Section 7 Gaussian Discriminant Analysis

1. Gaussian Discriminant Analysis
2. Review Gaussian distribution
3. Linear Discriminant Analysis (LDA)
4. LDA v.s. Logistics regression

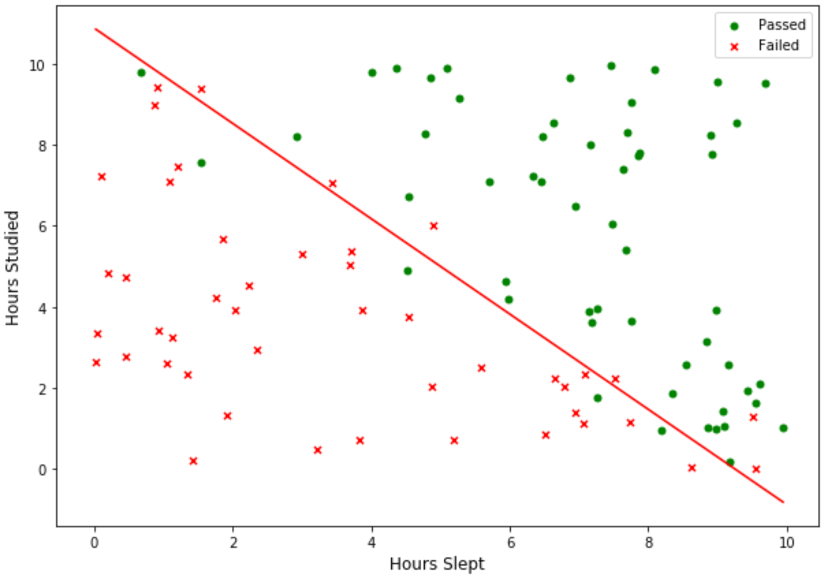
Question:

Classification Data:

$$D = \{(\vec{x}^{(i)}, y^{(i)}), i = 1, \ldots n\} \qquad y^{(i)} \in \{1,2, \ldots, K\},$$

Goal: Find conditional (posterior) probability

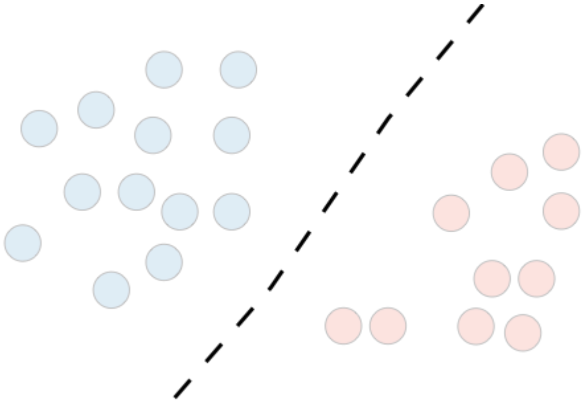
$$P(Y = k \,|\vec{X} = \vec{x}) \qquad \text{for } k = 1, 2, \ldots, K$$



➢ Discriminative learning algorithms.

- Based on a model of the conditional probability.

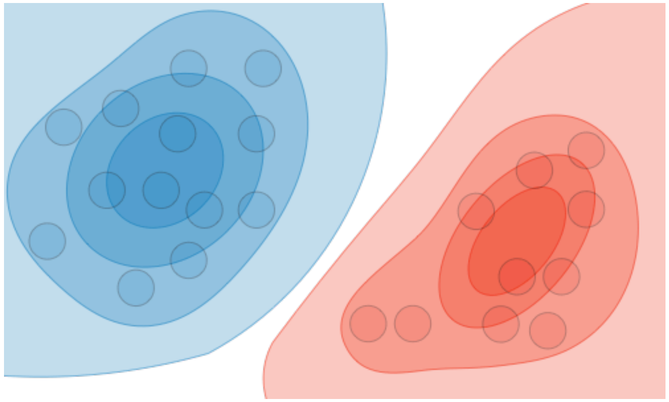
$$P(Y = k \,|\vec{X} = \vec{x}) := some\ model$$



Examples: logistics/softmax, SVM,

➢ Generative learning algorithms.

- Based on models of the distributions of the dataset:
    - Prior probability $P(Y)$
    - Likelihood probability $P(X|Y = k)$



Examples: Gauss discriminant analysis (GDA: LDA/QDA), Navis Bayes,

➢ Gauss discriminant analysis

By **Bayes Rule:**

$$P(Y = k \mid \vec{X} = \vec{x}) = \frac{P(\vec{X} = \vec{x} \mid Y = k)P(Y = k)}{\sum_{all\ i} P(\vec{X} = \vec{x} \mid Y = i)P(Y = i)}$$

$$= \frac{P(\vec{X} = \vec{x} \mid Y = k)P(Y = k)}{P(\vec{X} = \vec{x})}$$

$$Posterior = \frac{Likelihood \ \times Prior}{Evidence}$$

$$\underset{y}{\operatorname{argmax}} P(Y = k \mid \vec{X} = \vec{x}) = \underset{y}{\operatorname{argmax}} P(\vec{X} = \vec{x} \mid Y = k)P(Y = k)$$

Gaussian Discriminant Analysis (GDA) assumptions

- Prior probability $P(Y)$

Assume $Y \backsim \mathrm{Bernouli}\ (\phi)$ or $\mathrm{Categorical}(\phi_1, \dots, \phi_K)$

- Likelihood probability $P(X|Y = k)$

Assume $X|Y = k$ is a normal distribution for each k.

- **Binary** Classification $y \in \{0,1\}$

Assume $Y \backsim \mathrm{Bernouli}\ (\phi)$

Pdf function $p(y) = \phi^{y}(1-\phi)^{1-y} = \begin{cases} \phi & if\ y=1 \\ 1-\phi & if\ y=0 \end{cases}$

- **Multiclass** Classification $y \in \{1,2,\dots,K\}$

Assume $Y \backsim \mathrm{Categorical}(\phi_1, \dots, \phi_K)$ such that $\phi_1 + \cdots + \phi_K = 1$

Pdf function $p(y) = \phi_1^{\mathbb{I}(\mathrm{y}=1)} \phi_2^{\mathbb{I}(\mathrm{y}=2)} \dots \phi_K^{\mathbb{I}(\mathrm{y}=\mathrm{K})}$

