

CEE 616: Probabilistic Machine Learning

M2 Linear Methods: L2a Linear Discriminant Analysis

Jimi Oke

UMass **Amherst**

College of Engineering

Tue, Sep 23, 2025

Outline

① Introduction

② LDA model

③ LD derivation

④ QDA

⑤ Summary

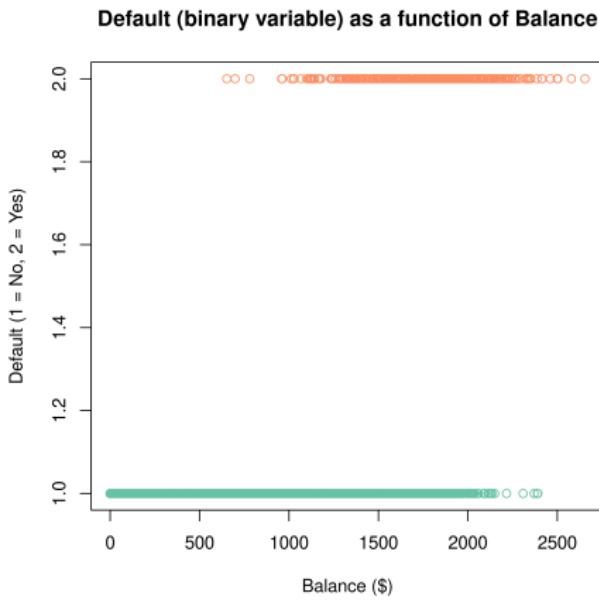
⑥ Appendix: GLMs

Why classify?

Consider the following plot:

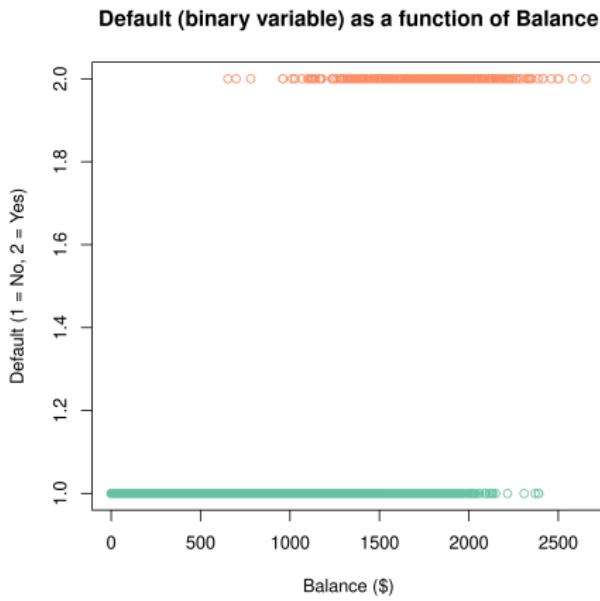
Why classify?

Consider the following plot:



Why classify?

Consider the following plot:



What problems would we face if we tried to fit a linear model to these data?

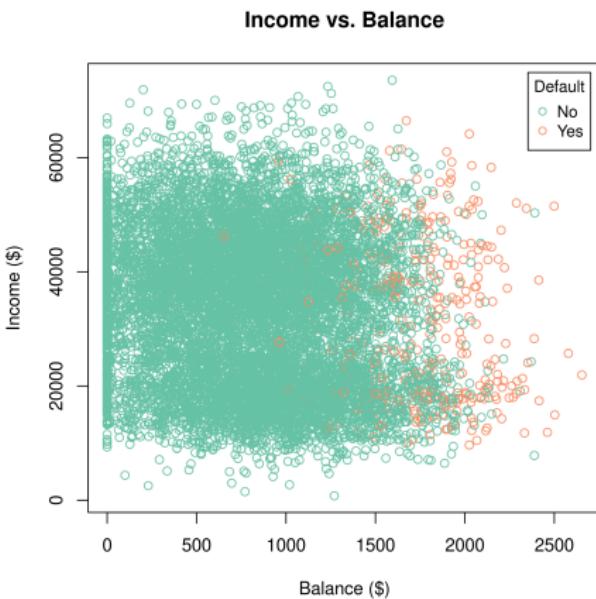
The classification problem

The classification problem

Consider the following plot

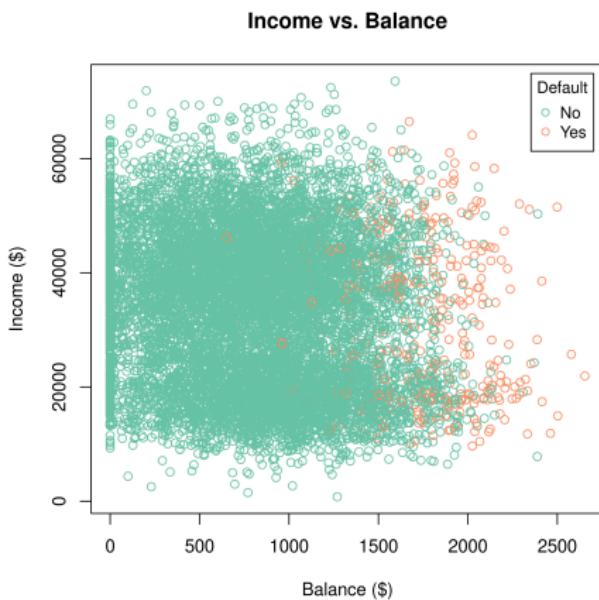
The classification problem

Consider the following plot



The classification problem

Consider the following plot



How would you estimate a model to classify a point $(x_{(i,\text{balance})}, x_{(i,\text{income})})$?

Bayes theorem for classification

Bayes theorem for classification

Given:

Bayes theorem for classification

Given:

- The conditional density of a variable x in class c :

Bayes theorem for classification

Given:

- The conditional density of a variable x in class c : $p_c(x)$

Bayes theorem for classification

Given:

- The conditional density of a variable \mathbf{x} in class c : $p_c(\mathbf{x})$
- The prior probability of class c :

Bayes theorem for classification

Given:

- The conditional density of a variable \mathbf{x} in class c : $p_c(\mathbf{x})$
- The prior probability of class c : π_c

Bayes theorem for classification

Given:

- The conditional density of a variable \mathbf{x} in class c : $p_c(\mathbf{x})$
- The prior probability of class c : π_c

such that:

Bayes theorem for classification

Given:

- The conditional density of a variable \mathbf{x} in class c : $p_c(\mathbf{x})$
- The prior probability of class c : π_c

$$\text{such that: } \sum_{c=1}^C \pi_c = 1 \quad (1)$$

Bayes theorem for classification

Given:

- The conditional density of a variable \mathbf{x} in class c : $p_c(\mathbf{x})$
- The prior probability of class c : π_c

$$\text{such that: } \sum_{c=1}^C \pi_c = 1 \quad (1)$$

Then, using Bayes theorem, we write **class posterior** as:

Bayes theorem for classification

Given:

- The conditional density of a variable \mathbf{x} in class c : $p_c(\mathbf{x})$
- The prior probability of class c : π_c

$$\text{such that: } \sum_{c=1}^C \pi_c = 1 \quad (1)$$

Then, using Bayes theorem, we write **class posterior** as:

$$p(y = c | \mathbf{x})$$

Bayes theorem for classification

Given:

- The conditional density of a variable \mathbf{x} in class c : $p_c(\mathbf{x})$
- The prior probability of class c : π_c

$$\text{such that: } \sum_{c=1}^C \pi_c = 1 \quad (1)$$

Then, using Bayes theorem, we write **class posterior** as:

$$p(y = c | \mathbf{x}) = \frac{p_c(\mathbf{x})\pi_c}{\sum_{c'=1}^C p_{c'}(\mathbf{x})\pi_{c'}} \quad (2)$$

Bayes theorem for classification

Given:

- The conditional density of a variable \mathbf{x} in class c : $p_c(\mathbf{x})$
- The prior probability of class c : π_c

$$\text{such that: } \sum_{c=1}^C \pi_c = 1 \quad (1)$$

Then, using Bayes theorem, we write **class posterior** as:

$$p(y = c | \mathbf{x}) = \frac{p_c(\mathbf{x})\pi_c}{\sum_{c'=1}^C p_{c'}(\mathbf{x})\pi_{c'}} \quad (2)$$

With the class posteriors, we can then assign an observation i using the Bayes' classifier:

Bayes theorem for classification

Given:

- The conditional density of a variable \mathbf{x} in class c : $p_c(\mathbf{x})$
- The prior probability of class c : π_c

$$\text{such that: } \sum_{c=1}^C \pi_c = 1 \quad (1)$$

Then, using Bayes theorem, we write **class posterior** as:

$$p(y = c | \mathbf{x}) = \frac{p_c(\mathbf{x})\pi_c}{\sum_{c'=1}^C p_{c'}(\mathbf{x})\pi_{c'}} \quad (2)$$

With the class posteriors, we can then assign an observation i using the Bayes' classifier:

$$y_i^* = \arg \max_k p(y = c | \mathbf{x}_i) \quad (3)$$

Bayes theorem for classification

Given:

- The conditional density of a variable \mathbf{x} in class c : $p_c(\mathbf{x})$
- The prior probability of class c : π_c

$$\text{such that: } \sum_{c=1}^C \pi_c = 1 \quad (1)$$

Then, using Bayes theorem, we write **class posterior** as:

$$p(y = c | \mathbf{x}) = \frac{p_c(\mathbf{x})\pi_c}{\sum_{c'=1}^C p_{c'}(\mathbf{x})\pi_{c'}} \quad (2)$$

With the class posteriors, we can then assign an observation i using the Bayes' classifier:

$$y_i^* = \arg \max_k p(y = c | \mathbf{x}_i) \quad (3)$$

(i.e. we assign to the class with the maximum probability)

Generative classifiers

Generative classifiers

A generative classifier is a model of the form:

Generative classifiers

A generative classifier is a model of the form:

$$p(y = c|\mathbf{x}; \theta) = \frac{p(\mathbf{x}|y = c; \theta)p(y = c; \theta)}{\sum_{c'} p(\mathbf{x}|y = c'; \theta)p(y = c; \theta)} \quad (4)$$

Generative classifiers

A generative classifier is a model of the form:

$$p(y = c|\mathbf{x}; \theta) = \frac{p(\mathbf{x}|y = c; \theta)p(y = c; \theta)}{\sum_{c'} p(\mathbf{x}|y = c'; \theta)p(y = c; \theta)} \quad (4)$$

where:

- $p(y = c; \theta)$ is the class prior

Generative classifiers

A generative classifier is a model of the form:

$$p(y = c|\mathbf{x}; \theta) = \frac{p(\mathbf{x}|y = c; \theta)p(y = c; \theta)}{\sum_{c'} p(\mathbf{x}|y = c'; \theta)p(y = c; \theta)} \quad (4)$$

where:

- $p(y = c; \theta)$ is the class prior
- $p(\mathbf{x}|y = c; \theta)$ is the class conditional density for class c

Generative classifiers

A generative classifier is a model of the form:

$$p(y = c|\mathbf{x}; \theta) = \frac{p(\mathbf{x}|y = c; \theta)p(y = c; \theta)}{\sum_{c'} p(\mathbf{x}|y = c'; \theta)p(y = c; \theta)} \quad (4)$$

where:

- $p(y = c; \theta)$ is the class prior
- $p(\mathbf{x}|y = c; \theta)$ is the class conditional density for class c
- Generative classifiers generate features for each class by sampling from the class conditional density

Generative classifiers

A generative classifier is a model of the form:

$$p(y = c|\mathbf{x}; \theta) = \frac{p(\mathbf{x}|y = c; \theta)p(y = c; \theta)}{\sum_{c'} p(\mathbf{x}|y = c'; \theta)p(y = c'; \theta)} \quad (4)$$

where:

- $p(y = c; \theta)$ is the class prior
- $p(\mathbf{x}|y = c; \theta)$ is the class conditional density for class c
- Generative classifiers generate features for each class by sampling from the class conditional density
- Discriminate classifiers model the class posterior $p(y|\mathbf{x}; \theta)$ directly

Generative classifiers

A generative classifier is a model of the form:

$$p(y = c|x; \theta) = \frac{p(x|y = c; \theta)p(y = c; \theta)}{\sum_{c'} p(x|y = c'; \theta)p(y = c'; \theta)} \quad (4)$$

where:

- $p(y = c; \theta)$ is the class prior
- $p(x|y = c; \theta)$ is the class conditional density for class c
- Generative classifiers generate features for each class by sampling from the class conditional density
- Discriminate classifiers model the class posterior $p(y|x; \theta)$ directly
- Linear discriminant analysis (LDA) arises from the log posterior being a linear function of x

Why linear discriminant analysis?

Why linear discriminant analysis?

Ideal situation

We know the following for each class:

- True [conditional] probability densities: $p_c(x)$

Why linear discriminant analysis?

Ideal situation

We know the following for each class:

- True [conditional] probability densities: $p_c(x)$
- True parameters: θ_c

Why linear discriminant analysis?

Ideal situation

We know the following for each class:

- True [conditional] probability densities: $p_c(x)$
- True parameters: θ_c
- True prior probabilities: π_c

Why linear discriminant analysis?

Ideal situation

We know the following for each class:

- True [conditional] probability densities: $p_c(x)$
- True parameters: θ_c
- True prior probabilities: π_c

Why linear discriminant analysis?

Ideal situation

We know the following for each class:

- True [conditional] probability densities: $p_c(x)$
- True parameters: θ_c
- True prior probabilities: π_c

If these were known, the Bayes decision boundary could be computed exactly

Why linear discriminant analysis?

Ideal situation

We know the following for each class:

- True [conditional] probability densities: $p_c(x)$
- True parameters: θ_c
- True prior probabilities: π_c

If these were known, the Bayes decision boundary could be computed exactly

In reality

Why linear discriminant analysis?

Ideal situation

We know the following for each class:

- True [conditional] probability densities: $p_c(x)$
- True parameters: θ_c
- True prior probabilities: π_c

If these were known, the Bayes decision boundary could be computed exactly

In reality

We are not certain!

Why linear discriminant analysis?

Ideal situation

We know the following for each class:

- True [conditional] probability densities: $p_c(\mathbf{x})$
- True parameters: θ_c
- True prior probabilities: π_c

If these were known, the Bayes decision boundary could be computed exactly

In reality

We are not certain!

- Assume Gaussian/normal conditional densities $p_c(\mathbf{x})$

This outlines the **linear discriminant analysis (LDA)** method, which approximates the Bayes classifier

Why linear discriminant analysis?

Ideal situation

We know the following for each class:

- True [conditional] probability densities: $p_c(\mathbf{x})$
- True parameters: θ_c
- True prior probabilities: π_c

If these were known, the Bayes decision boundary could be computed exactly

In reality

We are not certain!

- Assume Gaussian/normal conditional densities $p_c(\mathbf{x})$
- Estimate the parameters $\hat{\mu}_c$ and $\hat{\sigma}^2$ from the sample data

This outlines the **linear discriminant analysis (LDA)** method, which approximates the Bayes classifier

Why linear discriminant analysis?

Ideal situation

We know the following for each class:

- True [conditional] probability densities: $p_c(\mathbf{x})$
- True parameters: θ_c
- True prior probabilities: π_c

If these were known, the Bayes decision boundary could be computed exactly

In reality

We are not certain!

