Linear and Quadratic Discriminant Analysis

1. Regression Model

Ex: $\begin{cases} \beta_0 + \beta_1 x \\ \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p \end{cases}$ $\quad y = g(x) \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$

$PDF = f(y|\beta, x) \xrightarrow{-\log} -\log f(y|\beta, x)$

$\xrightarrow{\text{minimize}}$ to estimate β

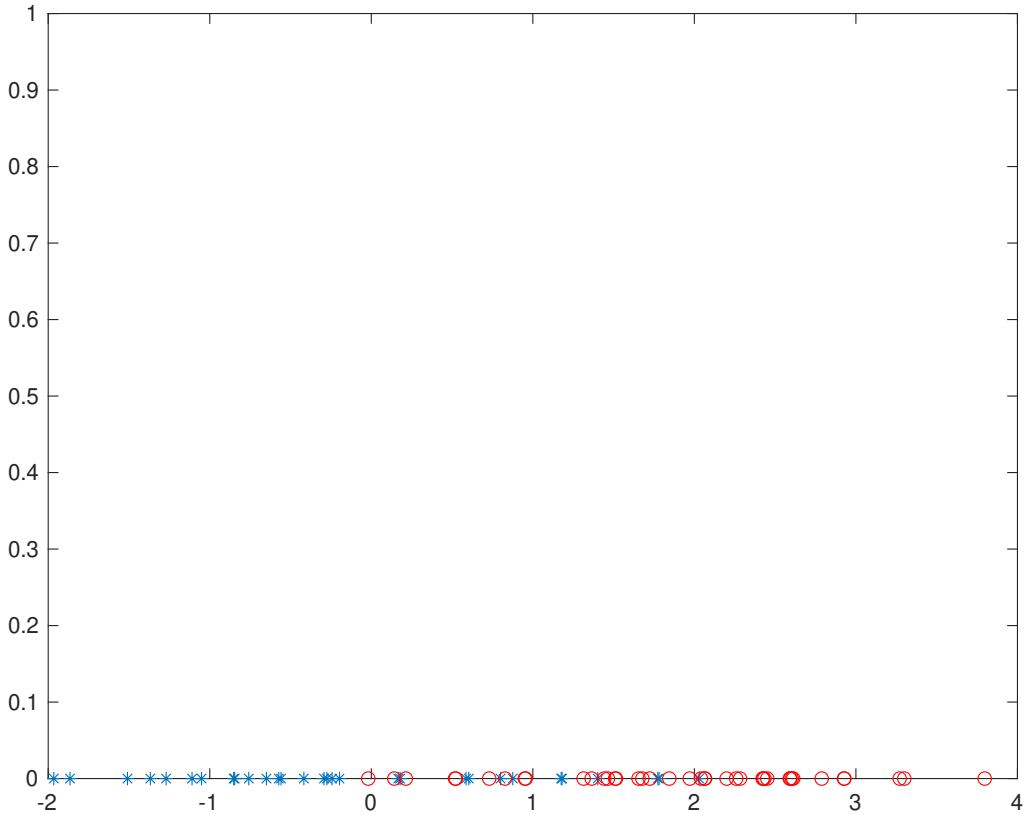
2. Classification $\quad y \in \{0, 1, 2, \dots, L\}$

$P(y=l \mid \beta^{(l)}, x), \quad l = 0, 1, 2, \dots, L$

$\xrightarrow{-\log} -\log P(y=l \mid \beta^{(l)}, x)$

$\xrightarrow{\text{minimize}}$ estimate $\beta^{(l)}$

LDA/QDA Story in Simple Words

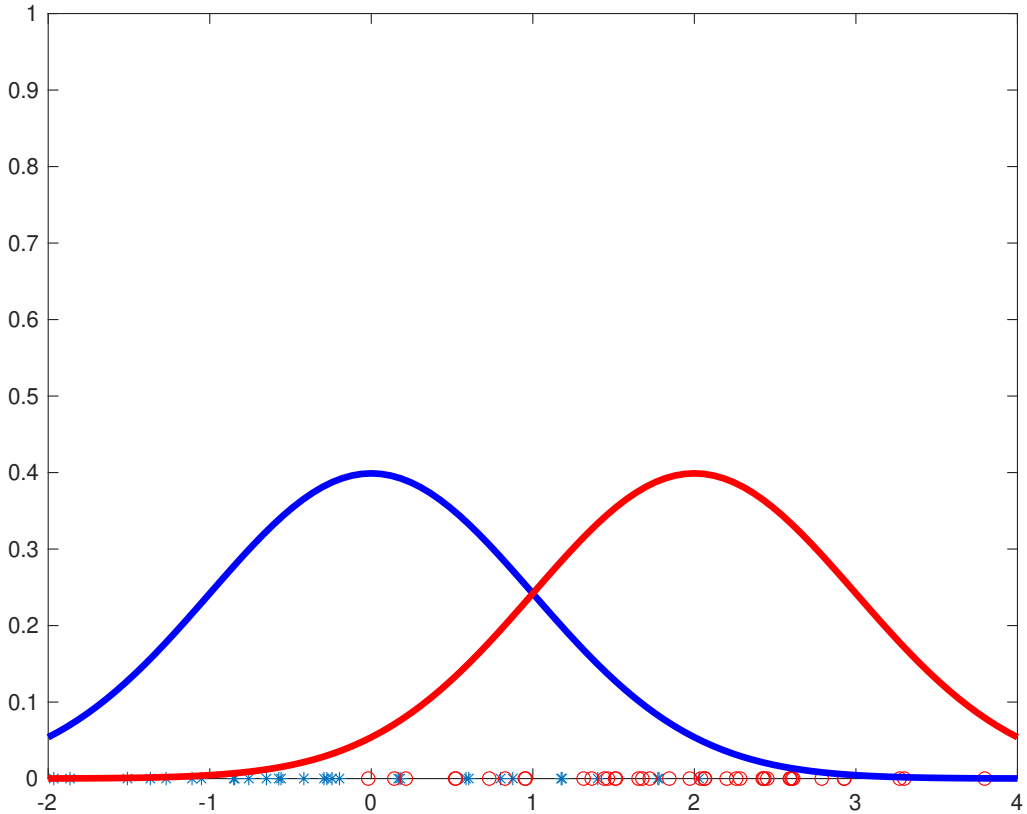


Let's do an exercise. We are given 30 blue points and 40 red points as above. Let's fit a Normal pdf to each cloud

$$X \sim (\mu, \sigma^2)$$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2}\right)$$