➢ **Normal (Gaussian) distribution (single variable).**

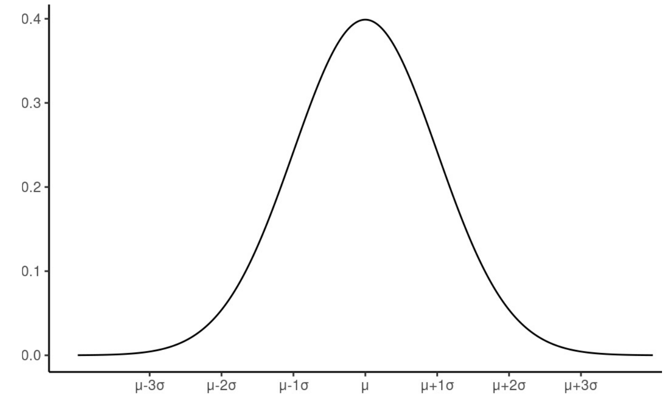
Random variable $X \sim \text{Normal}(\mu, \sigma^2)$

- The probability density function (**pdf**) for X is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad -\infty < x < \infty$$

- The **mean** of X is $E(X) = \mu$
- The **variance** of X is $Var(X) = \sigma^2$
- Probability

$$P(a < X < b) = \int_a^b f_X(x)dx$$



➢ **Multivariate normal distribution.**

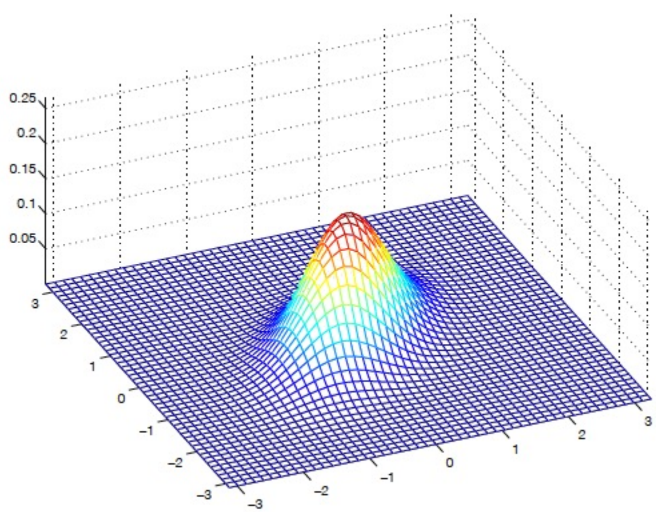
Vector random variable $\vec{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_d \end{bmatrix} \sim \mathrm{Normal}(\vec{\mu}, \Sigma)$

Here $\vec{\mu} \in \mathbb{R}^d$ and Σ is an $d \times d$ symmetric, positive definite matrix.

- The joint probability density function (**pdf**) for $\vec{X}$ is

$$f_{\vec{X}}(\vec{x}) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})\right)$$

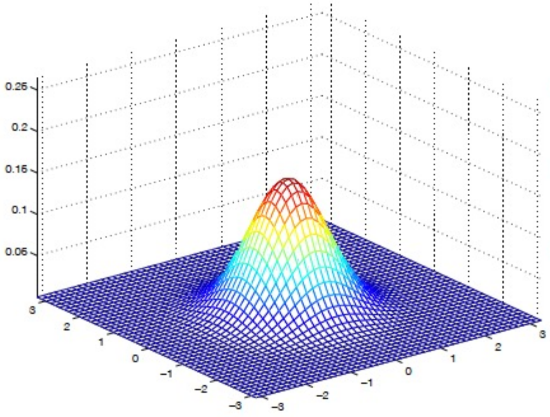
$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_{\vec{X}}(\vec{x}) \; d\vec{x} = 1$$



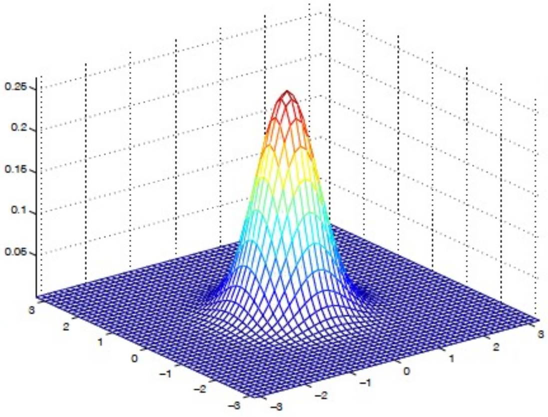
- The **mean vector** of $\vec{X}$ is $E(\vec{X}) = \vec{\mu}$

