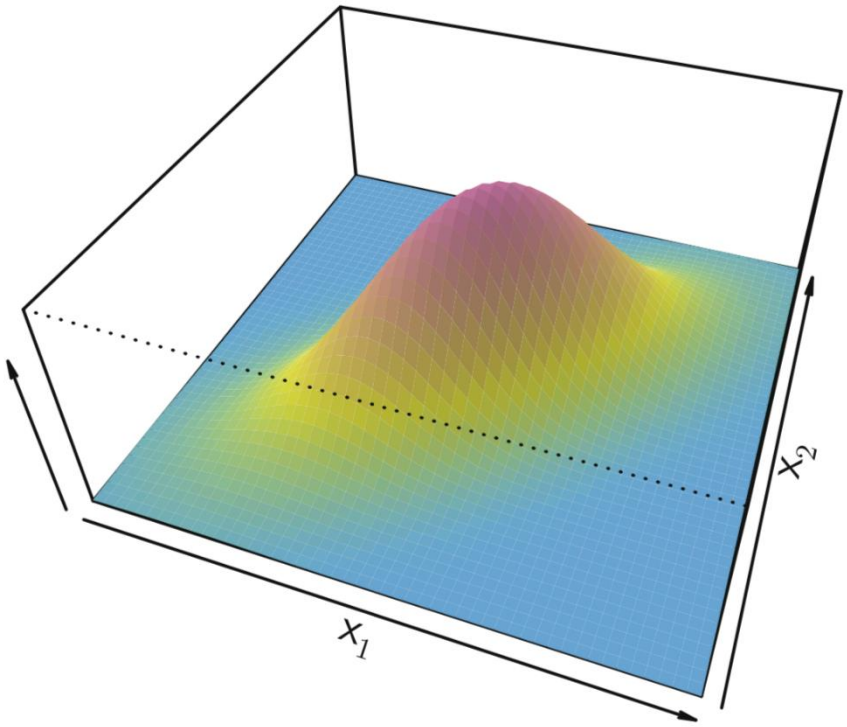


Linear discriminant analysis

- What if we have more features than just one? i.e. $x = (x_1, \cdots, x_m)$.
- We assume that these feature vectors are coming from class conditional multivariate Gaussian distributions, i.e. class specific means but with common covariance structure.



No correlation



0.7 correlation

Linear discriminant analysis

- Gaussian distribution:

$$f(x) = \frac{1}{(2\pi)^{m/2}|\Sigma|^{1/2}} exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