LDA/QDA Story in Simple Words

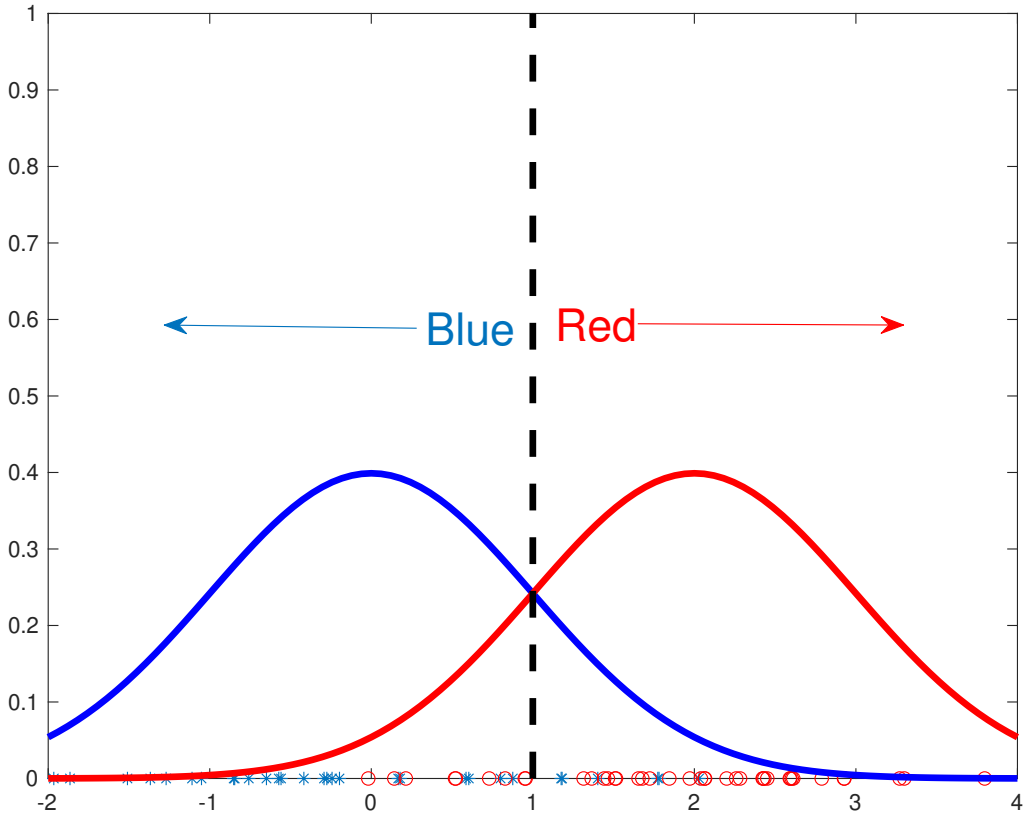


Let's assume both variances turn out similar. Find the intersection point. What can you say about this point?

LDA/QDA Story in Simple Words

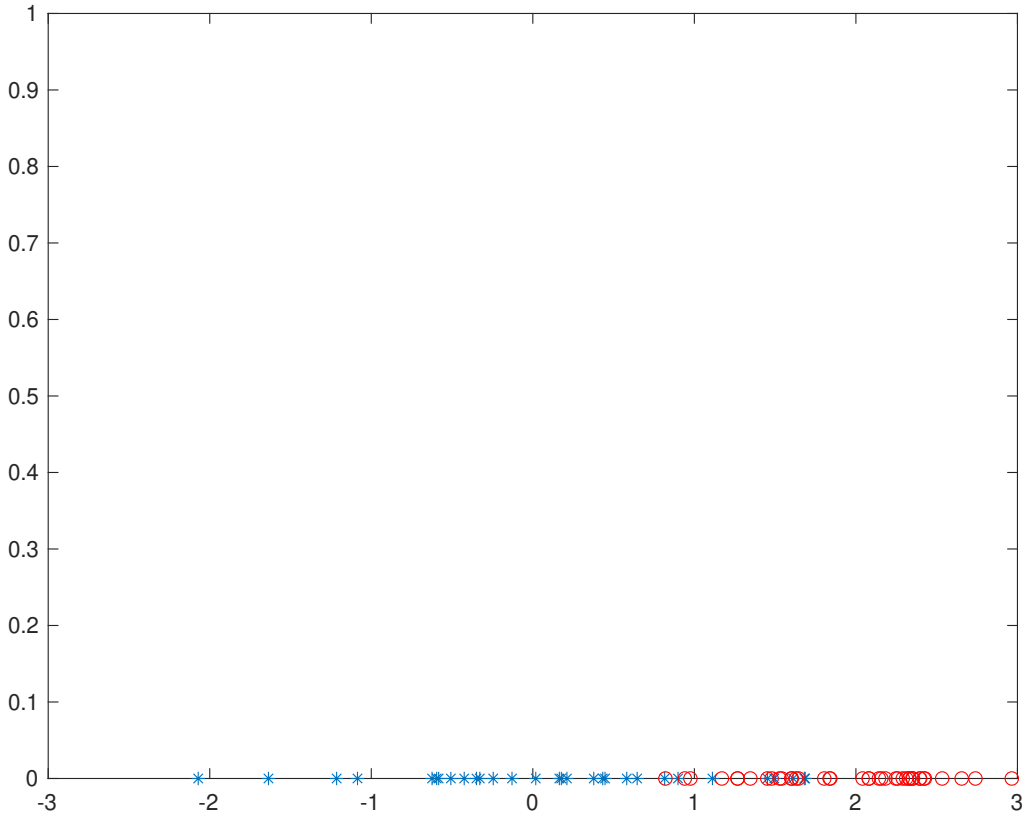
One dimension

1. Just Average the samples



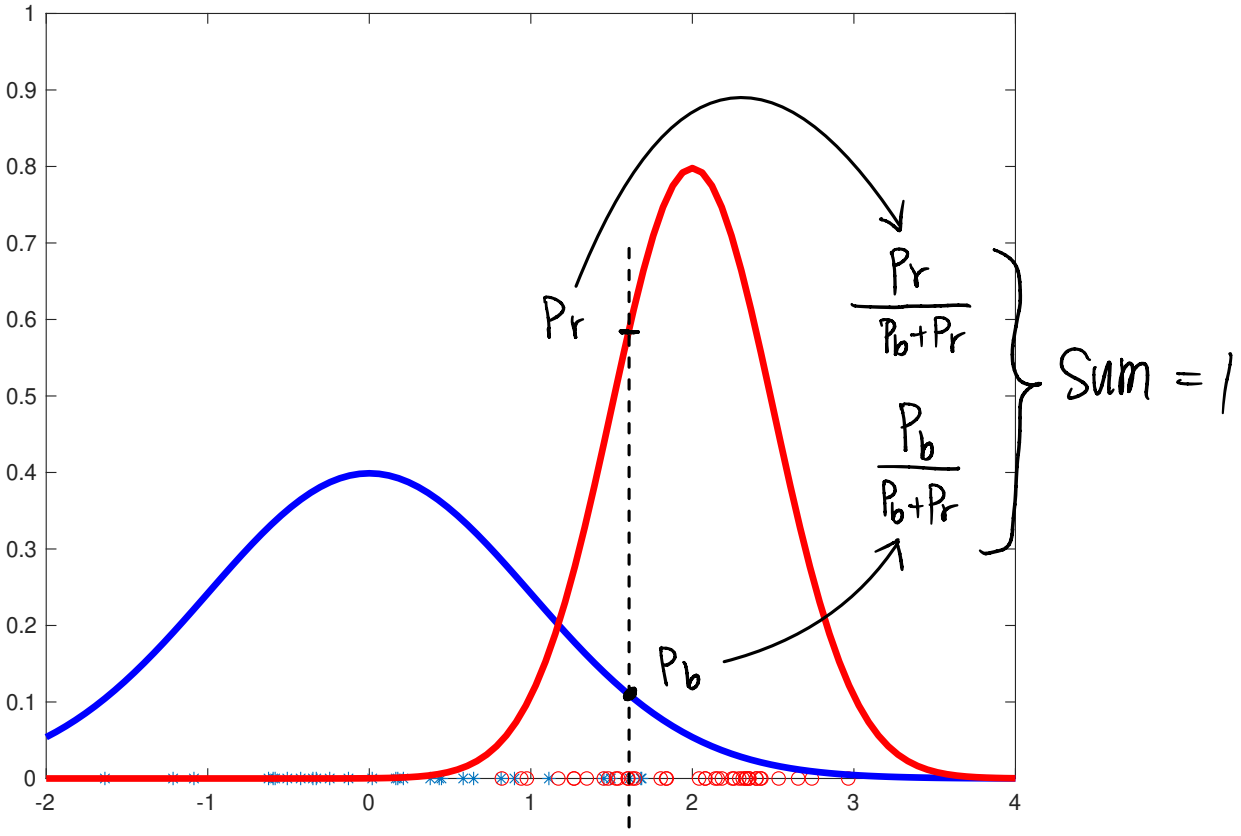
We have found a rule to classify each point on the real line