- The **(co)variance matrix** is Cov($\vec{X}$)= Σ

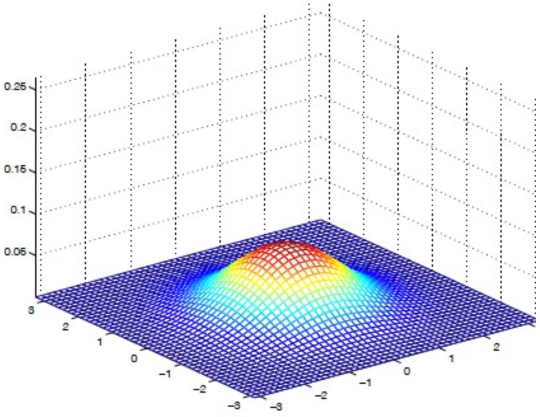
Standard normal

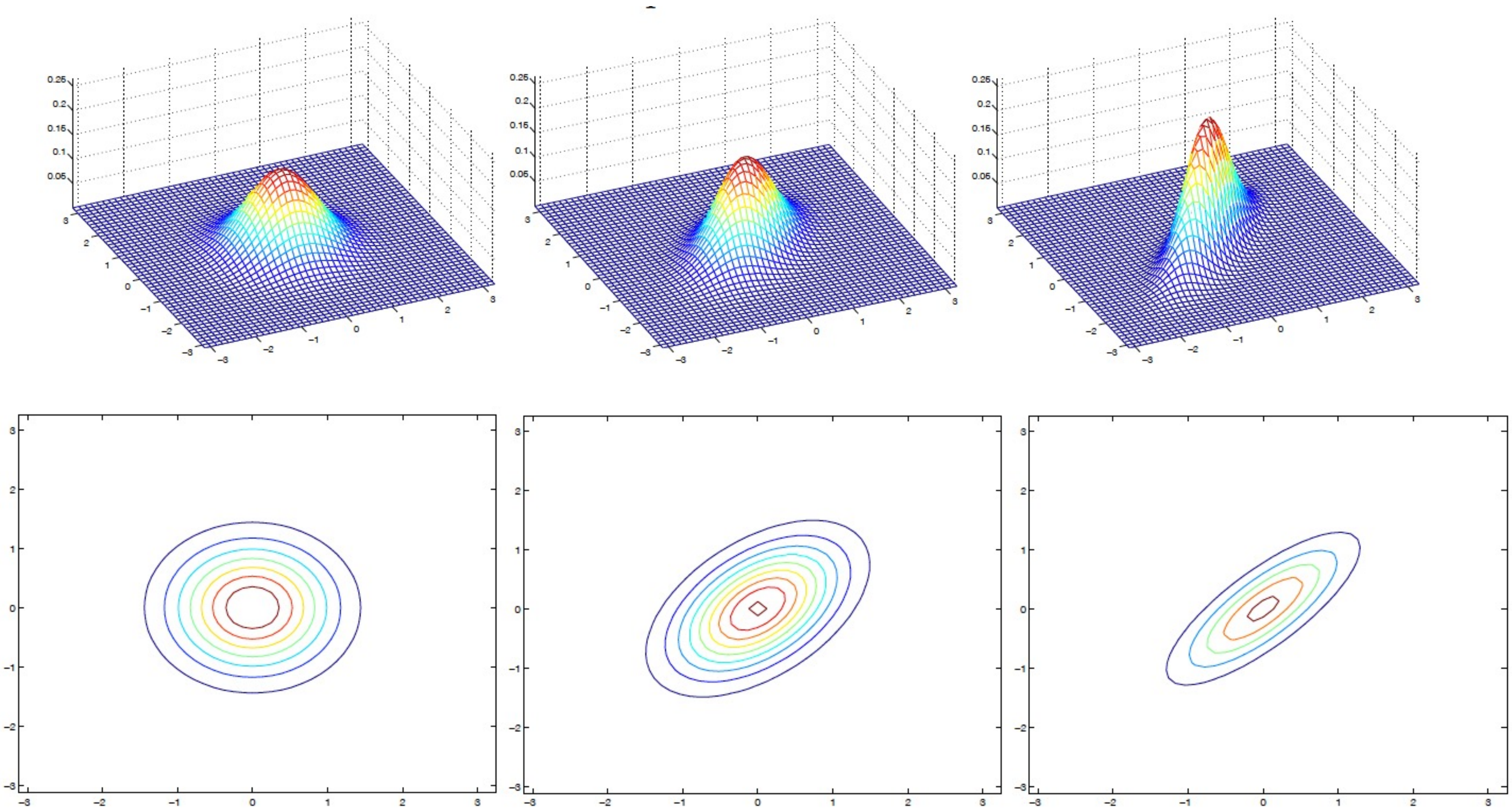


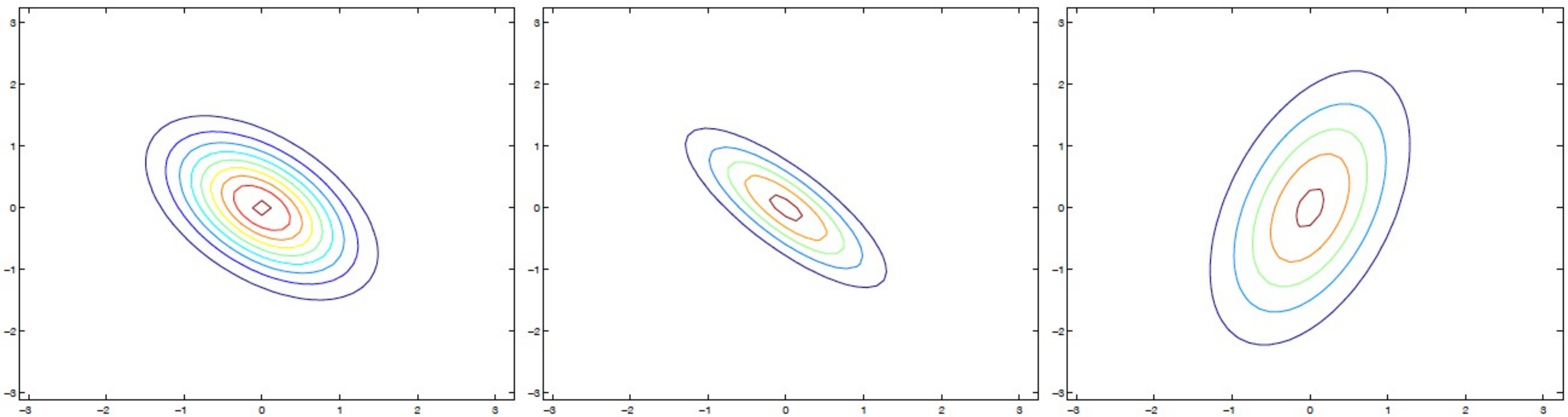
Compressed

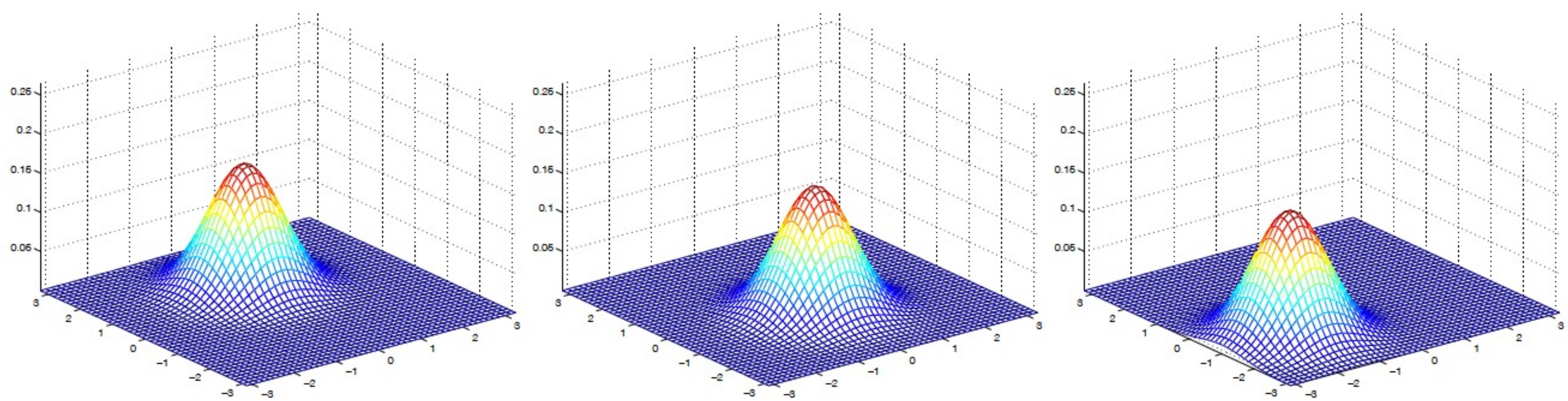


Spread-out









See more the probability review or Chapter 2“Pattern Recognition and Machine Learning”-Chris Bishop

- Gaussian (**Linear)** Discriminant Analysis (LDA).

