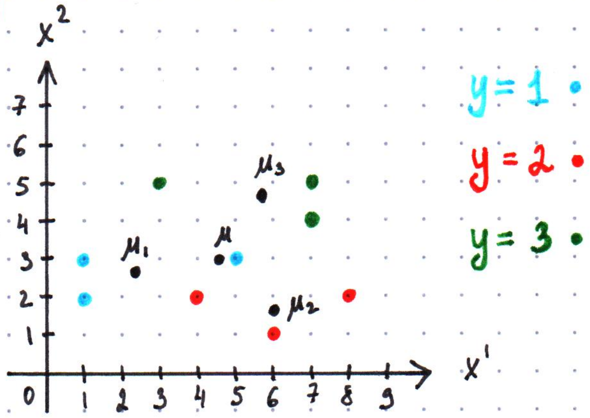
Palina Krukovich in group with Ilya Marozau and Ziad Al Sarih

1. Mean and covariance

$$S = \{((1,2)^T,1),((1,3)^T,1),((5,3)^T,1),((4,2)^T,2),((6,1)^T,2),$$
$$((8,2)^T,2),((3,5)^T,3),((7,5)^T,3),((7,4)^T,3)\}$$

(a)



(b) $\mu = \frac{1}{n}\sum_{i=0}^{n} x_i = \frac{1}{9}\left(\begin{pmatrix}1\\2\end{pmatrix}+\begin{pmatrix}1\\3\end{pmatrix}+\begin{pmatrix}5\\3\end{pmatrix}+\begin{pmatrix}4\\2\end{pmatrix}+\begin{pmatrix}6\\1\end{pmatrix}+\begin{pmatrix}8\\2\end{pmatrix}+\begin{pmatrix}3\\5\end{pmatrix}+\begin{pmatrix}7\\5\end{pmatrix}+\begin{pmatrix}7\\4\end{pmatrix}\right) = \begin{pmatrix}4,67\\3\end{pmatrix}$

$\mu_1 = \frac{1}{3}\left(\begin{pmatrix}1\\2\end{pmatrix}+\begin{pmatrix}1\\3\end{pmatrix}+\begin{pmatrix}5\\3\end{pmatrix}\right) = \begin{pmatrix}2,33\\2,67\end{pmatrix}$

$\mu_2 = \frac{1}{3}\left(\begin{pmatrix}4\\2\end{pmatrix}+\begin{pmatrix}6\\1\end{pmatrix}+\begin{pmatrix}8\\2\end{pmatrix}\right) = \begin{pmatrix}6\\1,67\end{pmatrix}$

$\mu_3 = \frac{1}{3}\left(\begin{pmatrix}3\\5\end{pmatrix}+\begin{pmatrix}7\\5\end{pmatrix}+\begin{pmatrix}7\\4\end{pmatrix}\right) = \begin{pmatrix}5,67\\4,67\end{pmatrix}$

$\Sigma_{11} = \frac{1}{8}\sum_{i=0}^{9}(x_i^1-\mu^1)^2 = \frac{1}{8}\big((1-4,67)^2+(1-4,67)^2+(5-4,67)^2+(4-4,67)^2+$
$+(6-4,67)^2+(8-4,67)^2+(3-4,67)^2+(7-4,67)^2+(7-4,67)^2\big) =$
$= 6,75$

$\Sigma_{12} = \Sigma_{21} = \frac{1}{8}\sum_{i=0}^{9}(x_i^1-\mu^1)(x_i^2-\mu^2) = \frac{1}{8}\big((1-4,67)(2-3)+(1-4,67)(3-3)+$
$+(5-4,67)(3-3)+(4-4,67)(2-3)+(6-4,67)(1-3)+(8-4,67)(2-3)+$
$+(3-4,67)(5-3)+(7-4,67)(5-3)+(7-4,67)(4-3)\big) = 0,25$