LDA/QDA Story in Simple Words



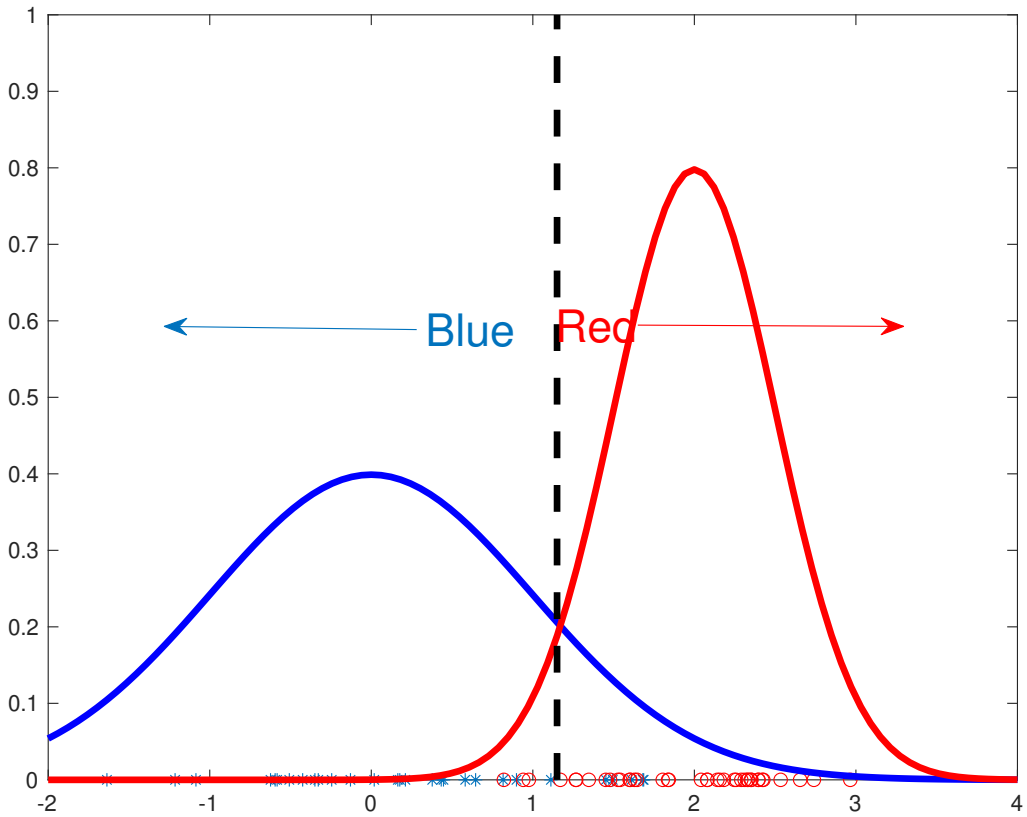
Let's do a different exercise with another set of points

LDA/QDA Story in Simple Words



The intersection point changes since the variances are no more the same.

LDA/QDA Story in Simple Words



Again we have a rule. If you want to assign a probability of being blue and being red to each point, how would you calculate that probability?

$$\mathbb{P}(Y = B|x) = p_B(x) = \frac{f_B(x)}{f_B(x) + f_R(x)}, \quad \mathbb{P}(Y = R|x) = p_R(x) = \frac{f_R(x)}{f_B(x) + f_R(x)}$$

LDA/QDA Story in Simple Words

$f_\ell = f(x|y=\ell)$

- Say you want to incorporate the probability of each class in your calculation of $\mathbb{P}(Y|x)$. We can use the Bayes formula

– Probabilistically, suppose that our y can take K distinct values. By the Bayes' theorem we have

Given Data $f(x|Y=\ell)$ → PDF

$$\mathbb{P}(Y=\ell|\boldsymbol{X}=\boldsymbol{x}) = \frac{\mathbb{P}(Y=\ell)\mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|Y=\ell)}{\sum_{k=1}^{K}\mathbb{P}(Y=k)\mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|Y=k)}$$