- Assume Gaussian/normal conditional densities $p_c(\mathbf{x})$
- Estimate the parameters $\hat{\mu}_c$ and $\hat{\sigma}^2$ from the sample data
- Estimate the priors from the data

This outlines the **linear discriminant analysis (LDA)** method, which approximates the Bayes classifier

Linear discriminant analysis (LDA)

Linear discriminant analysis (LDA)

Let N be the total number of training observations and N_c the number of training observations in Class c .

Linear discriminant analysis (LDA)

Let N be the total number of training observations and N_c the number of training observations in Class c .

Parameter estimates

Linear discriminant analysis (LDA)

Let N be the total number of training observations and N_c the number of training observations in Class c .

Parameter estimates

Class sample mean:

Linear discriminant analysis (LDA)

Let N be the total number of training observations and N_c the number of training observations in Class c .

Parameter estimates

$$\text{Class sample mean: } \hat{\mu}_c = \frac{1}{n_c} \sum_{i:y_i=c} x_i \quad (5)$$

Linear discriminant analysis (LDA)

Let N be the total number of training observations and N_c the number of training observations in Class c .

Parameter estimates

$$\text{Class sample mean: } \hat{\mu}_c = \frac{1}{n_c} \sum_{i:y_i=c} x_i \quad (5)$$

Common sample variance:

Linear discriminant analysis (LDA)

Let N be the total number of training observations and N_c the number of training observations in Class c .

Parameter estimates

$$\text{Class sample mean: } \hat{\mu}_c = \frac{1}{n_c} \sum_{i:y_i=c} x_i \quad (5)$$

$$\text{Common sample variance: } \hat{\sigma}^2 = \frac{1}{N - C} \sum_{c=1}^C \sum_{i:y_i=c} (x_i - \hat{\mu}_c)^2 \quad (6)$$

Class priors

Linear discriminant analysis (LDA)

Let N be the total number of training observations and N_c the number of training observations in Class c .

Parameter estimates

$$\text{Class sample mean: } \hat{\mu}_c = \frac{1}{n_c} \sum_{i:y_i=c} x_i \quad (5)$$

$$\text{Common sample variance: } \hat{\sigma}^2 = \frac{1}{N - C} \sum_{c=1}^C \sum_{i:y_i=c} (x_i - \hat{\mu}_c)^2 \quad (6)$$

Class priors

Computed as the fraction of observations belonging to Class c :

Linear discriminant analysis (LDA)

Let N be the total number of training observations and N_c the number of training observations in Class c .

Parameter estimates

$$\text{Class sample mean: } \hat{\mu}_c = \frac{1}{n_c} \sum_{i:y_i=c} x_i \quad (5)$$

$$\text{Common sample variance: } \hat{\sigma}^2 = \frac{1}{N - C} \sum_{c=1}^C \sum_{i:y_i=c} (x_i - \hat{\mu}_c)^2 \quad (6)$$

Class priors

Computed as the fraction of observations belonging to Class c :

$$\hat{\pi}_c = \frac{N_c}{N} \quad (7)$$

Introduction
ooooo

LDA model
○●ooooo

LD derivation
oooooooo

QDA
oooo

Summary
○

Appendix: GLMs
ooooo

LDA classifier

LDA classifier

- Like Bayes, the LDA classifier assigns an observation to the class with the maximizing posterior probability:

LDA classifier

- Like Bayes, the LDA classifier assigns an observation to the class with the maximizing posterior probability:

$$y_i^* = \arg \max_c P(Y = c | X = x_i) \quad (8)$$

- However, this is equivalent to maximizing the RHS of Bayes theorem:

LDA classifier

- Like Bayes, the LDA classifier assigns an observation to the class with the maximizing posterior probability:

$$y_i^* = \arg \max_c P(Y = c | X = x_i) \quad (8)$$

- However, this is equivalent to maximizing the RHS of Bayes theorem:

$$y_i^*$$

LDA classifier

- Like Bayes, the LDA classifier assigns an observation to the class with the maximizing posterior probability:

$$y_i^* = \arg \max_c P(Y = c | X = x_i) \quad (8)$$

- However, this is equivalent to maximizing the RHS of Bayes theorem:

$$y_i^* = \arg \max_c \frac{f_c(x_i)\pi_c}{\sum_{\ell=1}^C f_\ell(x_i)\pi_\ell} \quad (9)$$

- Since the **denominator** is the same for all classes, the LDA rule assigns an observation to the class which maximizes the **discriminant function**, δ_c :

LDA classifier

- Like Bayes, the LDA classifier assigns an observation to the class with the maximizing posterior probability:

$$y_i^* = \arg \max_c P(Y = c | X = x_i) \quad (8)$$

- However, this is equivalent to maximizing the RHS of Bayes theorem:

$$y_i^* = \arg \max_c \frac{f_c(x_i)\pi_c}{\sum_{\ell=1}^C f_\ell(x_i)\pi_\ell} \quad (9)$$

- Since the **denominator** is the same for all classes, the LDA rule assigns an observation to the class which maximizes the **discriminant function**, δ_c :

$$\delta_c \sim \log [f_c(x_i)\pi_c] =$$

LDA classifier

- Like Bayes, the LDA classifier assigns an observation to the class with the maximizing posterior probability:

$$y_i^* = \arg \max_c P(Y = c | X = x_i) \quad (8)$$

- However, this is equivalent to maximizing the RHS of Bayes theorem:

$$y_i^* = \arg \max_c \frac{f_c(x_i)\pi_c}{\sum_{\ell=1}^C f_\ell(x_i)\pi_\ell} \quad (9)$$

- Since the **denominator** is the same for all classes, the LDA rule assigns an observation to the class which maximizes the **discriminant function**, δ_c :

$$\delta_c \sim \log [f_c(x_i)\pi_c] = \log \pi_c + \log f_c(x_i) \quad (10)$$

LDA classifier

- Like Bayes, the LDA classifier assigns an observation to the class with the maximizing posterior probability:

$$y_i^* = \arg \max_c P(Y = c | X = x_i) \quad (8)$$

- However, this is equivalent to maximizing the RHS of Bayes theorem:

$$y_i^* = \arg \max_c \frac{f_c(x_i)\pi_c}{\sum_{\ell=1}^C f_\ell(x_i)\pi_\ell} \quad (9)$$

- Since the **denominator** is the same for all classes, the LDA rule assigns an observation to the class which maximizes the **discriminant function**, δ_c :

$$\delta_c \sim \log [f_c(x_i)\pi_c] = \log \pi_c + \log f_c(x_i) \quad (10)$$

In LDA, we assume that f_c is Gaussian.

LDA classifier

- Like Bayes, the LDA classifier assigns an observation to the class with the maximizing posterior probability:

$$y_i^* = \arg \max_c P(Y = c | X = x_i) \quad (8)$$

- However, this is equivalent to maximizing the RHS of Bayes theorem:

$$y_i^* = \arg \max_c \frac{f_c(x_i)\pi_c}{\sum_{\ell=1}^C f_\ell(x_i)\pi_\ell} \quad (9)$$

- Since the **denominator** is the same for all classes, the LDA rule assigns an observation to the class which maximizes the **discriminant function**, δ_c :

$$\delta_c \sim \log [f_c(x_i)\pi_c] = \log \pi_c + \log f_c(x_i) \quad (10)$$

In LDA, we assume that f_c is Gaussian. In the univariate case, this is:

LDA classifier

- Like Bayes, the LDA classifier assigns an observation to the class with the maximizing posterior probability:

$$y_i^* = \arg \max_c P(Y = c | X = x_i) \quad (8)$$

- However, this is equivalent to maximizing the RHS of Bayes theorem:

$$y_i^* = \arg \max_c \frac{f_c(x_i)\pi_c}{\sum_{\ell=1}^C f_\ell(x_i)\pi_\ell} \quad (9)$$