- A classifier assign an observation x to the class for which

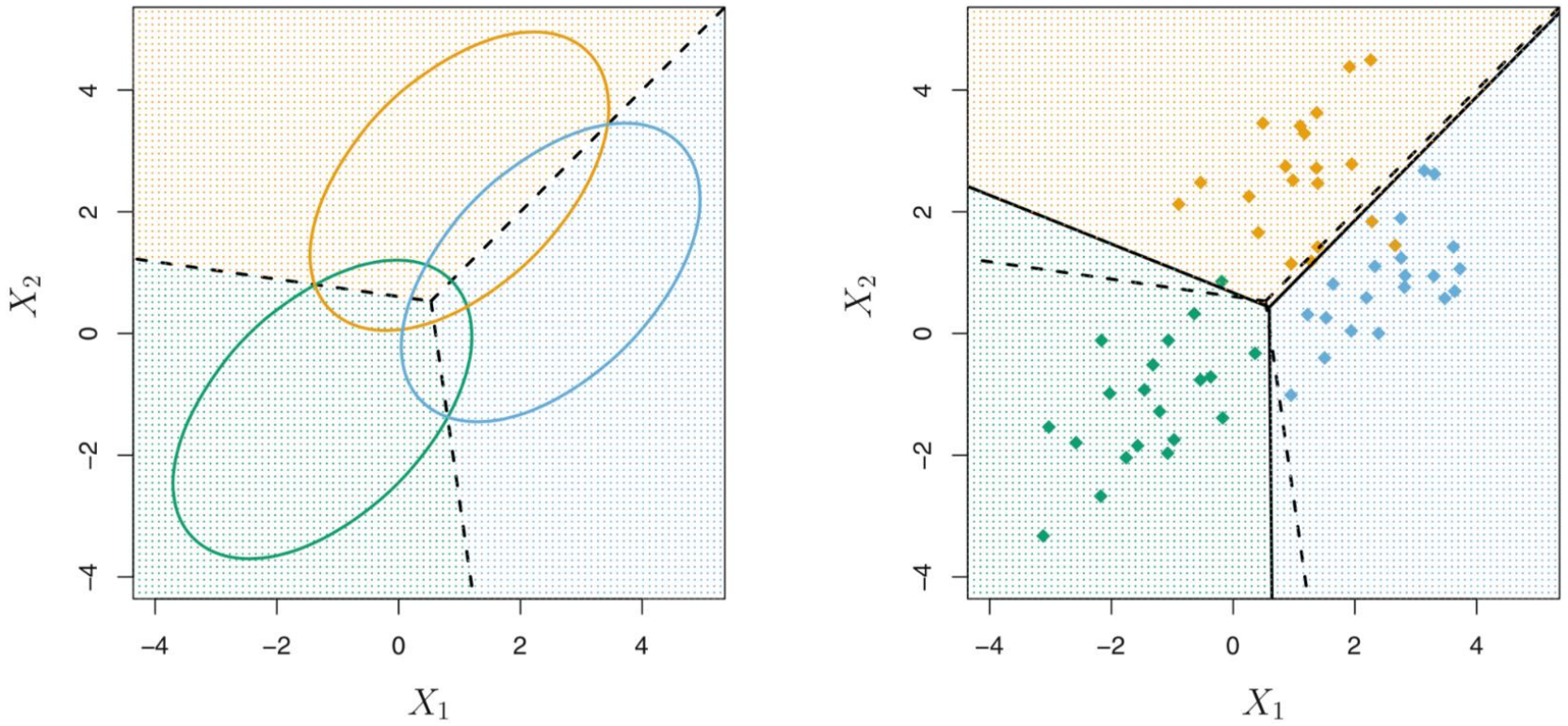
$$\delta_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \ln \pi_k$$

is largest.

- Decision boundary can be obtained by solving pairs of equations:

$$\delta_i(x) = \delta_j(x), i \neq j$$

Linear discriminant analysis



- Three equally- sized Gaussian classes are shown with class-specific mean vectors and a common covariance matrix. Dashed lines are true decision boundaries.
- Number of unknown parameters: $K \cdot m + m(m+1)/2$.
- If $m = 100$ and $K = 2$, we have 5250 unknown parameters!

Question 4

- Find decision boundary when:

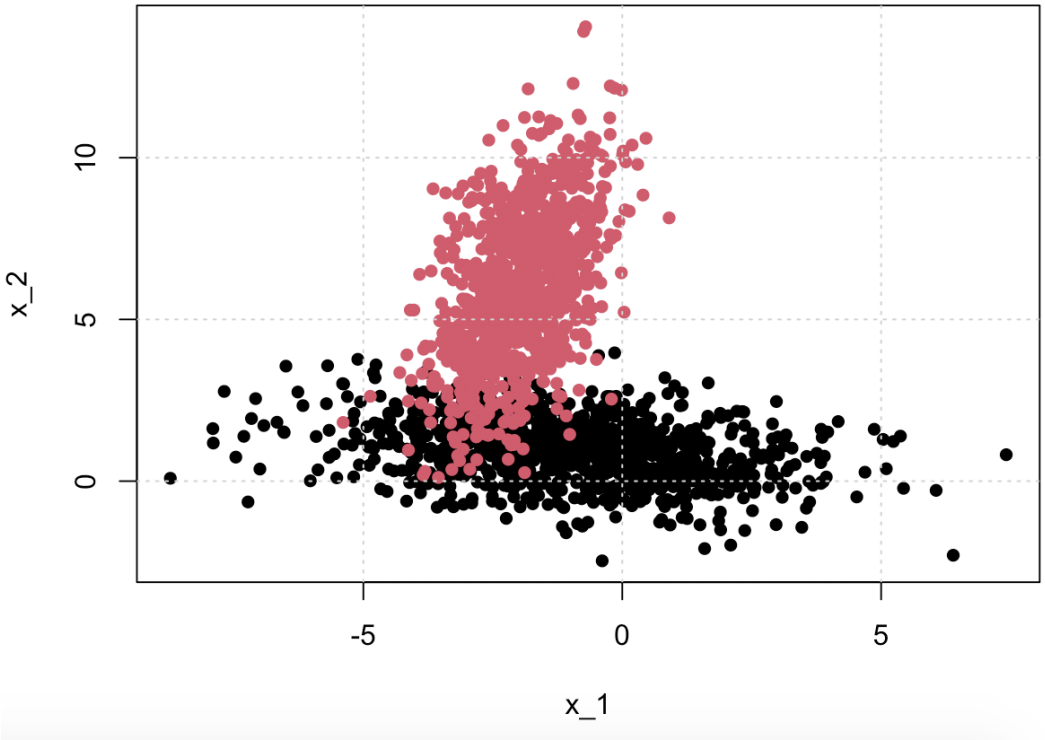
$$K = 2; \mu_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \mu_1 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \Sigma = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$$

Prior probabilities are equal to 0.5.

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

Quadratic discriminant analysis

- Assuming the same covariance structures for all classes is probably unrealistic.



- Can we relax this assumption?

Quadratic discriminant analysis

- We can:

$$x|y = k \sim \mathcal{N}(\mu_k, \Sigma_k)$$

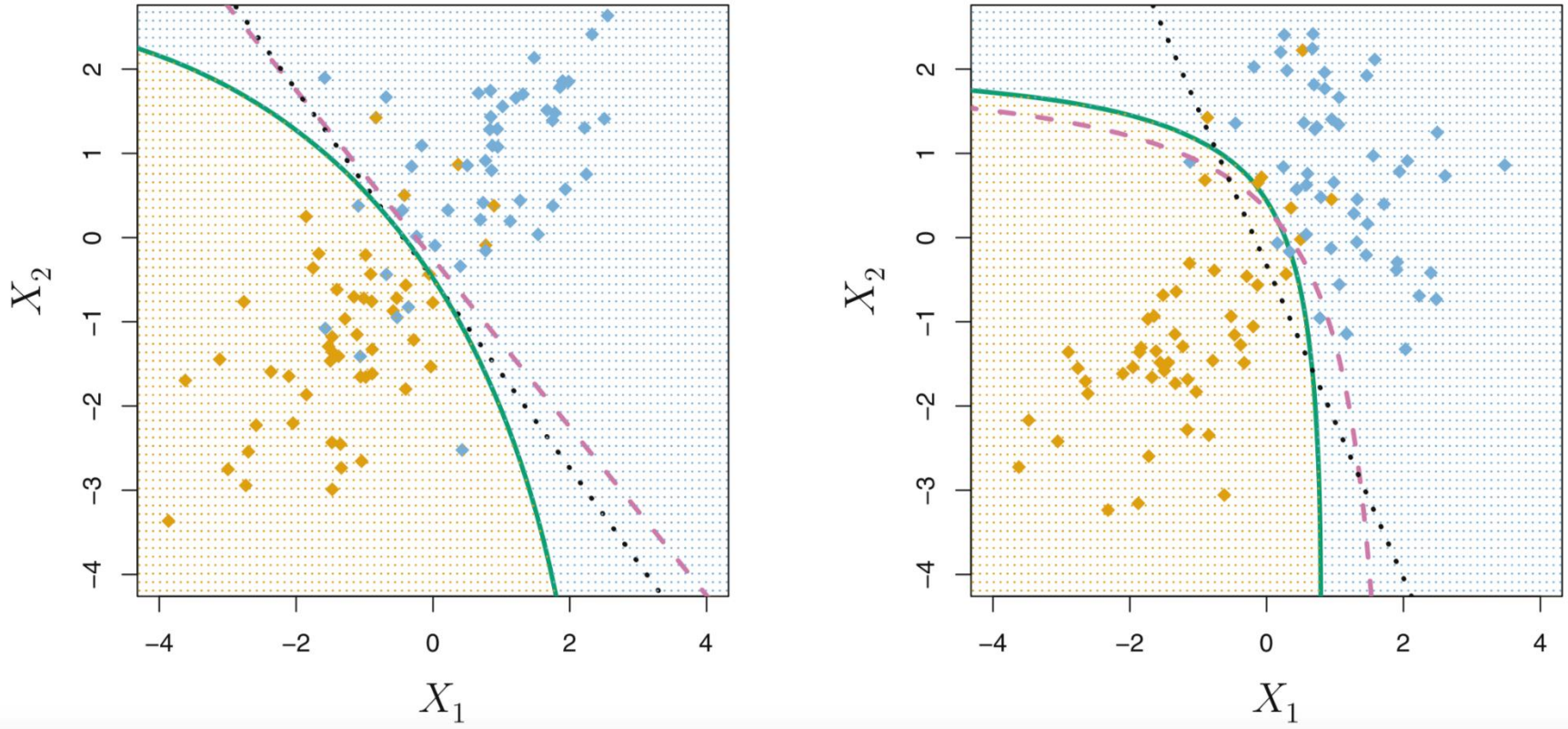
- Then, the classifier assigns an observation x to the class for which

$$\delta_k(x) = -\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k) - \frac{1}{2}\ln|\Sigma_k| + \ln \pi_k$$

is largest.

- QDA plugs in the estimates for μ_k, Σ_k and π_k. Unlike in LDA, x appears as a quadratic function and therefore the decision boundary is not linear.
- Number of unknown parameters $K \cdot m + K\, m(m+1)/2$. If $m = 100$ and $K = 2$, we have 10300 unknown parameters.
- QDA is recommended if training sample is very large. Because for small samples it has high variance.

Quadratic discriminant analysis



- **Left**: The Bayes (purple dashed), LDA (black dotted), and QDA (green solid) decision boundaries for a two-class problem with $\Sigma_1 = \Sigma_2$. The shading indicates the QDA decision rule. Since the Bayes decision boundary is linear, it is more accurately approximated by LDA than by QDA.
- **Right**: Details are as given in the left-hand panel, except that $\Sigma_1 \neq \Sigma_2$. Since the Bayes decision boundary is non-linear, it is more accurately approximated by QDA than by LDA.

Question 5

- Find decision boundary when:

$$K = 2; \mu_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \mu_1 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$$

Prior probabilities are equal to 0.5.

$$\delta_k(x) = -\frac{1}{2}(x - \mu_k)^T \Sigma_{\mathrm{k}}^{-1}(x - \mu_k) - \frac{1}{2}\ln|\Sigma_k| + \ln \pi_k$$

Summary

- LDA and QDA assumes normal distributions, conditioned on the class.
- LDA is less complex because it has a strict assumption, while QDA is more complex (more unknown parameters);
- LDA generates linear decision boundary, while QDA generates quadratic decision boundary.
- Parameters of distributions are obtained by ML method.
- LDA and QDA puts assumptions on generating process of the data, while logistic regression assumes a specific for for probability model.
- Normality is rarely satisfied. However, LDA and QDA are quite robust against the violations of these assumptions.