π_ℓ is the proportion $f(x|Y=k)$

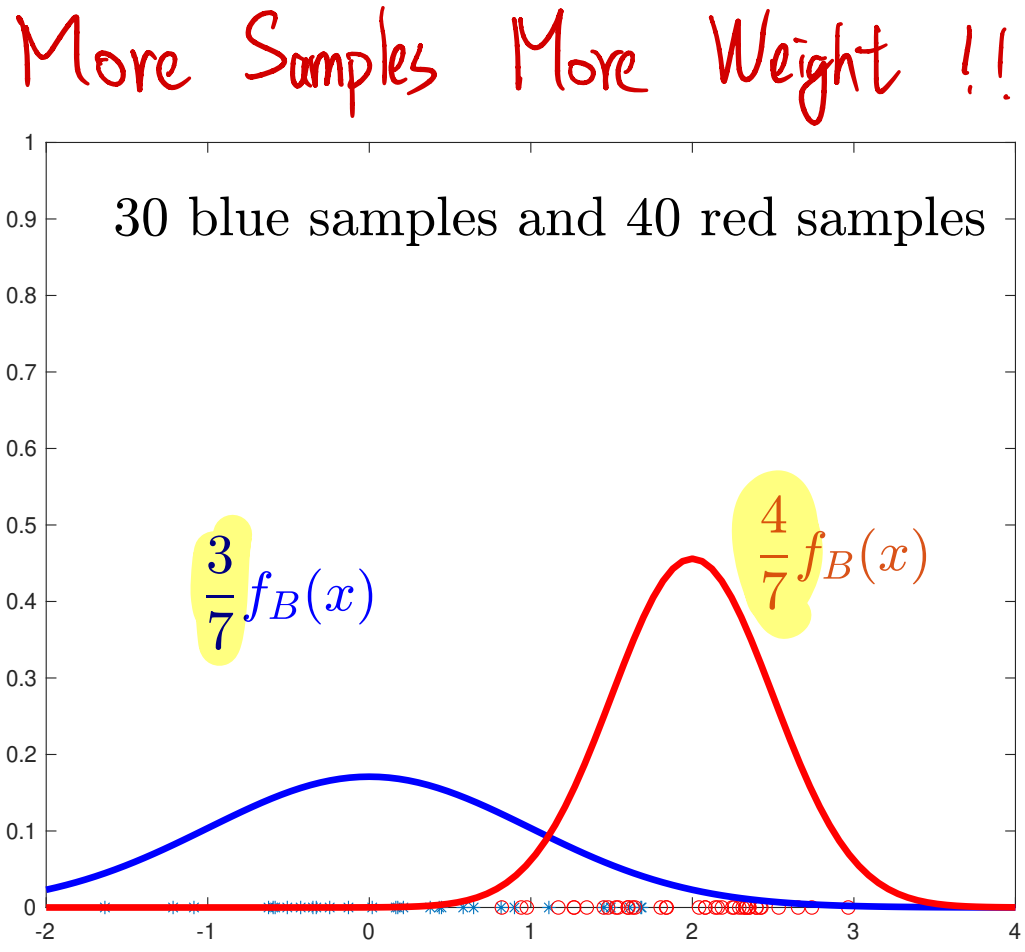
$$= \frac{\pi_\ell f_\ell(\boldsymbol{x})}{\sum_{k=1}^{K}\pi_k f_k(\boldsymbol{x})}$$

– Let's see why this equality holds, knowing the Bayes' equality

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)}$$

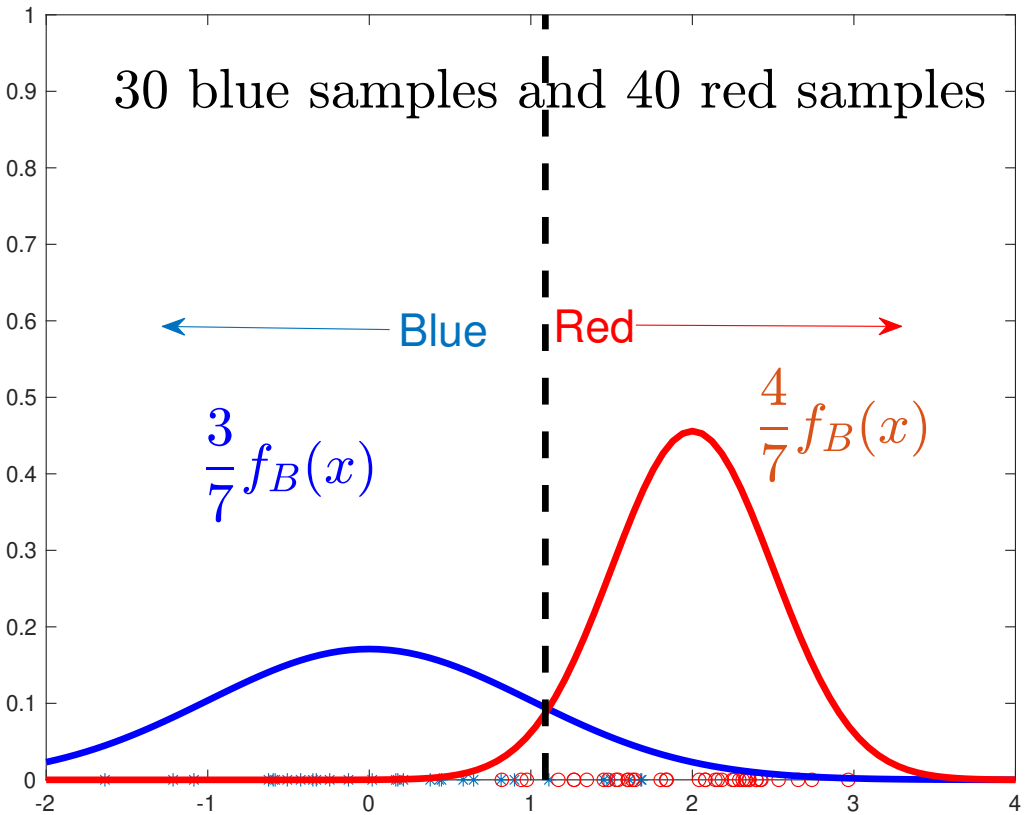
and $A_1 \cup A_2 \cdots \cup A_K$ covering the entire space, where $A_i \cap A_j = \emptyset$.