$\Sigma_{22} = \frac{1}{8}\sum_{i=0}^{9}(x_i^2-\mu^2)^2 = \frac{1}{8}\big((2-3)^2+(3-3)^2+(3-3)^2+(2-3)^2+(1-3)^2+$
$+(2-3)^2+(5-3)^2+(5-3)^2+(4-3)^2\big) = 2$

$$\Sigma = \begin{pmatrix}6,75 & 0,25\\ 0,25 & 2\end{pmatrix}$$

2. Constructing multi-class classifiers

$\mu_1 = (0,0)^T \quad \mu_2 = (0,6)^T \quad \mu_3 = (6,0)^T \qquad \Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{pmatrix} 1 & 0{,}5 \\ 0{,}5 & 1 \end{pmatrix}$

2.1. One versus one classifier

(a) Yes. The given dataset fulfills the assumptions of LDA, if we were to construct a binary classifier between every pair of classes.

Distribution of features $p(\underline{x}|y) \sim N(\mu_y, \Sigma)$ is Normal Gaussian distribution:

- Covariance matrixes are identical for all classes: $\Sigma_1 = \Sigma_2 = \Sigma_3 = \Sigma$
- μ_y varies for each class

(b) Calculating inverse covariance matrix

$$\Sigma^{-1} = \frac{1}{\Delta} A^T$$ where A is a matrix of cofactors, Δ is a determinant of Σ

$$\Sigma^{-1} = \frac{1}{\Delta} \begin{pmatrix} A_{11} & A_{21} \\ A_{12} & A_{22} \end{pmatrix}$$

$\Delta = 1 \cdot 1 - 0{,}5 \cdot 0{,}5 = 0{,}75$

$\Delta \neq 0 \rightarrow \Sigma^{-1}$ can be found

$A_{11} = (-1)^{1+1} \cdot 1 = 1 \qquad A_{21} = (-1)^{2+1} \cdot 0{,}5 = -0{,}5$

$A_{12} = (-1)^{1+2} \cdot 0{,}5 = -0{,}5 \qquad A_{22} = (-1)^{2+2} \cdot 1 = 1$

$$\Sigma^{-1} = \frac{1}{0{,}75} \begin{pmatrix} 1 & -0{,}5 \\ -0{,}5 & 1 \end{pmatrix} = \begin{pmatrix} 4/3 & -2/3 \\ -2/3 & 4/3 \end{pmatrix}$$

Class 1 – Class 2

$$\underline{w}_{12}^T = (\underline{\mu}_2 - \underline{\mu}_1)^T \Sigma^{-1} = (0,6) \cdot \begin{pmatrix} 4/3 & -2/3 \\ -2/3 & 4/3 \end{pmatrix} = (-4, 8)$$

$$b_{12} = \frac{1}{2} \left(\underline{\mu}_1^T \Sigma^{-1} \underline{\mu}_1 - \underline{\mu}_2^T \Sigma^{-1} \underline{\mu}_2 \right) = -\frac{1}{2} (0,6) \begin{pmatrix} 4/3 & -2/3 \\ -2/3 & 4/3 \end{pmatrix} \begin{pmatrix} 0 \\ 6 \end{pmatrix} = -24$$

$$c_{12} = \log \frac{P(y=1)}{P(y=2)} = \log \frac{0{,}5}{0{,}5} = 0$$

Class 1 – Class 3

$$\underline{w}_{13}^T = (\underline{\mu}_3 - \underline{\mu}_1)^T \Sigma^{-1} = (6,0) \cdot \begin{pmatrix} 4/3 & -2/3 \\ -2/3 & 4/3 \end{pmatrix} = (8, -4)$$

$$b_{13} = \frac{1}{2} \left(\underline{\mu}_1^T \Sigma^{-1} \underline{\mu}_1 - \underline{\mu}_3^T \Sigma^{-1} \underline{\mu}_3 \right) = -\frac{1}{2} (6,0) \begin{pmatrix} 4/3 & -2/3 \\ -2/3 & 4/3 \end{pmatrix} \begin{pmatrix} 6 \\ 0 \end{pmatrix} = -24$$