- Since the **denominator** is the same for all classes, the LDA rule assigns an observation to the class which maximizes the **discriminant function**, δ_c :

$$\delta_c \sim \log [f_c(x_i)\pi_c] = \log \pi_c + \log f_c(x_i) \quad (10)$$

In LDA, we assume that f_c is Gaussian. In the univariate case, this is:

$$f(x_c) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_c)^2\right) \quad (11)$$

LDA classifier

- Like Bayes, the LDA classifier assigns an observation to the class with the maximizing posterior probability:

$$y_i^* = \arg \max_c P(Y = c | X = x_i) \quad (8)$$

- However, this is equivalent to maximizing the RHS of Bayes theorem:

$$y_i^* = \arg \max_c \frac{f_c(x_i)\pi_c}{\sum_{\ell=1}^C f_\ell(x_i)\pi_\ell} \quad (9)$$

- Since the **denominator** is the same for all classes, the LDA rule assigns an observation to the class which maximizes the **discriminant function**, δ_c :

$$\delta_c \sim \log [f_c(x_i)\pi_c] = \log \pi_c + \log f_c(x_i) \quad (10)$$

In LDA, we assume that f_c is Gaussian. In the univariate case, this is:

$$f(x_c) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_c)^2\right) \quad (11)$$

LDA classifier: univariate case

LDA classifier: univariate case

The univariate linear discriminant functions used by the LDA classifier are:

LDA classifier: univariate case

The univariate linear discriminant functions used by the LDA classifier are:

$$\hat{\delta}_c(x) = \log(\hat{\pi}_c) + x \frac{\hat{\mu}_c}{\hat{\sigma}^2} - \frac{\hat{\mu}_c^2}{2\hat{\sigma}^2} \quad (12)$$

LDA classifier: univariate case

The univariate linear discriminant functions used by the LDA classifier are:

$$\hat{\delta}_c(x) = \log(\hat{\pi}_c) + x \frac{\hat{\mu}_c}{\hat{\sigma}^2} - \frac{\hat{\mu}_c^2}{2\hat{\sigma}^2} \quad (12)$$

Observations are assigned to the class c which maximizes $\hat{\delta}_c(x)$.

LDA classifier: univariate case

The univariate linear discriminant functions used by the LDA classifier are:

$$\hat{\delta}_c(x) = \log(\hat{\pi}_c) + x \frac{\hat{\mu}_c}{\hat{\sigma}^2} - \frac{\hat{\mu}_c^2}{2\hat{\sigma}^2} \quad (12)$$

Observations are assigned to the class c which maximizes $\hat{\delta}_c(x)$.

Example: LDA for two classes with equal priors

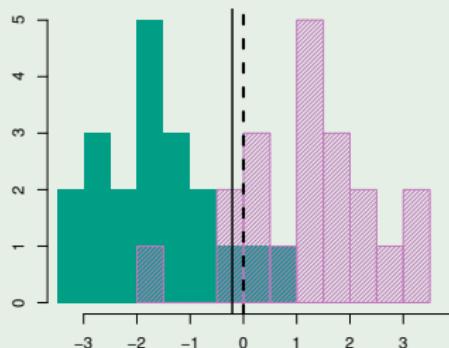


Figure: Bayes decision boundary (dashed)
and LDA decision boundary (solid)
estimated from the training data

LDA classifier: univariate case

The univariate linear discriminant functions used by the LDA classifier are:

$$\hat{\delta}_c(x) = \log(\hat{\pi}_c) + x \frac{\hat{\mu}_c}{\hat{\sigma}^2} - \frac{\hat{\mu}_c^2}{2\hat{\sigma}^2} \quad (12)$$

Observations are assigned to the class c which maximizes $\hat{\delta}_c(x)$.

Example: LDA for two classes with equal priors

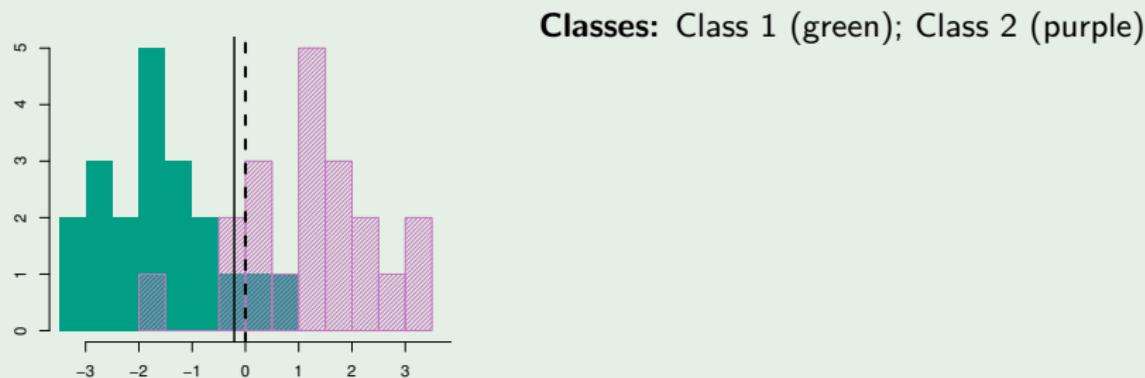


Figure: Bayes decision boundary (dashed)
and LDA decision boundary (solid)
estimated from the training data

LDA classifier: univariate case

The univariate linear discriminant functions used by the LDA classifier are:

$$\hat{\delta}_c(x) = \log(\hat{\pi}_c) + x \frac{\hat{\mu}_c}{\hat{\sigma}^2} - \frac{\hat{\mu}_c^2}{2\hat{\sigma}^2} \quad (12)$$

Observations are assigned to the class c which maximizes $\hat{\delta}_c(x)$.

Example: LDA for two classes with equal priors

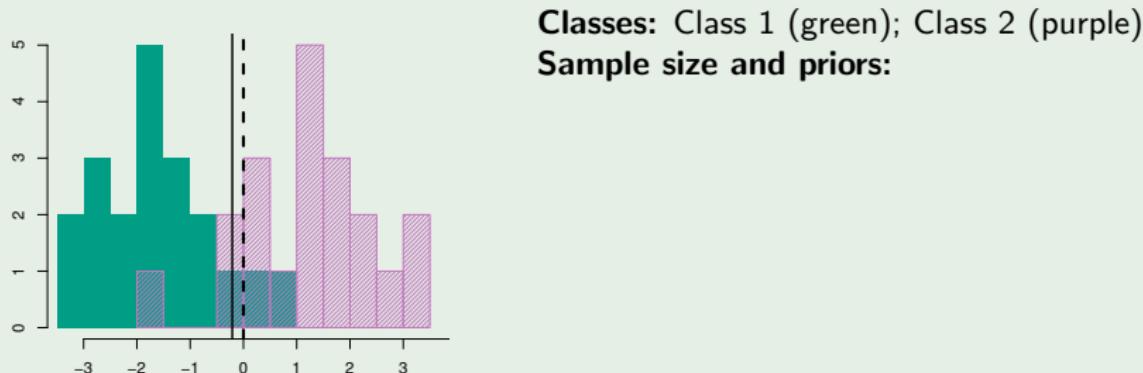


Figure: Bayes decision boundary (dashed)
and LDA decision boundary (solid)
estimated from the training data

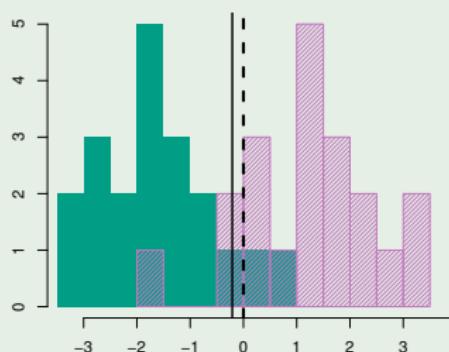
LDA classifier: univariate case

The univariate linear discriminant functions used by the LDA classifier are:

$$\hat{\delta}_c(x) = \log(\hat{\pi}_c) + x \frac{\hat{\mu}_c}{\hat{\sigma}^2} - \frac{\hat{\mu}_c^2}{2\hat{\sigma}^2} \quad (12)$$

Observations are assigned to the class c which maximizes $\hat{\delta}_c(x)$.