LDA/QDA Story in Simple Words



The intersection point yet changes

LDA/QDA Story in Simple Words

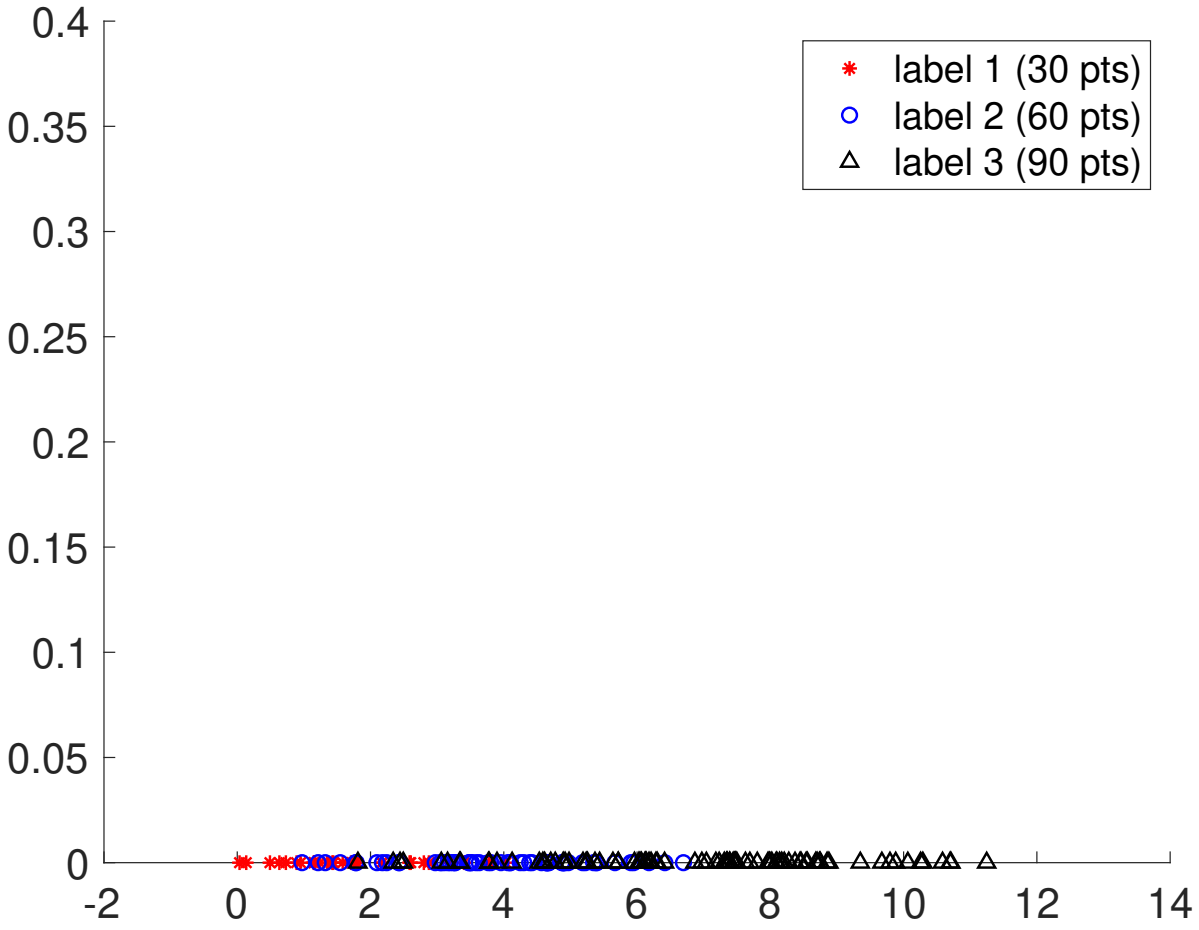


$$\mathbb{P}(Y = B|x) = p_B(x) = \frac{\frac{3}{7}f_B(x)}{\frac{3}{7}f_B(x) + \frac{4}{7}f_R(x)},$$

$$\mathbb{P}(Y = R|x) = p_R(x) = \frac{\frac{4}{7}f_R(x)}{\frac{3}{7}f_B(x) + \frac{4}{7}f_R(x)}$$

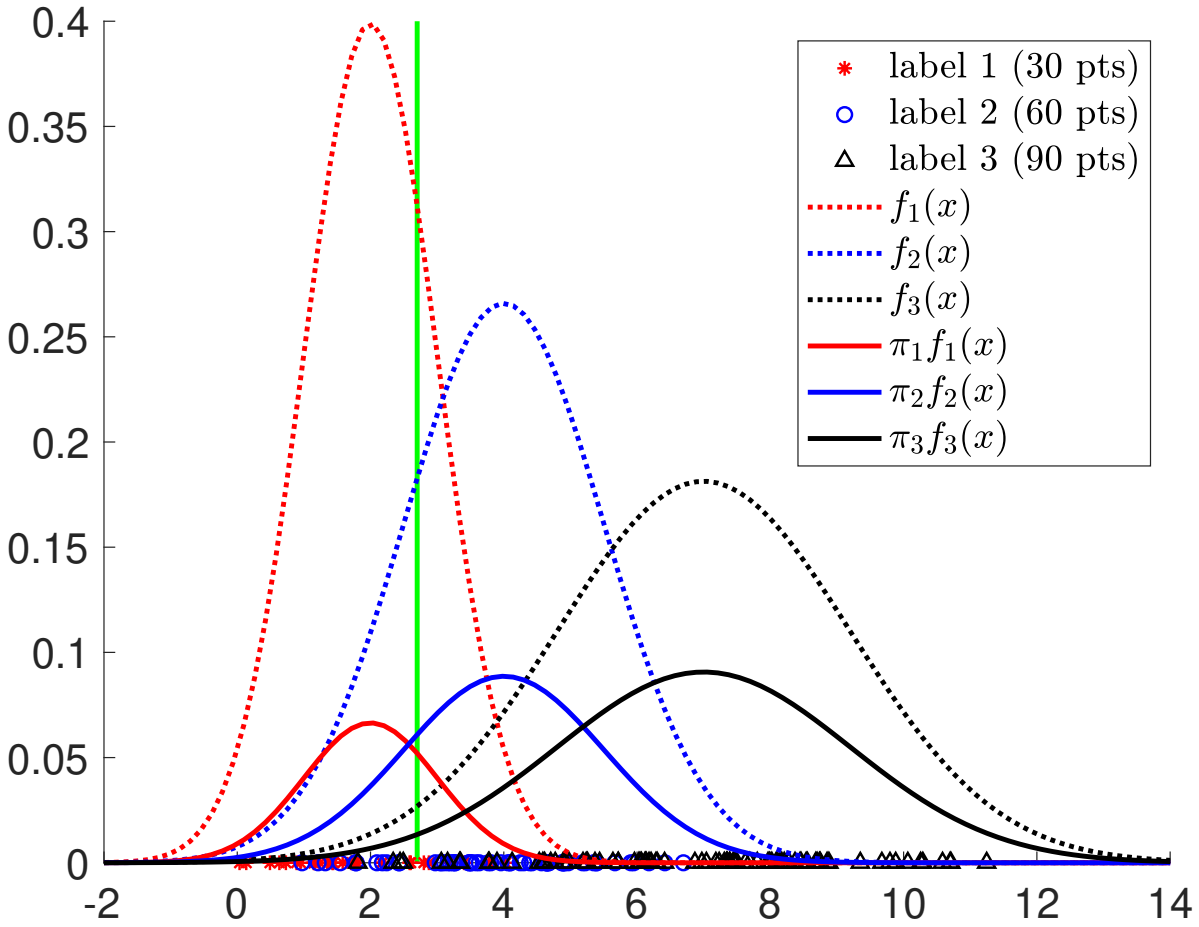
LDA/QDA Story in Simple Words

Question. What if we have more than two classes?



Story in Simple Words

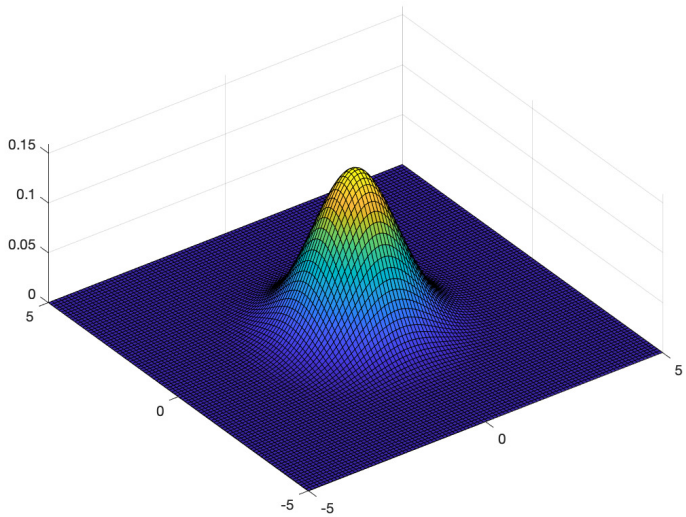
Answer. Same idea!