$c_{13} = 0$

Class 2 - Class 3

$$\underline{w}_{23}^T = (\underline{\mu}_3 - \underline{\mu}_2)^T \Sigma^{-1} = (6,-6) \cdot \begin{pmatrix} 4/3 & -2/3 \\ -2/3 & 4/3 \end{pmatrix} = (12,-12)$$

$$b_{23} = \frac{1}{2}\left(\underline{\mu}_2^T \Sigma^{-1} \underline{\mu}_2 - \underline{\mu}_3^T \Sigma^{-1} \underline{\mu}_3 \right) = (\text{from } b_{12} \text{ and } b_{13} \text{ calc.}) = \frac{1}{2}(48-48) = 0$$

$$c_{23} = 0$$

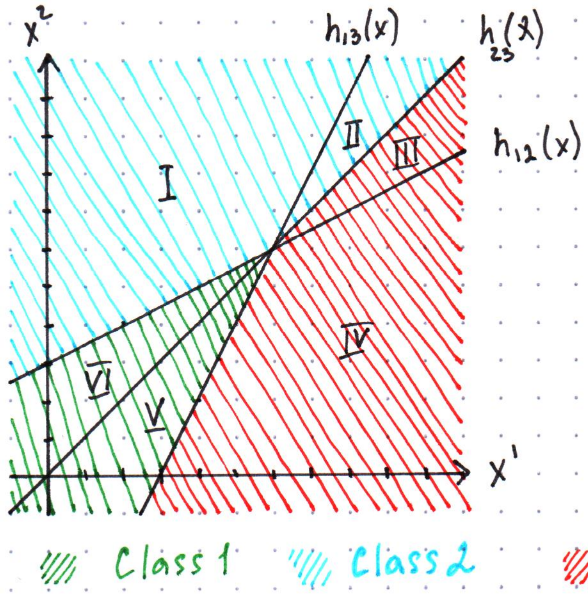
Binary classifiers

$$h_{12}(x) = 1_{\{(-4,8)x - 24 < 0\}} + 2_{\{(-4,8)x - 24 \geq 0\}}$$

$$h_{13}(x) = 1_{\{(8,-4)x - 24 < 0\}} + 3_{\{(8,-4)x - 24 \geq 0\}}$$

$$h_{23}(x) = 2_{\{(12,-12)x < 0\}} + 3_{\{(12,-12)x \geq 0\}}$$

(c)



	I	II	III	IV	V	VI
h_{12}	2	2	2	1	1	1
h_{23}	2	2	3	3	3	2
h_{13}	1	3	3	3	1	1
h_{ovo}	2	2	3	3	1	1

(d)

$x_1 = (10,0)^T$ Values are calculated according to the formulas from 2.1(b) and the plot from 2.1(c)

$(-4,8)\binom{10}{0} - 24 = -64 < 0 \Rightarrow h_{12}(x_1) = 1$

$(8,-4)\binom{10}{0} - 24 = 56 > 0 \Rightarrow h_{13}(x_1) = 3$

$(12,-12)\binom{10}{0} = 120 > 0 \Rightarrow h_{23}(x_1) = 3$

$h_{ovo}(x_1) = 3$

$x_2 = (-2,6)^T$ $h_{12}(x_2) = 2$ $h_{13}(x_2) = 1$ $h_{23}(x_2) = 2$ $h_{ovo}(x_2) = 2$

$x_3 = (-2,-6)^T$ $h_{12}(x_3) = 1$ $h_{13}(x_3) = 1$ $h_{ovo}(x_3) = 1$

$x_4 = (9,8)^T$ $h_{12}(x_4) = 2$ $h_{13}(x_4) = 3$ $h_{23}(x_4) = 3$ $h_{ovo}(x_4) = 3$