Example: LDA for two classes with equal priors



Classes: Class 1 (green); Class 2 (purple)
Sample size and priors:

$$n_1 = n_2 = 20$$

Figure: Bayes decision boundary (dashed)
and LDA decision boundary (solid)
estimated from the training data

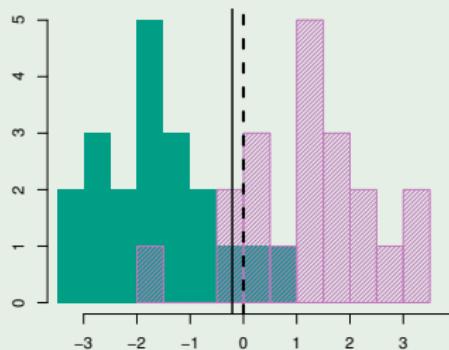
LDA classifier: univariate case

The univariate linear discriminant functions used by the LDA classifier are:

$$\hat{\delta}_c(x) = \log(\hat{\pi}_c) + x \frac{\hat{\mu}_c}{\hat{\sigma}^2} - \frac{\hat{\mu}_c^2}{2\hat{\sigma}^2} \quad (12)$$

Observations are assigned to the class c which maximizes $\hat{\delta}_c(x)$.

Example: LDA for two classes with equal priors



Classes: Class 1 (green); Class 2 (purple)
Sample size and priors:

$$n_1 = n_2 = 20$$

$$\hat{\pi}_1 = \hat{\pi}_2 =$$

Figure: Bayes decision boundary (dashed)
and LDA decision boundary (solid)
estimated from the training data

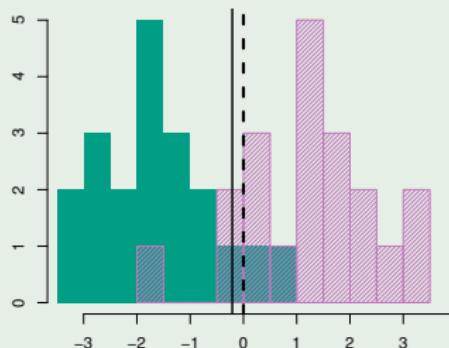
LDA classifier: univariate case

The univariate linear discriminant functions used by the LDA classifier are:

$$\hat{\delta}_c(x) = \log(\hat{\pi}_c) + x \frac{\hat{\mu}_c}{\hat{\sigma}^2} - \frac{\hat{\mu}_c^2}{2\hat{\sigma}^2} \quad (12)$$

Observations are assigned to the class c which maximizes $\hat{\delta}_c(x)$.

Example: LDA for two classes with equal priors



Classes: Class 1 (green); Class 2 (purple)
Sample size and priors:

$$\begin{aligned} n_1 &= n_2 = 20 \\ \hat{\pi}_1 &= \hat{\pi}_2 = \frac{20}{40} = \end{aligned}$$

Figure: Bayes decision boundary (dashed)
and LDA decision boundary (solid)
estimated from the training data

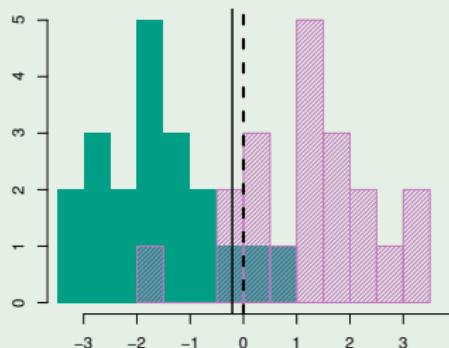
LDA classifier: univariate case

The univariate linear discriminant functions used by the LDA classifier are:

$$\hat{\delta}_c(x) = \log(\hat{\pi}_c) + x \frac{\hat{\mu}_c}{\hat{\sigma}^2} - \frac{\hat{\mu}_c^2}{2\hat{\sigma}^2} \quad (12)$$

Observations are assigned to the class c which maximizes $\hat{\delta}_c(x)$.

Example: LDA for two classes with equal priors



Classes: Class 1 (green); Class 2 (purple)
Sample size and priors:

$$\begin{aligned} n_1 &= n_2 = 20 \\ \hat{\pi}_1 &= \hat{\pi}_2 = \frac{20}{40} = 0.5 \end{aligned}$$

Figure: Bayes decision boundary (dashed)
and LDA decision boundary (solid)
estimated from the training data

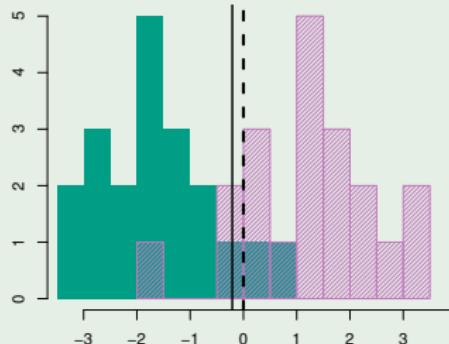
LDA classifier: univariate case

The univariate linear discriminant functions used by the LDA classifier are:

$$\hat{\delta}_c(x) = \log(\hat{\pi}_c) + x \frac{\hat{\mu}_c}{\hat{\sigma}^2} - \frac{\hat{\mu}_c^2}{2\hat{\sigma}^2} \quad (12)$$

Observations are assigned to the class c which maximizes $\hat{\delta}_c(x)$.

Example: LDA for two classes with equal priors



Classes: Class 1 (green); Class 2 (purple)
Sample size and priors:

$$\begin{aligned} n_1 &= n_2 = 20 \\ \hat{\pi}_1 &= \hat{\pi}_2 = \frac{20}{40} = 0.5 \end{aligned}$$

Decision boundary:

Figure: Bayes decision boundary (dashed)
and LDA decision boundary (solid)
estimated from the training data

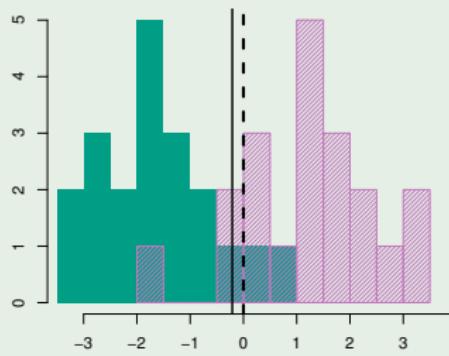
LDA classifier: univariate case

The univariate linear discriminant functions used by the LDA classifier are:

$$\hat{\delta}_c(x) = \log(\hat{\pi}_c) + x \frac{\hat{\mu}_c}{\hat{\sigma}^2} - \frac{\hat{\mu}_c^2}{2\hat{\sigma}^2} \quad (12)$$

Observations are assigned to the class c which maximizes $\hat{\delta}_c(x)$.

Example: LDA for two classes with equal priors



Classes: Class 1 (green); Class 2 (purple)

Sample size and priors:

$$n_1 = n_2 = 20$$

$$\hat{\pi}_1 = \hat{\pi}_2 = \frac{20}{40} = 0.5$$

Decision boundary:

$$x = \frac{\hat{\mu}_1 + \hat{\mu}_2}{2}$$

Figure: Bayes decision boundary (dashed)
and LDA decision boundary (solid)
estimated from the training data

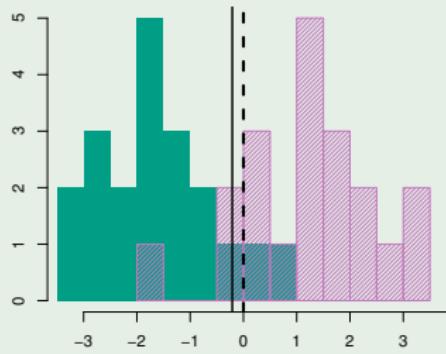
LDA classifier: univariate case

The univariate linear discriminant functions used by the LDA classifier are:

$$\hat{\delta}_c(x) = \log(\hat{\pi}_c) + x \frac{\hat{\mu}_c}{\hat{\sigma}^2} - \frac{\hat{\mu}_c^2}{2\hat{\sigma}^2} \quad (12)$$

Observations are assigned to the class c which maximizes $\hat{\delta}_c(x)$.