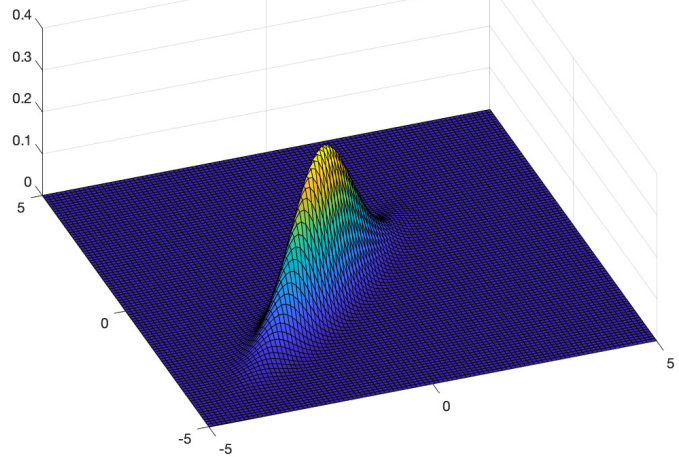


Story in Simple Words

Goal. Higher dimensions,

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$$





Also, in higher dimensions the intersection between two Normal pdfs is no more just a point, it would be a boundary

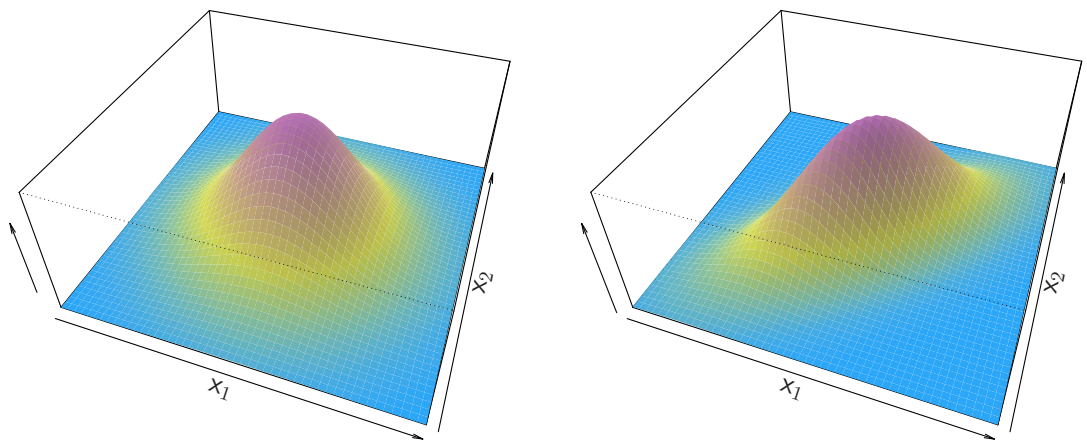
Little Introduction about Multivariate Normal

- Recall the normal distribution for a random variable x:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

- Similar to the scalar case, we can define a distribution for the random vector $\boldsymbol{x} = [x_1, \cdots, x_p]^T$ as

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



Linear Discriminant Analysis (LDA)

- Recall

$$\mathbb{P}(Y = \ell | \boldsymbol{X} = \boldsymbol{x}) = \frac{\mathbb{P}(Y = \ell)\mathbb{P}(\boldsymbol{X} = \boldsymbol{x} | Y = \ell)}{\sum_{k=1}^{K} \mathbb{P}(Y = k)\mathbb{P}(\boldsymbol{X} = \boldsymbol{x} | Y = k)} = \frac{\pi_\ell f_\ell(\boldsymbol{x})}{\sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x})}$$

- The purpose of LDA is learning a model for $\mathbb{P}(Y = \ell | \boldsymbol{X} = \boldsymbol{x})$
- In the formulation above, $f_\ell(\boldsymbol{x})$ is in a sense the distribution we consider for the data points in class ℓ, and π_ℓ is the probability that we pick some random sample and it belongs to class ℓ
- In LDA, we assume that all $f_\ell(\boldsymbol{x})$ have a multivariate normal distribution with similar covariances and different means, i.e.

$$f_\ell(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_\ell)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_\ell)\right)$$

- Unlike logistic regression, which involved a rather complicated maximization for learning, in LDA we have closed form expressions for π_ℓ, $\boldsymbol{\mu}_\ell$ and $\boldsymbol{\Sigma}$ and classifying new test points becomes very easy

Linear Discriminant Analysis (LDA)

– Given a training set $(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)$ where the responses y can take K distinct class values $1, 2, \cdots, K$, we can easily learn the LDA model by calculating π_ℓ, $\boldsymbol{\mu}_\ell$ and $\boldsymbol{\Sigma}$ via (considering c_ℓ to be the index of samples in class ℓ)

1. Calculate Proportion

$$\hat{\pi}_\ell = \frac{\#\text{ of elements in } c_\ell}{n}$$

2. Just Average samples

$$\hat{\boldsymbol{\mu}}_\ell = \frac{1}{\#\text{ of elements in } c_\ell} \sum_{i \in c_\ell} \boldsymbol{x}_i$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N - K} \sum_{k=1}^{K} \sum_{i \in c_\ell} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_\ell)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_\ell)^\top$$

Covariance matrix

? (N − K)

– After this point for a new test point $\boldsymbol{x}_t$ we have all that is needed to calculate $\mathbb{P}(Y = \ell | \boldsymbol{X} = \boldsymbol{x}_t)$ for $\ell = 1, \cdots, K$ and pick as the label the one that is largest

X_1	X_2	y
1	2	r
-1	-1	b
0	3	r
-2	1	r

$$\mu_r = \frac{\begin{bmatrix}1\\2\end{bmatrix} + \begin{bmatrix}0\\3\end{bmatrix} + \begin{bmatrix}-2\\1\end{bmatrix}}{3} = \begin{bmatrix}\frac{-1}{3}\\2\end{bmatrix}$$

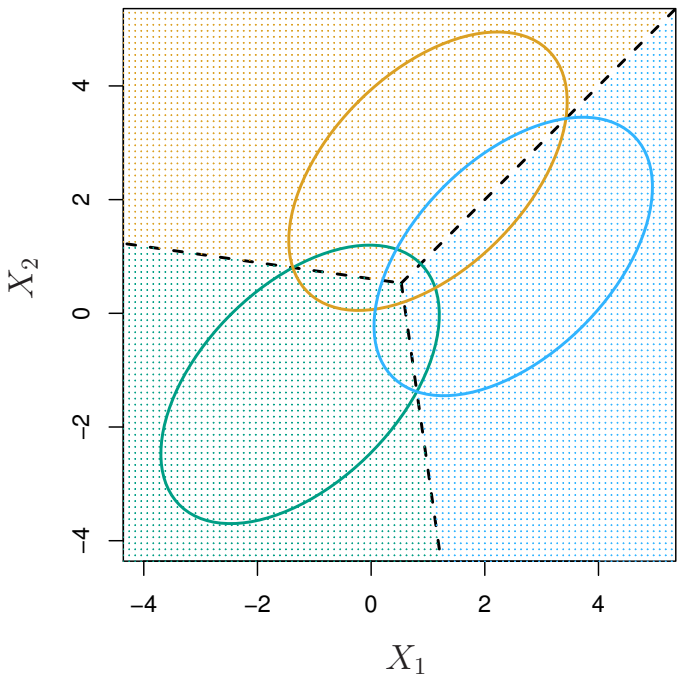
$$\mu_b = \begin{bmatrix}-1\\-1\end{bmatrix}$$