$x_5 = (9,11)^T$ $h_{12}(x_5) = 2$ $h_{13}(x_5) = 3$ $h_{23}(x_5) = 2$ $h_{ovo}(x_5) = 2$

2.2. One versus all classifier (OVA)

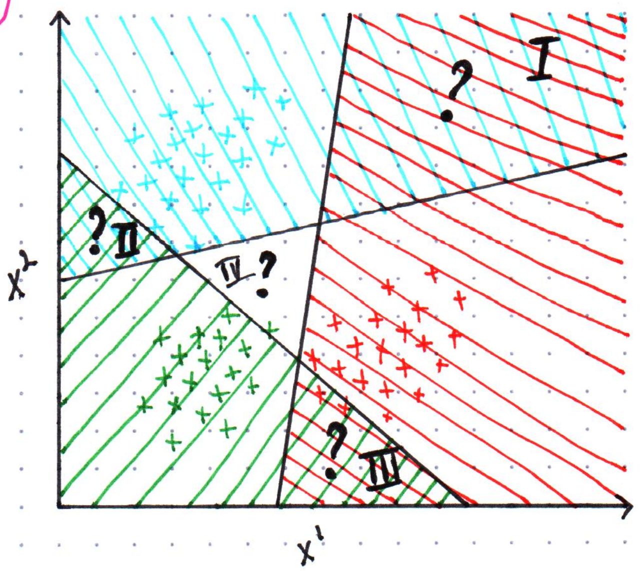
(a) No. The dataset does not fulfill the assumptions of LDA, if we were to construct a binary classifier between one class and all the remaining classes.

In LDA we assume that the distribution of features is normal and covariance matrixes for all classes are identical.

But in this dataset we'll get different covariance matrixes for one class and all the remaining classes.
The collection of classes (12, 23, 13) is not normally distributed.

$$\Sigma_1 \neq \Sigma_{23} \quad \Sigma_2 \neq \Sigma_{13} \quad \Sigma_3 \neq \Sigma_{12}$$

(b)



Class 1 ×
Class 2 ×
Class 3 ×

If we just look at which side of the hyperplane x is, we get the whole uncertainty regions (I, II, III, IV).

For regions I, II, III - equal probability for two classes

Region IV - equal probability of all three classes.

(c) Let $p(y=j \mid x) \Rightarrow p(y=1 \mid x)$ - probability of class j given x
$p(y=\bar{j} \mid x) \Rightarrow p(y=0 \mid x)$ - probability of any other class given x

$$p(y=1 \mid x) = \frac{p(x \mid y=1)\, p(y=1)}{p(x)} = \frac{p(x \mid y=1)\, p(y=1)}{p(x \mid y=1)\, p(y=1) + p(x \mid y=0)\, p(y=0)} =$$

$$= \frac{1}{1 + \underbrace{\frac{p(x \mid y=0)\, p(y=0)}{p(x \mid y=1)\, p(y=1)}}}$$

$$= \frac{\cancel{\frac{1}{\sqrt{2\pi}\cdot|\Sigma|^{1/2}}} \cdot e^{-\frac{1}{2}(x-\mu_0)^T \Sigma^{-1}(x-\mu_0)}}{\cancel{\frac{1}{\sqrt{2\pi}\cdot|\Sigma|^{1/2}}} \cdot e^{-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1)}} \cdot \frac{p(y=0)}{p(y=1)} =$$

$$= e^{-\frac{1}{2}(x-\mu_0)^T \Sigma^{-1}(x-\mu_0) + \frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1)} \cdot e^{\log \frac{p(y=0)}{p(y=1)}} =$$

BRUNNEN

$$= e^{-\left(\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0) - \frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - \log\frac{P(y=0)}{P(y=1)}\right)} =$$