Example: LDA for two classes with equal priors



Classes: Class 1 (green); Class 2 (purple)

Sample size and priors:

$$n_1 = n_2 = 20$$

$$\hat{\pi}_1 = \hat{\pi}_2 = \frac{20}{40} = 0.5$$

Decision boundary:

$$x = \frac{\hat{\mu}_1 + \hat{\mu}_2}{2}$$

Figure: Bayes decision boundary (dashed) and LDA decision boundary (solid) estimated from the training data

Performance (error rate):

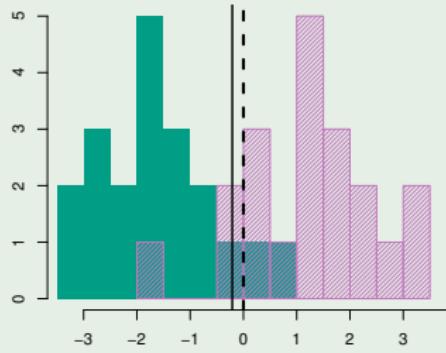
LDA classifier: univariate case

The univariate linear discriminant functions used by the LDA classifier are:

$$\hat{\delta}_c(x) = \log(\hat{\pi}_c) + x \frac{\hat{\mu}_c}{\hat{\sigma}^2} - \frac{\hat{\mu}_c^2}{2\hat{\sigma}^2} \quad (12)$$

Observations are assigned to the class c which maximizes $\hat{\delta}_c(x)$.

Example: LDA for two classes with equal priors



Classes: Class 1 (green); Class 2 (purple)

Sample size and priors:

$$n_1 = n_2 = 20$$

$$\hat{\pi}_1 = \hat{\pi}_2 = \frac{20}{40} = 0.5$$

Decision boundary:

$$x = \frac{\hat{\mu}_1 + \hat{\mu}_2}{2}$$

Figure: Bayes decision boundary (dashed) and LDA decision boundary (solid) estimated from the training data

Performance (error rate):
Bayes:

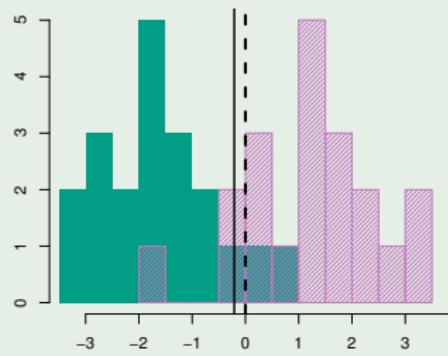
LDA classifier: univariate case

The univariate linear discriminant functions used by the LDA classifier are:

$$\hat{\delta}_c(x) = \log(\hat{\pi}_c) + x \frac{\hat{\mu}_c}{\hat{\sigma}^2} - \frac{\hat{\mu}_c^2}{2\hat{\sigma}^2} \quad (12)$$

Observations are assigned to the class c which maximizes $\hat{\delta}_c(x)$.

Example: LDA for two classes with equal priors



Classes: Class 1 (green); Class 2 (purple)

Sample size and priors:

$$n_1 = n_2 = 20$$

$$\hat{\pi}_1 = \hat{\pi}_2 = \frac{20}{40} = 0.5$$

Decision boundary:

$$x = \frac{\hat{\mu}_1 + \hat{\mu}_2}{2}$$

Figure: Bayes decision boundary (dashed) and LDA decision boundary (solid) estimated from the training data

Performance (error rate):
Bayes: 10.6%;

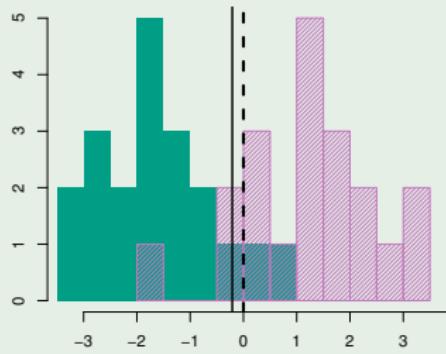
LDA classifier: univariate case

The univariate linear discriminant functions used by the LDA classifier are:

$$\hat{\delta}_c(x) = \log(\hat{\pi}_c) + x \frac{\hat{\mu}_c}{\hat{\sigma}^2} - \frac{\hat{\mu}_c^2}{2\hat{\sigma}^2} \quad (12)$$

Observations are assigned to the class c which maximizes $\hat{\delta}_c(x)$.

Example: LDA for two classes with equal priors



Classes: Class 1 (green); Class 2 (purple)

Sample size and priors:

$$n_1 = n_2 = 20$$

$$\hat{\pi}_1 = \hat{\pi}_2 = \frac{20}{40} = 0.5$$

Decision boundary:

$$x = \frac{\hat{\mu}_1 + \hat{\mu}_2}{2}$$

Figure: Bayes decision boundary (dashed) and LDA decision boundary (solid) estimated from the training data

Performance (error rate):
Bayes: 10.6%; LDA:

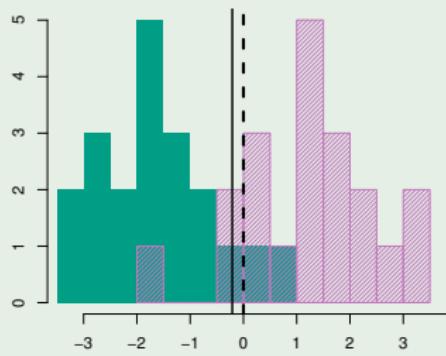
LDA classifier: univariate case

The univariate linear discriminant functions used by the LDA classifier are:

$$\hat{\delta}_c(x) = \log(\hat{\pi}_c) + x \frac{\hat{\mu}_c}{\hat{\sigma}^2} - \frac{\hat{\mu}_c^2}{2\hat{\sigma}^2} \quad (12)$$

Observations are assigned to the class c which maximizes $\hat{\delta}_c(x)$.

Example: LDA for two classes with equal priors



Classes: Class 1 (green); Class 2 (purple)

Sample size and priors:

$$n_1 = n_2 = 20$$

$$\hat{\pi}_1 = \hat{\pi}_2 = \frac{20}{40} = 0.5$$

Decision boundary:

$$x = \frac{\hat{\mu}_1 + \hat{\mu}_2}{2}$$

Figure: Bayes decision boundary (dashed) and LDA decision boundary (solid) estimated from the training data

Performance (error rate):
Bayes: 10.6%; LDA: 11.1%

Multiple predictors

Multiple predictors

For a **single** predictor x , we assumed a ***univariate*** Gaussian distribution.

Multiple predictors

For a **single** predictor x , we assumed a ***univariate*** Gaussian distribution.

If we have **multiple predictors**, then:

Multiple predictors

For a **single** predictor x , we assumed a *univariate* Gaussian distribution.

If we have **multiple predictors**, then:

$$\mathbf{x} = (x_1, x_2, \dots, x_D) \quad (13)$$

Multiple predictors

For a **single** predictor x , we assumed a *univariate* Gaussian distribution.

If we have **multiple predictors**, then:

$$\mathbf{x} = (x_1, x_2, \dots, x_D) \quad (13)$$

We then assume sample \mathbf{x} has a *multivariate* Gaussian distribution.