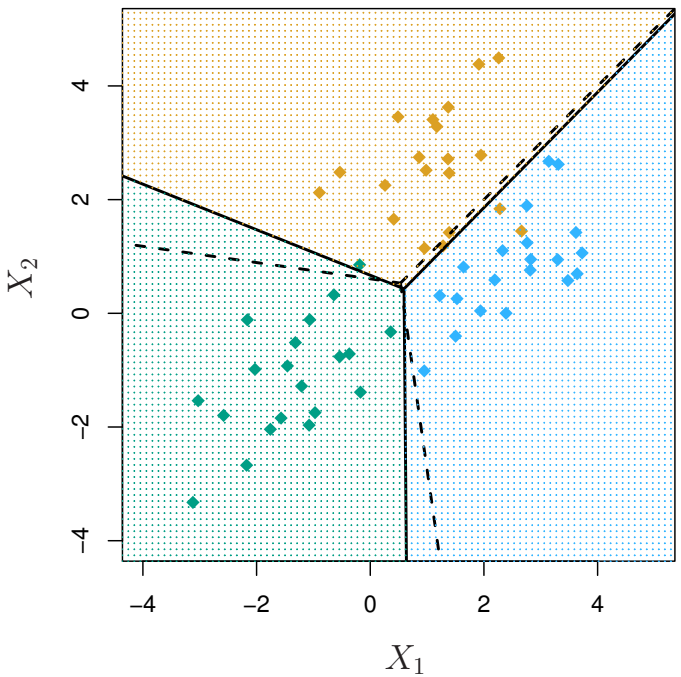
$$\Sigma_r = \left(\begin{bmatrix}1\\2\end{bmatrix} - \begin{bmatrix}\frac{-1}{3}\\2\end{bmatrix}\right)\left(\begin{bmatrix}1\\2\end{bmatrix} - \begin{bmatrix}\frac{-1}{3}\\2\end{bmatrix}\right)^T$$

$$+ \left(\begin{bmatrix}0\\3\end{bmatrix} - \begin{bmatrix}\frac{-1}{3}\\2\end{bmatrix}\right)\left(\begin{bmatrix}0\\3\end{bmatrix} - \begin{bmatrix}\frac{-1}{3}\\2\end{bmatrix}\right)^T$$

$$+ \left(\begin{bmatrix}-2\\1\end{bmatrix} - \begin{bmatrix}\frac{-1}{3}\\2\end{bmatrix}\right)\left(\begin{bmatrix}-2\\1\end{bmatrix} - \begin{bmatrix}\frac{-1}{3}\\2\end{bmatrix}\right)^T$$

Linear Discriminant Analysis (LDA)





Assume there are 2 classes

$$\pi_1 f_1(x) \longrightarrow \frac{\pi_1}{(2\pi)^{p/2}\,|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{(x-\mu_1)^T \Sigma^{-1} (x-\mu_1)}{2}\right)$$

$$\pi_2 f_2(x) \longrightarrow \frac{\pi_2}{(2\pi)^{p/2}\,|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{(x-\mu_2)^T \Sigma^{-1} (x-\mu_2)}{2}\right)$$

Decision Boundary : $\{x : \pi_1 f_1(x) = \pi_2 f_2(x)\}$

$$\pi_1 \exp\left(\frac{-1}{2}(x-\mu_1)^T \Sigma^{-1} (x-\mu_1)\right) = \pi_2$$

$\xrightarrow{\log}$ $\log \pi_1 - \frac{1}{2}(x-\mu_1)^T \Sigma^{-1} (x-\mu_1)$

$$\rightarrow \log \pi_1 - \frac{1}{2}\left[\underbrace{x^T \Sigma^{-1} x}_{=} - \mu_1^T \Sigma^{-1} x - x^T \Sigma^{-1} \mu_1 + \mu_1^T \Sigma^{-1} \mu_1\right]$$

$$= \log \pi_2 - \frac{1}{2}\left[x^T \Sigma^{-1} x - \mu_2^T \Sigma^{-1} x - x^T \Sigma^{-1} \mu_2 + \mu_2^T \Sigma^{-1} \mu_2\right]$$

Linear Discriminant Analysis (LDA)

- In practice to assign a label to a given test point $\boldsymbol{x}_t$ we do not need to calculate

$$\mathbb{P}(Y = \ell | \boldsymbol{X} = \boldsymbol{x}_t) = \frac{\pi_\ell f_\ell(\boldsymbol{x}_t)}{\sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_t)}$$

  and only comparing $\pi_\ell f_\ell(\boldsymbol{x}_t)$ is enough

- This reduces to evaluate

$$\delta_\ell = \boldsymbol{x}_t^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_\ell - \frac{1}{2} \boldsymbol{\mu}_\ell^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_\ell + \log \pi_\ell$$

  and pick as the class ℓ corresponding to the largest δ_ℓ

- You can find the decision boundary between class i and j by finding the points for which $\delta_i = \delta_j$

- [see the sample Matlab code]

Quadratic Discriminant Analysis (QDA)

- Recall

$$\mathbb{P}(Y = \ell | \boldsymbol{X} = \boldsymbol{x}_t) = \frac{\pi_\ell f_\ell(\boldsymbol{x}_t)}{\sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_t)}$$

- The purpose of QDA is learning a model for $\mathbb{P}(Y = \ell | \boldsymbol{X} = \boldsymbol{x})$ in a more flexible way compared to LDA
- In QDA, we assume that all $f_\ell(\boldsymbol{x})$ have a multivariate normal distribution with similar covariances and different means, i.e.

$$f_\ell(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}_\ell|^{1/2}} \exp\left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_\ell)^\top \boldsymbol{\Sigma}_\ell^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_\ell) \right)$$

- The main difference between LDA and QDA is in LDA we consider a single $\boldsymbol{\Sigma}$ for all classes, but in QDA we allow more flexibility by having a different covariance matrix for each class
- Similar to LDA, QDA can be learned easily and we can obtain closed form expressions for π_ℓ, $\boldsymbol{\mu}_\ell$ and $\boldsymbol{\Sigma}_\ell$

Quadratic Discriminant Analysis (QDA)