Binary Classification Data: $D = \{(\vec{x}^{(i)}, y^{(i)}), i = 1, \ldots n\}$ $y^{(i)} \in \{0,1\}$,

Goal: Find conditional (posterior) probability

$$P(Y = k \,|\vec{X} = \vec{x}) \qquad \text{for } k = 0,1$$

GDA Method: We need to find

$$P(\vec{X} = \vec{x} | Y = k)$$

$$P(Y = k)$$

Assume $Y \backsim \text{Bernouli } (\phi)$

Assume $\vec{X} \,|Y = 0 \backsim \text{Normal}(\vec{\mu}_0, \Sigma_0)$

$$\vec{X} \,|Y = 1 \backsim \text{Normal}(\vec{\mu}_1, \Sigma_1)$$

LDA Assume: $\Sigma_0 = \Sigma_1 = \Sigma$

pdf functions:

$$p_Y(y) = \phi^y (1-\phi)^{1-y}$$

$$p(\vec{X}|Y=0) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu}_0)^T \Sigma^{-1} (\vec{x}-\vec{\mu}_0)\right)$$

$$p(\vec{X}|Y=1) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu}_1)^T \Sigma^{-1} (\vec{x}-\vec{\mu}_1)\right)$$

Given data $D = \left\{\left(\vec{x}^{(i)}, y^{(i)}\right), i = 1, \dots n\right\}$, we want to **maximize likelihood**

$$P(\text{data}) = P(\boldsymbol{X}, \vec{\boldsymbol{y}}) = \prod_{i=1}^{n} P\left(\vec{X} = \vec{x}^{(i)}, Y = y^{(i)}\right)$$

Equivalently, we **maximize likelihood**

$$L(\phi, \vec{\mu}_0, \vec{\mu}_1, \Sigma\,) = \prod_{i=1}^{n} p\big(\, \vec{X} = \vec{x}^{(i)} \mid Y = y^{(i)}\big)\, p_Y(y^{(i)})$$

Equivalently, we **maximize log likelihood**

$$l(\phi, \vec{\mu}_0, \vec{\mu}_1, \Sigma\,) = \log L(\phi, \vec{\mu}_0, \vec{\mu}_1, \Sigma\,)$$

$$= \sum_{i=1}^{n} \left(\log p \left(\, \vec{X} = \vec{x}^{(i)} \;\middle|\; Y = y^{(i)} \right) + \log p_Y\big(y^{(i)}\big) \right)$$

Calculate $\nabla\; l(\phi, \vec{\mu}_0, \vec{\mu}_1, \Sigma\,) = 0$ and find critical points. (Practice.)

Maximum Likelihood estimates:

We obtain formulas for the parameters maximizing the likelihood:

$$\phi = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(y^{(i)} = 1)$$

$$\vec{\mu}_0 = \frac{\sum_{i=1}^{n} \mathbb{1}\left(y^{(i)} = 0\right)\vec{x}^{(i)}}{\sum_{i=1}^{n} \mathbb{1}(y^{(i)} = 0)}$$

$$\vec{\mu}_1 = \frac{\sum_{i=1}^{n} \mathbb{1}\left(y^{(i)} = 1\right)\vec{x}^{(i)}}{\sum_{i=1}^{n} \mathbb{1}(y^{(i)} = 1)}$$

$$\Sigma = \frac{1}{n - 2}\sum_{i=1}^{n}\left(\vec{x}^{(i)} - \vec{\mu}_{y^{(i)}}\right)\left(\vec{x}^{(i)} - \vec{\mu}_{y^{(i)}}\right)^T$$

We have the optimal distribution models with pdf functions:

$$p_Y(k) = \phi^k(1-\phi)^{1-k}$$

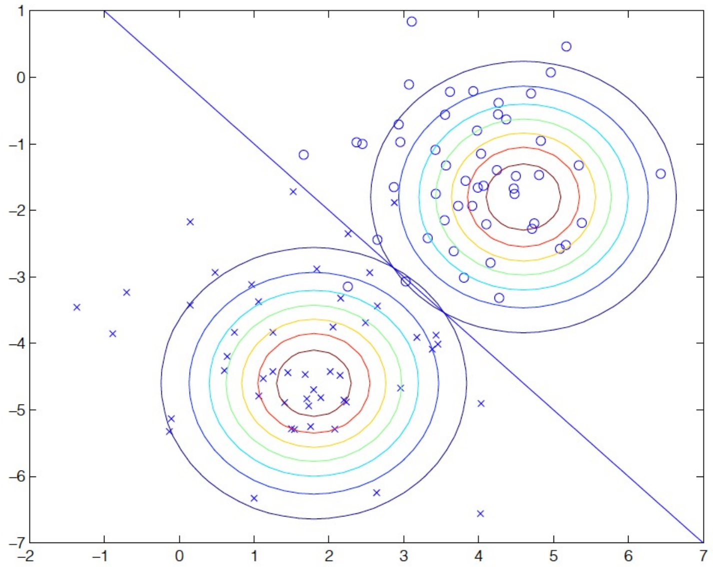
$$p(\vec{X} = \vec{x}|Y = 0) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu}_0)^T\Sigma^{-1}(\vec{x} - \vec{\mu}_0)\right)$$

$$p(\vec{X} = \vec{x}|Y = 1) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu}_1)^T\Sigma^{-1}(\vec{x} - \vec{\mu}_1)\right)$$

We have the formulas for the conditional pdf of Y give$n$ $\vec{X} = \vec{x}$

$$p(Y = k\ |\vec{X} = \vec{x}) = \frac{p\left(\vec{X} = \vec{x}\middle|Y = k\right)p_Y(k)}{\sum_{all\ i} p\left(\vec{X} = \vec{x}\middle|Y = i\right)p_Y(i)}$$

We can find the level curves of the distributions and the **boundary**:



On **boundary** points $\vec{x}$, we have

$$p(Y=0\,|\vec{X}=\vec{x}) = p(Y=1\,|\vec{X}=\vec{x})$$

So,

$$\log\frac{p(Y=0\,|\vec{X}=\vec{x})}{p(Y=1\,|\vec{X}=\vec{x})} = 0$$

$$p(Y = k \mid \vec{X} = \vec{x}) = \frac{p(\vec{X} = \vec{x} \mid Y = k)\, p_Y(k)}{p(\vec{X} = \vec{x})}$$

$$\log p(Y = k \mid \vec{X} = \vec{x}) = \log p(\vec{X} = \vec{x} \mid Y = k) + \log p_Y(k) - \log p(\vec{X} = \vec{x})$$

$$= -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(\vec{x} - \vec{\mu}_k)^T \Sigma_k^{-1} (\vec{x} - \vec{\mu}_k) + \log \phi_k + constant$$

Quadratic discriminant function $\delta_k^Q(\vec{x})$

LDA assumption:
$\Sigma_0 = \Sigma_1 = \Sigma$

$$= \vec{x}^T \Sigma^{-1} \vec{\mu}_k - \frac{1}{2}\vec{\mu}_k^T \Sigma^{-1} \vec{\mu}_k + \log \phi_k + constant$$

Linear discriminant function $\delta_k(\vec{x})$

Boundary formula

$$\log p(Y = 0 \,|\vec{X} = \vec{x}) = \log p(Y = 1 \,|\vec{X} = \vec{x})$$

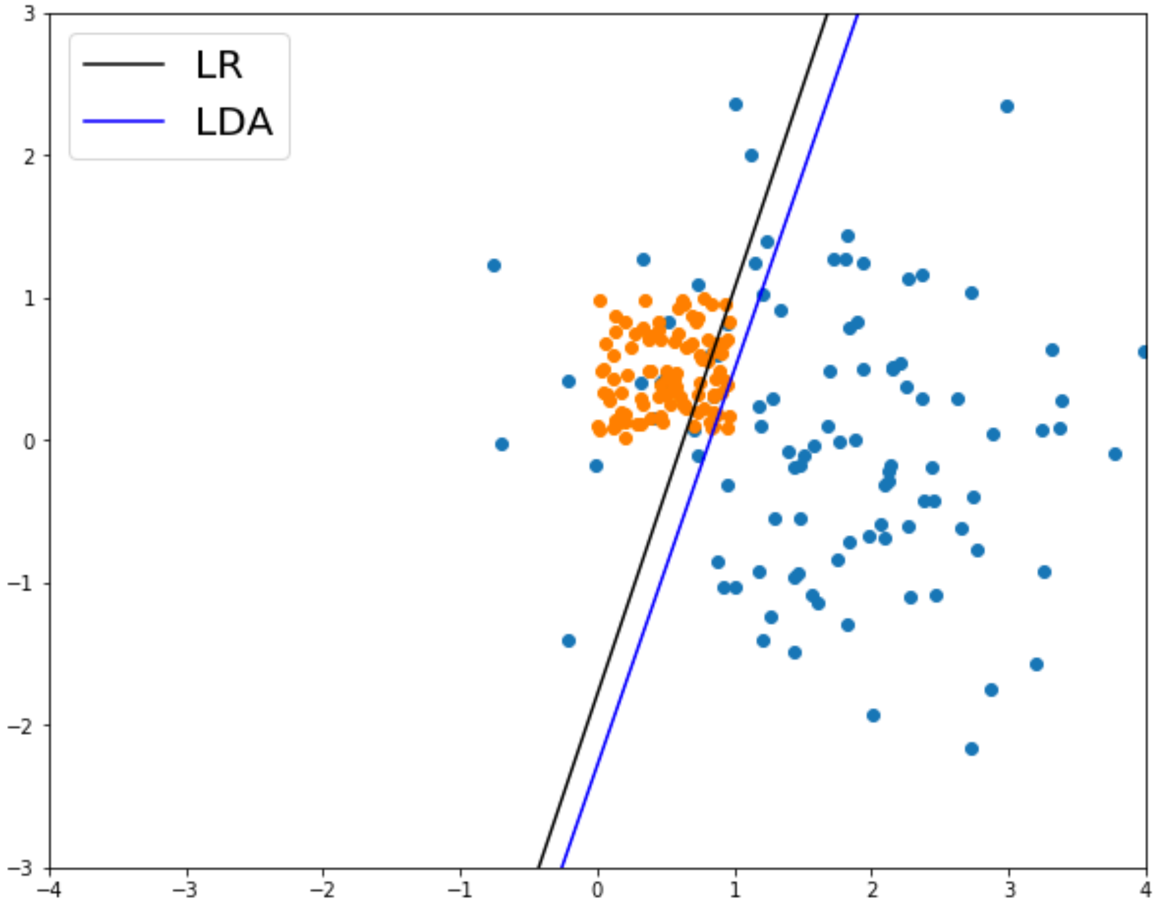
- LDA boundary Equivalent to $\delta_0(\vec{x}) = \delta_1(\vec{x})$

$$\vec{x}^T \Sigma^{-1} \vec{\mu}_0 - \frac{1}{2} \vec{\mu}_0^T \Sigma^{-1} \vec{\mu}_0 + \log \phi_0 = \vec{x}^T \Sigma^{-1} \vec{\mu}_1 - \frac{1}{2} \vec{\mu}_1^T \Sigma^{-1} \vec{\mu}_1 + \log \phi_1$$

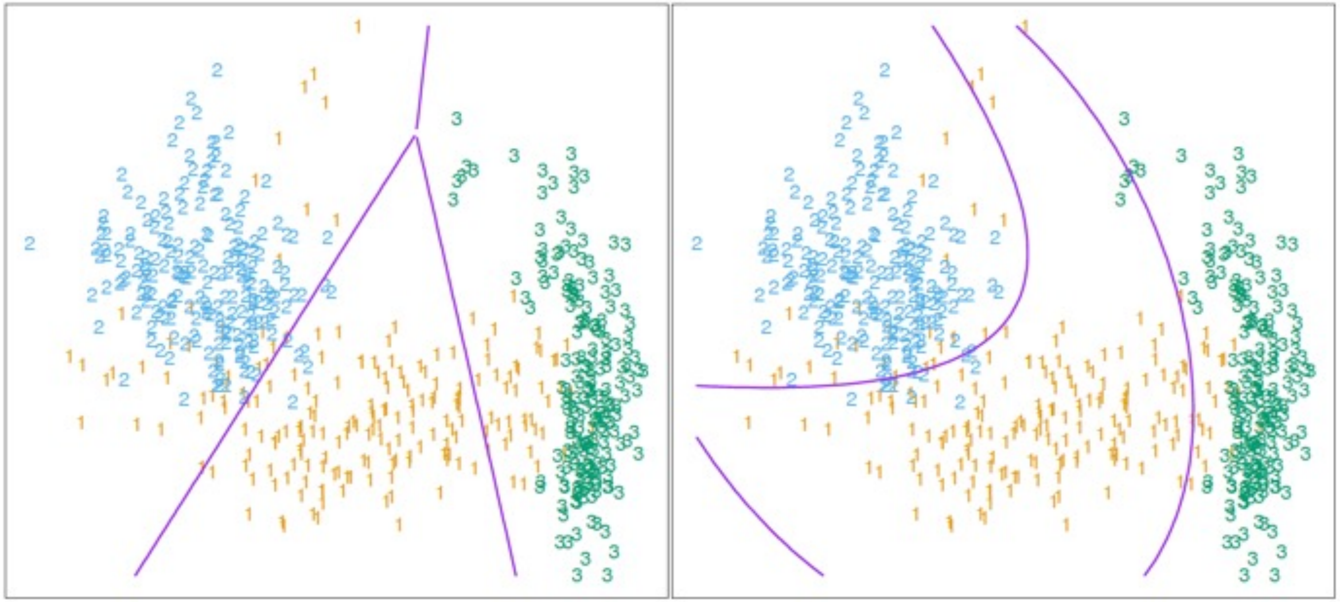
- QDA boundary Equivalent to $\delta_0^Q(\vec{x}) = \delta_1^Q(\vec{x})$