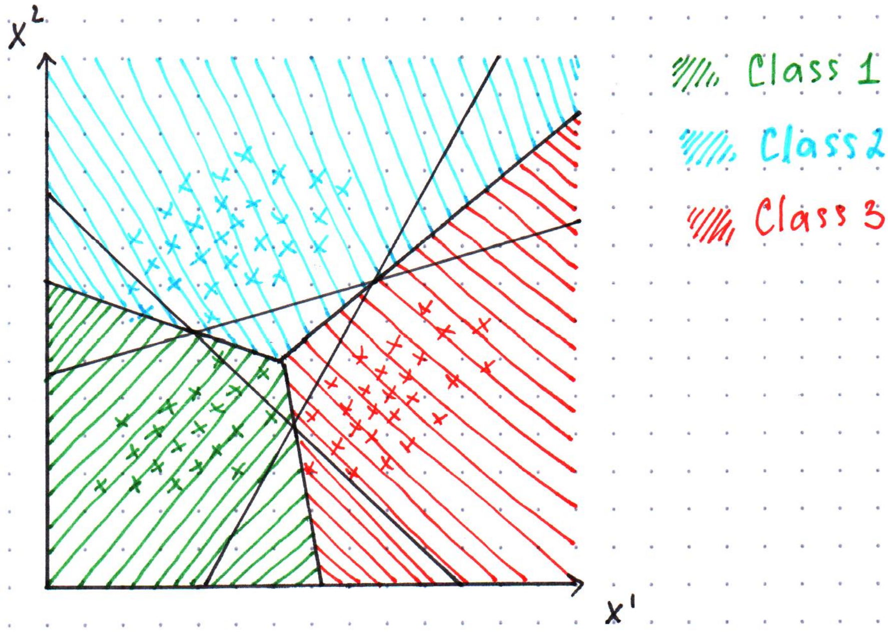
$$= e^{-\left(\underbrace{(\mu_1-\mu_0)^T\Sigma^{-1}}_{w^T} x + \underbrace{\frac{1}{2}\left(\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1\right)}_{b} - \underbrace{\log\frac{P(y=0)}{P(y=1)}}_{c}\right)}$$

$$= e^{-(w^T x + b - c)}$$

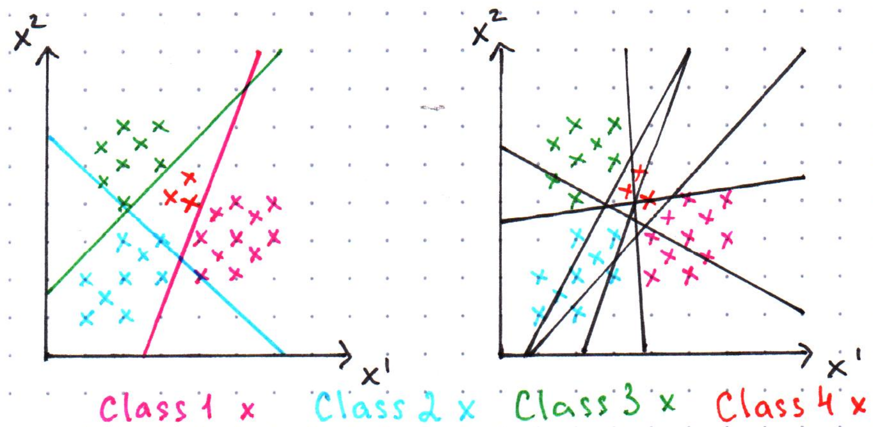
$$P(y=1 \mid x) = \frac{1}{1+e^{-(w^T x + b - c)}} = \sigma(w^T x + b - c)$$

$$P(y=j \mid x=x_i) = \sigma(w_j^T x_i + b_j - c_j)$$

(d)



(e)



Class 4 can't be separated in OVA.

OVO classifier can handle such situation because it constructs classifiers in pairs And all pairs are separable.