– Given a training set $(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)$ where the responses y can take K distinct class values $1, 2, \cdots, K$, we can easily learn the QDA model by calculating π_ℓ, $\boldsymbol{\mu}_\ell$ and $\boldsymbol{\Sigma}_\ell$ via (considering c_ℓ to be the index of samples in class ℓ)

$$\hat{\pi}_\ell = \frac{\#\text{ of elements in } c_\ell}{n}$$

$$\hat{\boldsymbol{\mu}}_\ell = \frac{1}{\#\text{ of elements in } c_\ell} \sum_{i \in c_\ell} \boldsymbol{x}_i$$

$$\hat{\boldsymbol{\Sigma}}_\ell = \frac{1}{\#\text{ of elements in } c_\ell - 1} \sum_{i \in c_\ell} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_\ell)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_\ell)^\top$$

– After this point for a new test point $\boldsymbol{x}_t$ we have all that is needed to calculate $\mathbb{P}(Y = \ell | \boldsymbol{X} = \boldsymbol{x}_t)$ for $\ell = 1, \cdots, K$ and pick as the label the one that is largest

Assume there are 2 classes

$$\pi_1 f_1(x) \longrightarrow \frac{\pi_1}{(2\pi)^{p/2}\,|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{(x-\mu_1)^T \Sigma^{-1} (x-\mu_1)}{2}\right)$$

$$\pi_2 f_2(x) \longrightarrow \frac{\pi_2}{(2\pi)^{p/2}\,|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{(x-\mu_2)^T \Sigma^{-1} (x-\mu_2)}{2}\right)$$

Decision Boundary : $\{x : \pi_1 f_1(x) = \pi_2 f_2(x)\}$

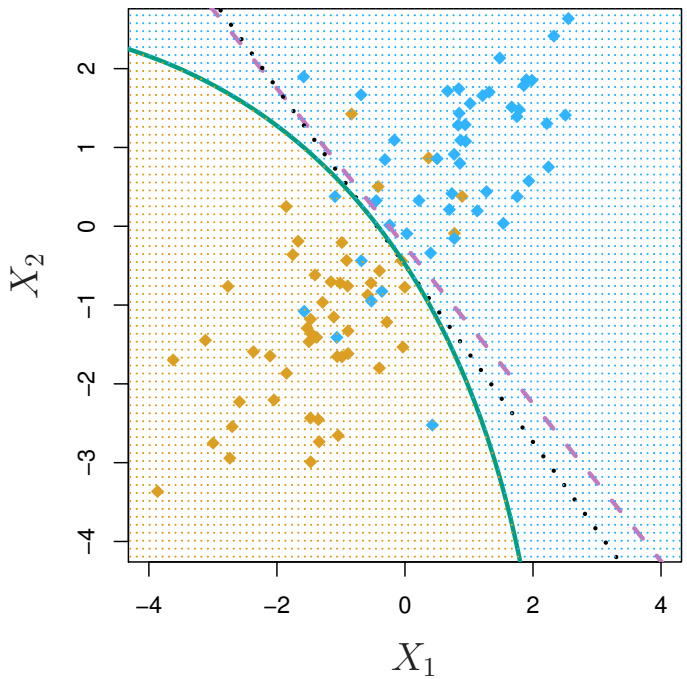
$$\pi_1 \exp\left(\frac{-1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1)\right) = \pi_2$$

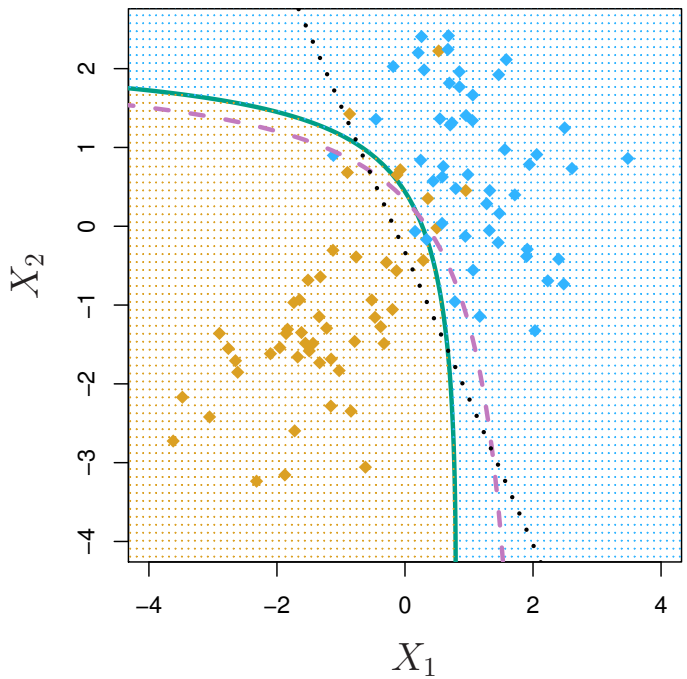
$$\xrightarrow{\log} \log \pi_1 - \frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1)$$

$$\rightarrow \log \pi_1 - \frac{1}{2}\left[x^T \Sigma_1^{-1} x - \mu_1^T \Sigma_1^{-1} x - x^T \Sigma_1^{-1} \mu_1 + \mu_1^T \Sigma_1^{-1} \mu_1\right]$$

$$= \log \pi_2 - \frac{1}{2}\left[x^T \Sigma_2^{-1} x - \mu_2^T \Sigma_2^{-1} x - x^T \Sigma_2^{-1} \mu_2 + \mu_2^T \Sigma_2^{-1} \mu_2\right]$$

Quadratic Discriminant Analysis (QDA)





Quadratic Discriminant Analysis (QDA)

- In practice to assign a label to a given test point $\boldsymbol{x}_t$ we do not need to calculate

$$\mathbb{P}(Y = \ell | \boldsymbol{X} = \boldsymbol{x}_t) = \frac{\pi_\ell f_\ell(\boldsymbol{x}_t)}{\sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_t)}$$

  and only comparing $\pi_\ell f_\ell(\boldsymbol{x}_t)$ is enough

- This reduces to evaluate

$$\delta_\ell = -\frac{1}{2} \log |\boldsymbol{\Sigma}_\ell| - \frac{1}{2} (\boldsymbol{x}_t - \boldsymbol{\mu}_\ell)^\top \boldsymbol{\Sigma}_\ell^{-1} (\boldsymbol{x}_t - \boldsymbol{\mu}_\ell) + \log \pi_\ell$$

  and pick as the class the ℓ corresponding to the largest δ_ℓ

- [see the sample Matlab code]

Some R Simulations

– Let's preform some basic classification tasks in R!

Summary

- Logistic regression is very popular for classification, especially for binary classification
- LDA is especially useful when $K > 2$, the number of training samples is small, or the classes are well separated, and Gaussian assumptions are reasonable.
- QDA presents more flexibility in shaping the partitions compared to LDA
- Logistic regression can also fit quadratic boundaries like QDA, by explicitly including quadratic terms in the model

References


S. Boyd, S. P. Boyd, and L. Vandenberghe.
Convex optimization.
Cambridge university press, 2004.


I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio.
Deep learning, volume 1.
MIT press Cambridge, 2016,
link:http://www.deeplearningbook.org/contents/convnets.html.