$$-\frac{1}{2}\log|\Sigma_0| - \frac{1}{2}(\vec{x} - \vec{\mu}_0)^T \Sigma_0^{-1}(\vec{x} - \vec{\mu}_0) + \log \phi_0 =$$

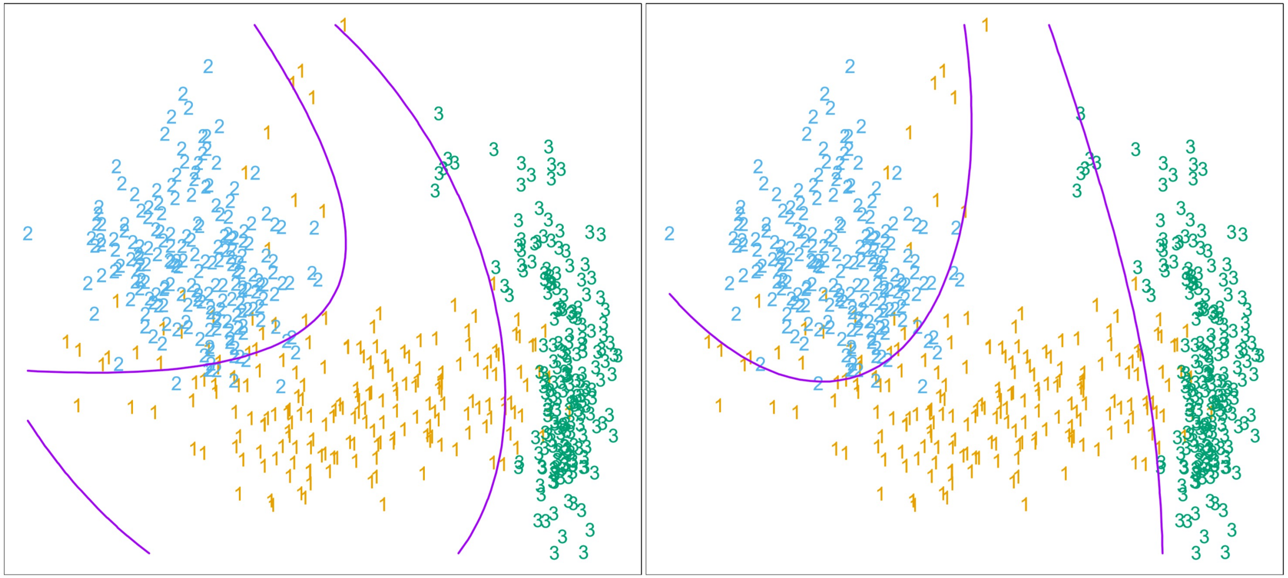
$$-\frac{1}{2}\log|\Sigma_1| - \frac{1}{2}(\vec{x} - \vec{\mu}_1)^T \Sigma_1^{-1}(\vec{x} - \vec{\mu}_1) + \log \phi_1$$


LR
LDA
3
2
1
0
−1
−2
−3
−4
−3
−2
−1
0
1
2
3
4

Example from the book.



- The left plot shows some data from three classes, with linear decision boundaries found by **linear discriminant analysis**.
- The right plot shows quadratic decision boundaries. These were obtained by finding linear boundaries in the five-dimensional space x_1 , $x_2, x_1x_2, x_1^2, x_2^2$. Linear inequalities in this space are quadratic inequalities in the original space.



- Two methods for fitting quadratic boundaries.
- The left plot shows the quadratic decision boundaries for the data (obtained using LDA in the five-dimensional space x_1 , $x_2, x_1 x_2, x_1^2, x_2^2$).
- The right plot shows the quadratic decision boundaries found by **QDA**. The differences are small, as is usually the case.

Compare to Logistic regression

- When these modeling assumptions are correct, then GDA will be better fits to the data.
- GDA will be a better algorithm than logistic regression for **small** training set sizes.
- logistic regression makes significantly **weaker** assumptions. So, it is more robust and less sensitive to incorrect modeling assumptions.
- GDA has closed formulas for the optimal points. Logistic regression need to use gradient descent or Newton’s methods.

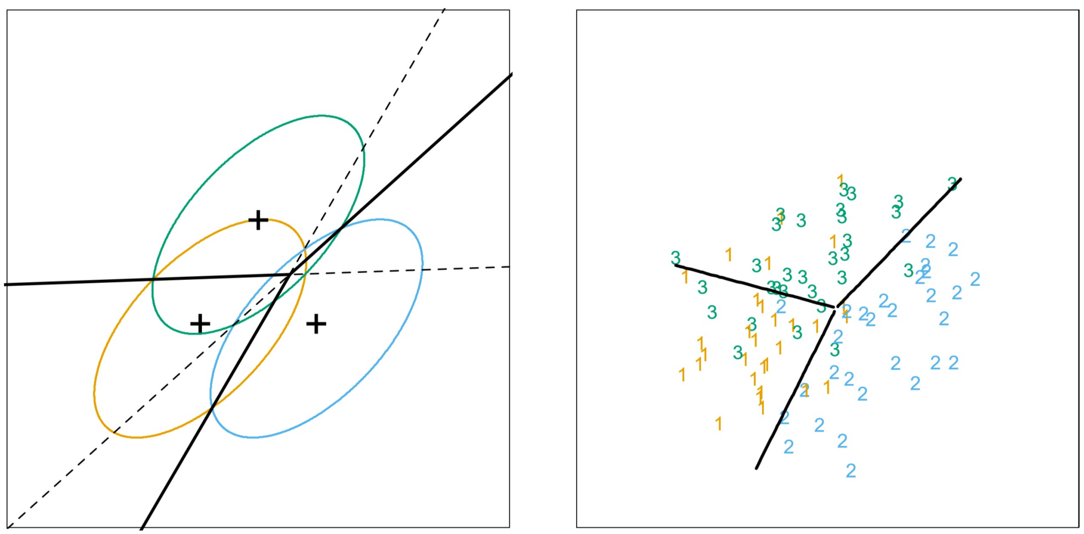
Example:

- **Multiclass Classification**

$y \in \{1,2,\dots,K\}$

Assume $Y \backsim \text{Categorical}(\phi_1, \dots, \phi_K)$ such that $\phi_1 + \cdots + \phi_K = 1$

Pdf function $p(y) = \phi_1^{\mathbb{I}(y=1)} \phi_2^{\mathbb{I}(y=2)} \dots \phi_K^{\mathbb{I}(y=K)}$



Maximum Likelihood estimates:

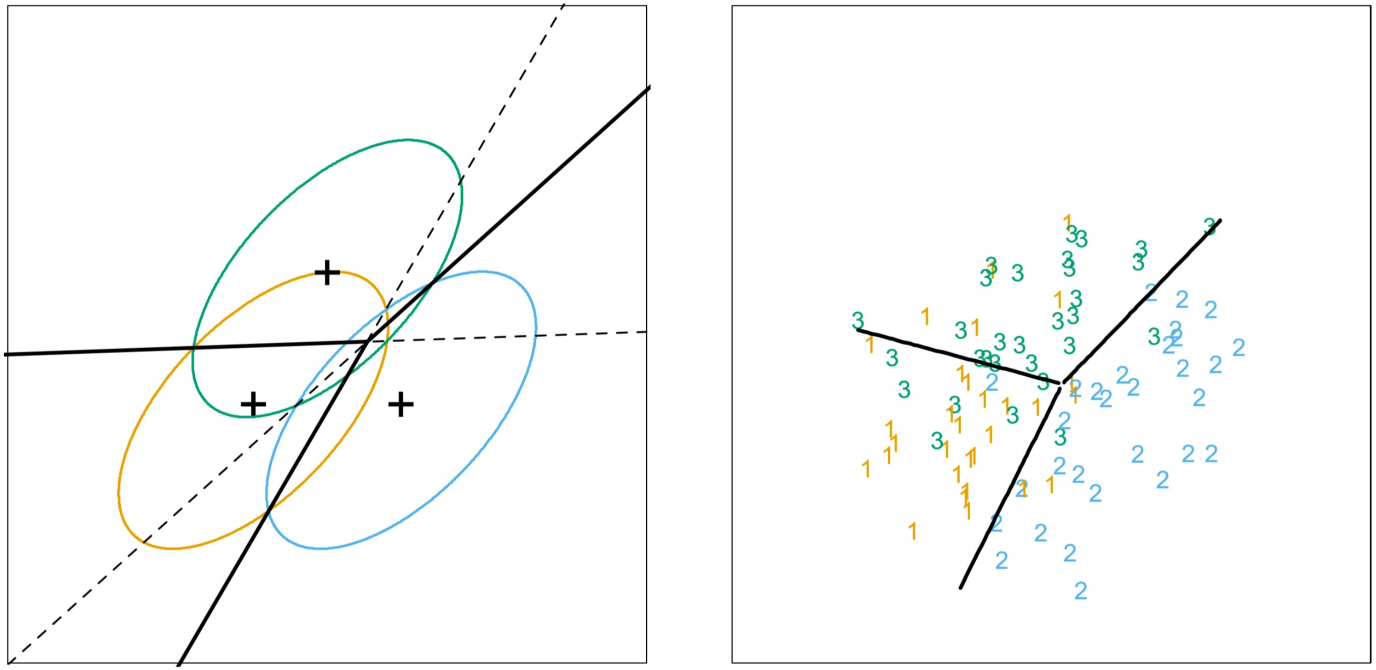
We obtain formulas for the parameters maximizing the likelihood:

$$\phi_j = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(y^{(i)} = j)$$

$$\vec{\mu}_{\mathrm{j}} = \frac{\sum_{i=1}^{n} \mathbb{1}\left(y^{(i)} = j\right)\vec{x}^{(i)}}{\sum_{i=1}^{n} \mathbb{1}(y^{(i)} = j)}$$