Multiple predictors

For a **single predictor x** , we assumed a *univariate* Gaussian distribution.

If we have **multiple predictors**, then:

$$\mathbf{x} = (x_1, x_2, \dots, x_D) \quad (13)$$

We then assume sample \mathbf{x} has a *multivariate* Gaussian distribution.

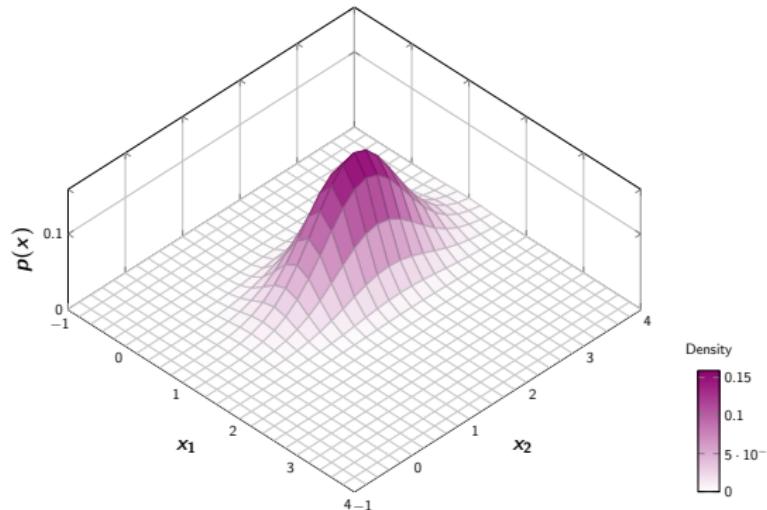


Figure: A multivariate Gaussian density function; $p = 2$ (bivariate Gaussian)

Multiple predictors

For a **single predictor** x , we assumed a *univariate* Gaussian distribution.

If we have **multiple predictors**, then:

$$\mathbf{x} = (x_1, x_2, \dots, x_D) \quad (13)$$

We then assume sample \mathbf{x} has a *multivariate* Gaussian distribution.

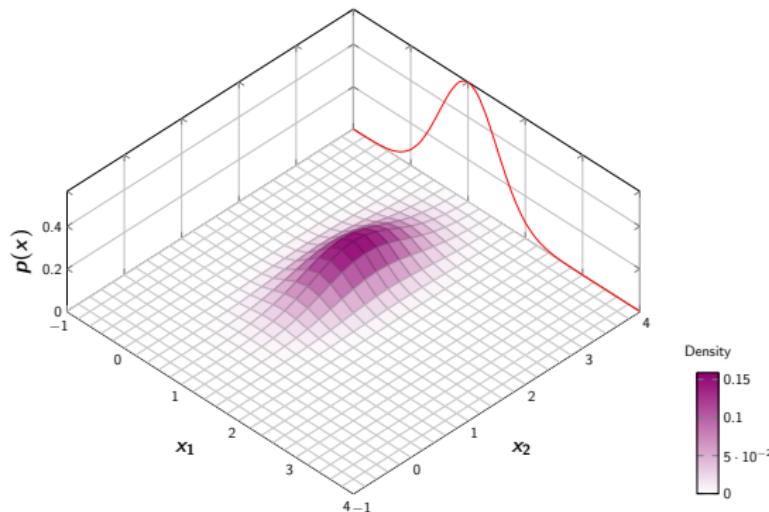


Figure: A multivariate Gaussian density function; $p = 2$ (bivariate Gaussian)

Multiple predictors

For a **single predictor** x , we assumed a *univariate* Gaussian distribution.

If we have **multiple predictors**, then:

$$\mathbf{x} = (x_1, x_2, \dots, x_D) \quad (13)$$

We then assume sample \mathbf{x} has a *multivariate* Gaussian distribution.

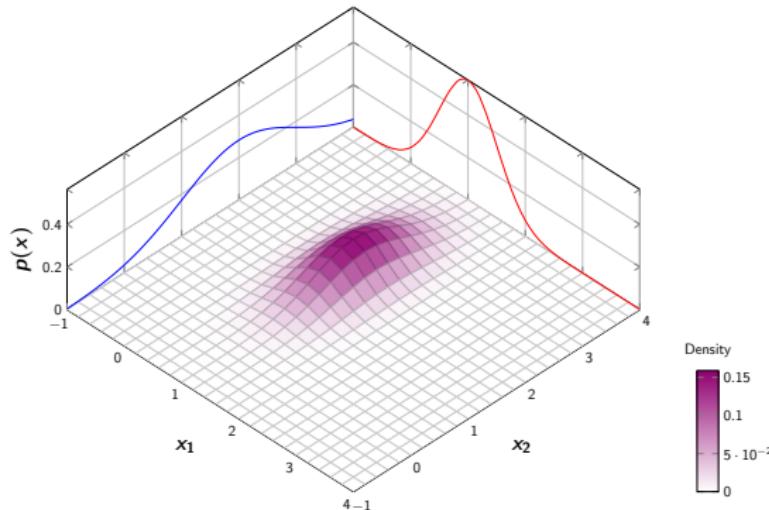


Figure: A multivariate Gaussian density function; $p = 2$ (bivariate Gaussian)

Multiple predictors

For a **single predictor x** , we assumed a *univariate* Gaussian distribution.

If we have **multiple predictors**, then:

$$\mathbf{x} = (x_1, x_2, \dots, x_D) \quad (13)$$

We then assume sample \mathbf{x} has a *multivariate* Gaussian distribution.

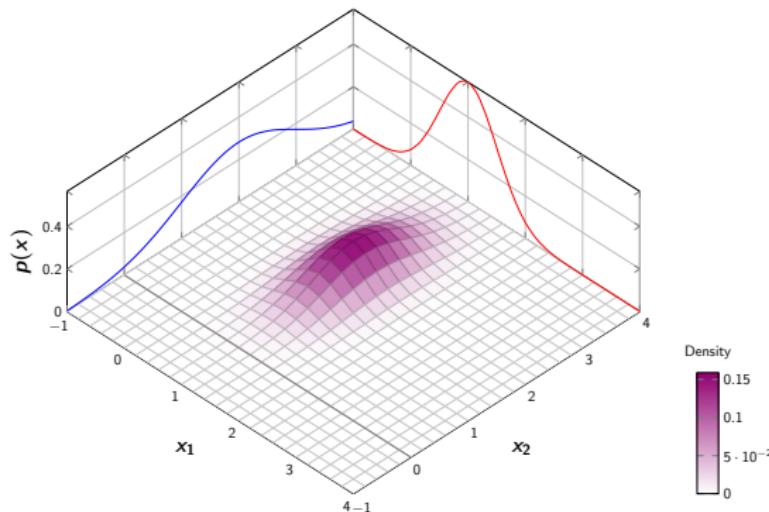


Figure: A multivariate Gaussian density function; $p = 2$ (bivariate Gaussian)

Multiple predictors

For a **single predictor x** , we assumed a *univariate* Gaussian distribution.

If we have **multiple predictors**, then:

$$\mathbf{x} = (x_1, x_2, \dots, x_D) \quad (13)$$

We then assume sample \mathbf{x} has a *multivariate* Gaussian distribution.