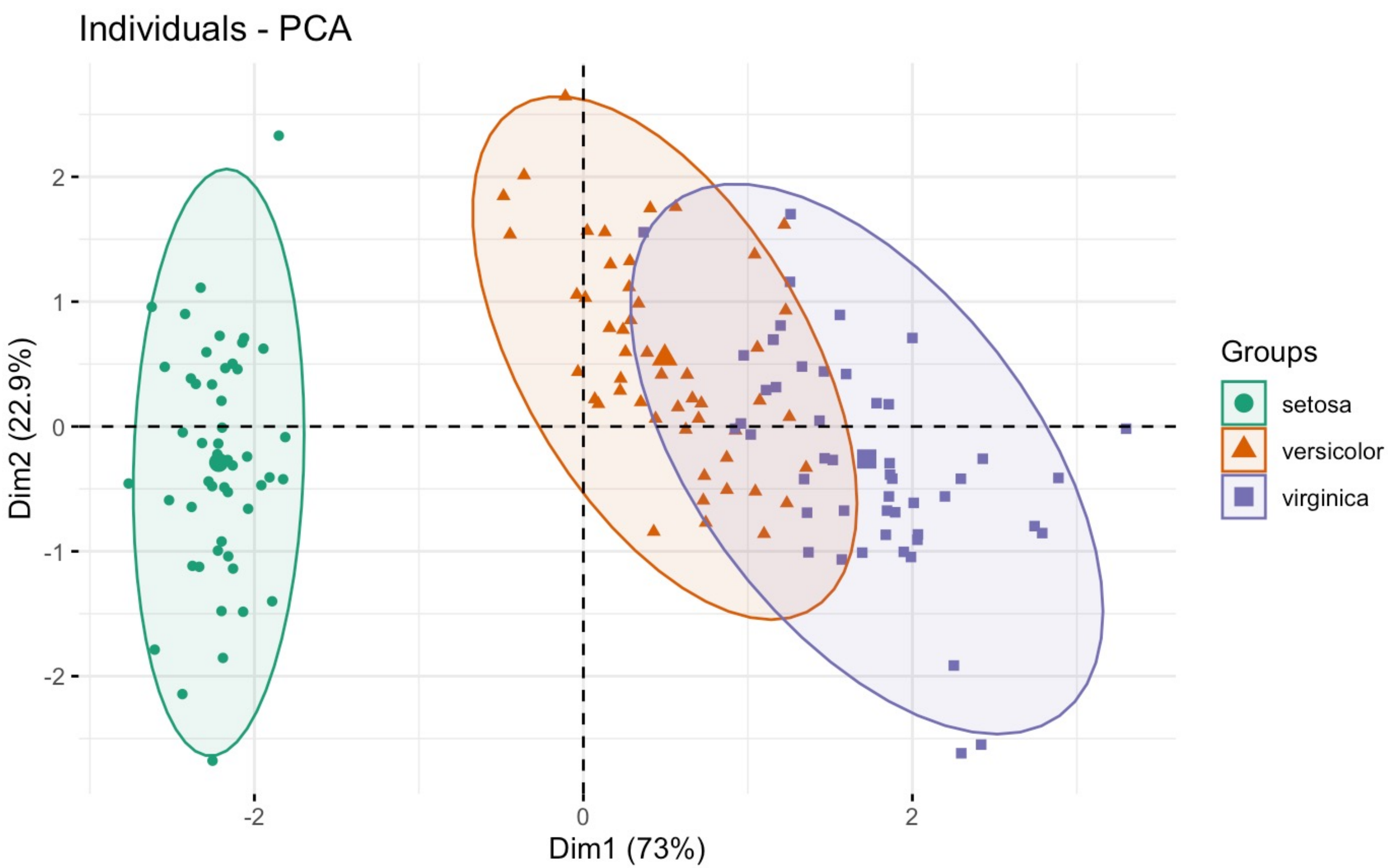
$$\Sigma = \frac{1}{n - K} \sum_{i=1}^{n} \left(\vec{x}^{(i)} - \vec{\mu}_{y^{(i)}}\right)\left(\vec{x}^{(i)} - \vec{\mu}_{y^{(i)}}\right)^T$$

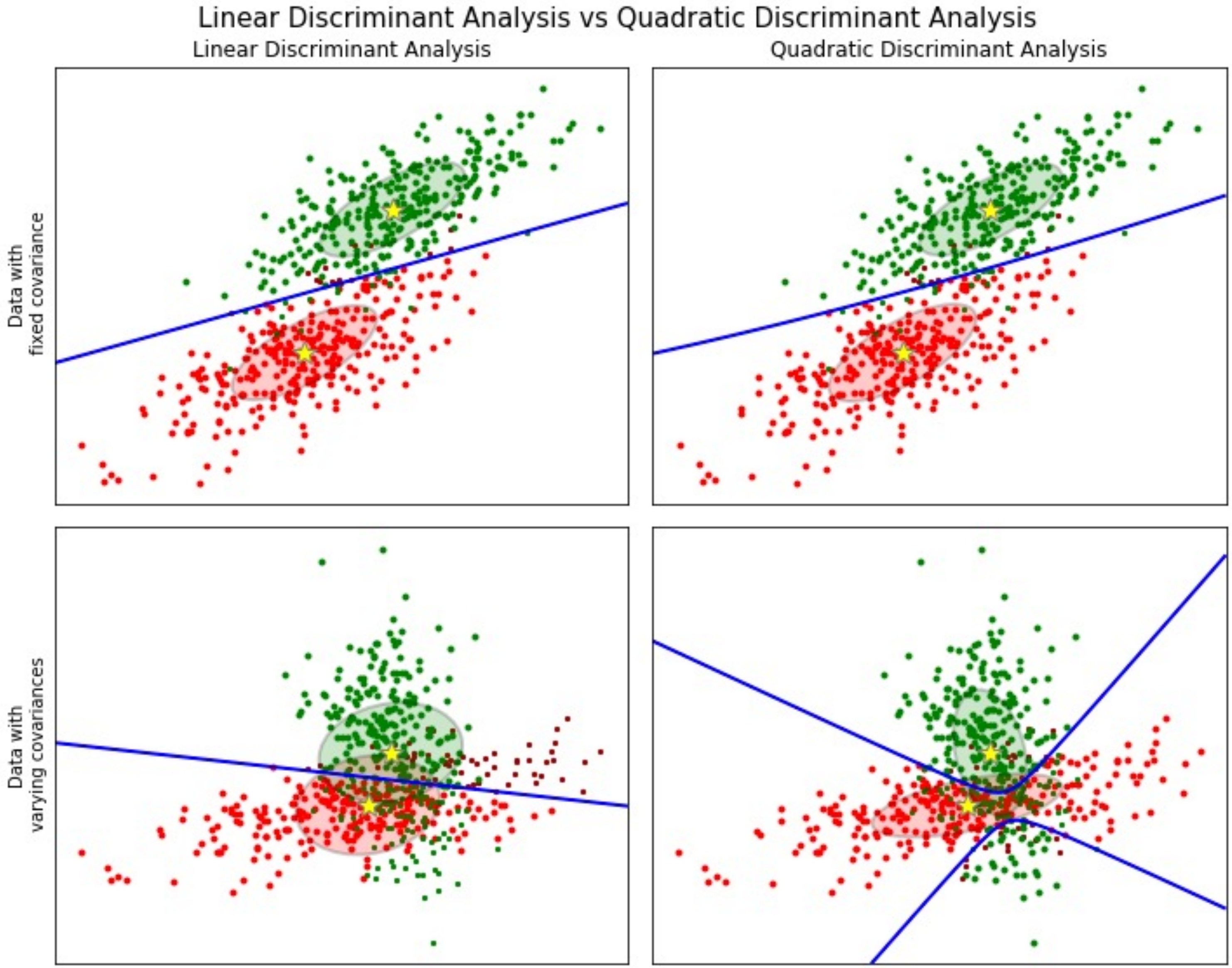
LDA assumption: $\Sigma_1 = \cdots = \Sigma_{\mathrm{K}} = \Sigma$



The left panel shows three Gaussian distributions, with the same covariance and different means. Included are the contours of constant density enclosing 95% of the probability in each case. The Bayes decision boundaries between each pair of classes are shown (broken straight lines), and the Bayes decision boundaries separating all three classes are the thicker solid lines (a subset of the former). On the right we see a sample of 30 drawn from each Gaussian distribution, and the fitted LDA decision boundaries.

Iris dataset





https://scikit-learn.org/stable/modules/lda_qda.html