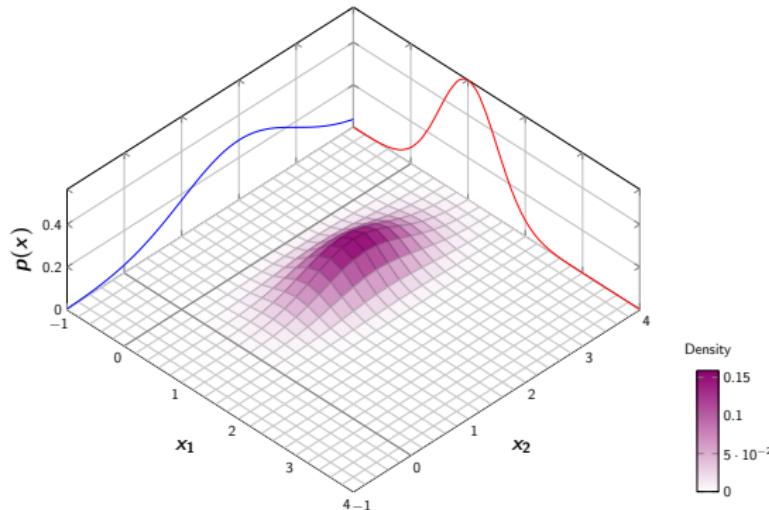


Figure: A multivariate Gaussian density function; $p = 2$ (bivariate Gaussian)

Multiple predictors

For a **single predictor** x , we assumed a *univariate* Gaussian distribution.

If we have **multiple predictors**, then:

$$\mathbf{x} = (x_1, x_2, \dots, x_D) \quad (13)$$

We then assume sample \mathbf{x} has a *multivariate* Gaussian distribution.

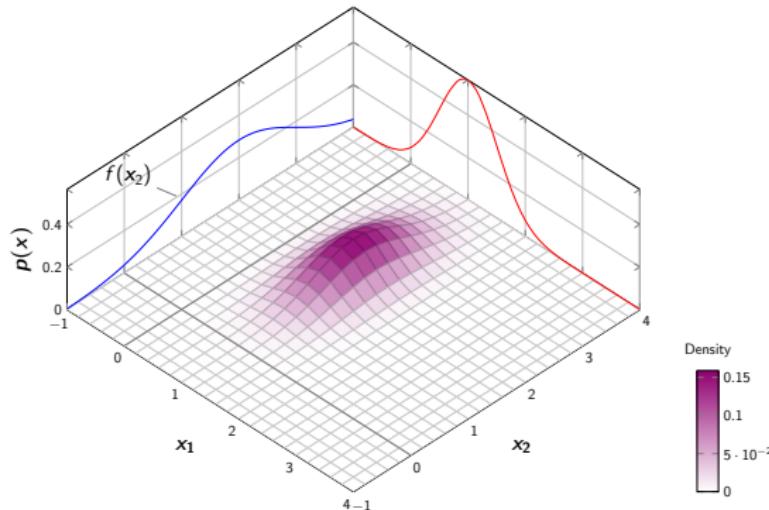


Figure: A multivariate Gaussian density function; $p = 2$ (bivariate Gaussian)

Multiple predictors

For a **single predictor x** , we assumed a ***univariate*** Gaussian distribution.

If we have **multiple predictors**, then:

$$\mathbf{x} = (x_1, x_2, \dots, x_D) \quad (13)$$

We then assume sample \mathbf{x} has a ***multivariate*** Gaussian distribution.

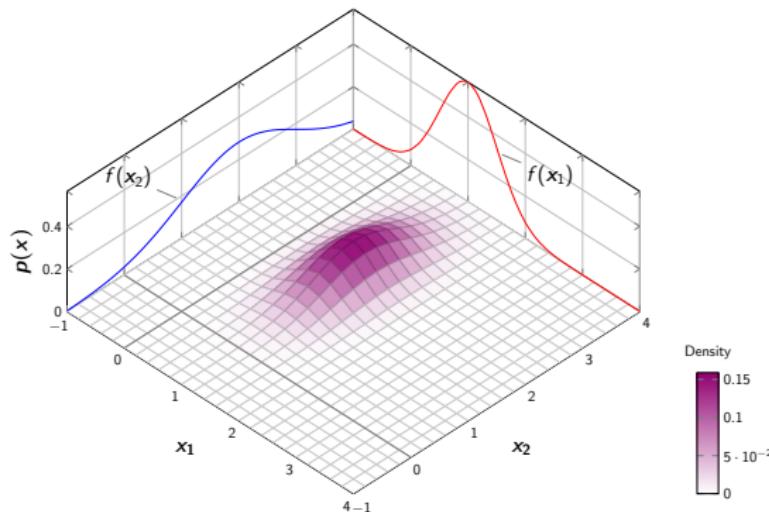


Figure: A multivariate Gaussian density function; $p = 2$ (bivariate Gaussian)

Multivariate Gaussian distributions: notation

Multivariate Gaussian distributions: notation

If X is a p -dimensional normally distributed random variable, then:

Multivariate Gaussian distributions: notation

If X is a p -dimensional normally distributed random variable, then:

$$X \sim \mathcal{N}(\mu, \Sigma) \quad (14)$$

Multivariate Gaussian distributions: notation

If X is a p -dimensional normally distributed random variable, then:

$$X \sim \mathcal{N}(\mu, \Sigma) \tag{14}$$

where $\mu = \mathbb{E}(x)$ and $\Sigma = \text{Cov}(x)$ is the $p \times p$ covariance matrix of X .

Multivariate Gaussian distributions: notation

If X is a p -dimensional normally distributed random variable, then:

$$X \sim \mathcal{N}(\mu, \Sigma) \quad (14)$$

where $\mu = \mathbb{E}(x)$ and $\Sigma = \text{Cov}(x)$ is the $p \times p$ covariance matrix of X .

$$\text{Cov}(x_j, X_I) = \mathbb{E}(x_j - \mathbb{E}(x_j)(x_I - \mathbb{E}(x_I)))$$

Multivariate Gaussian distributions: notation

If X is a p -dimensional normally distributed random variable, then:

$$X \sim \mathcal{N}(\mu, \Sigma) \quad (14)$$

where $\mu = \mathbb{E}(x)$ and $\Sigma = \text{Cov}(x)$ is the $p \times p$ covariance matrix of X .

$$\text{Cov}(x_j, X_l) = \mathbb{E}(x_j - \mathbb{E}(x_j)(x_l - \mathbb{E}(x_l))) = \mathbb{E}(x_j X_l) - \mathbb{E}(x_j)\mathbb{E}(x_l) \quad (15)$$

Multivariate Gaussian distributions: notation

If X is a p -dimensional normally distributed random variable, then:

$$X \sim \mathcal{N}(\mu, \Sigma) \quad (14)$$

where $\mu = \mathbb{E}(x)$ and $\Sigma = \text{Cov}(x)$ is the $p \times p$ covariance matrix of X .

$$\text{Cov}(x_j, X_l) = \mathbb{E}(x_j - \mathbb{E}(x_j)(x_l - \mathbb{E}(x_l)) = \mathbb{E}(x_j X_l) - \mathbb{E}(x_j)\mathbb{E}(x_l) \quad (15)$$

$$\text{Cov}(x_j, X_j) =$$

Multivariate Gaussian distributions: notation

If X is a p -dimensional normally distributed random variable, then:

$$X \sim \mathcal{N}(\mu, \Sigma) \quad (14)$$

where $\mu = \mathbb{E}(x)$ and $\Sigma = \text{Cov}(x)$ is the $p \times p$ covariance matrix of X .

$$\text{Cov}(x_j, X_l) = \mathbb{E}(x_j - \mathbb{E}(x_j)(x_l - \mathbb{E}(x_l)) = \mathbb{E}(x_j X_l) - \mathbb{E}(x_j)\mathbb{E}(x_l) \quad (15)$$

$$\text{Cov}(x_j, X_j) = \mathbb{E}(x_j^2) - (\mathbb{E}(x_j))^2 =$$

Multivariate Gaussian distributions: notation

If X is a p -dimensional normally distributed random variable, then:

$$X \sim \mathcal{N}(\mu, \Sigma) \quad (14)$$

where $\mu = \mathbb{E}(x)$ and $\Sigma = \text{Cov}(x)$ is the $p \times p$ covariance matrix of X .

$$\text{Cov}(x_j, X_l) = \mathbb{E}(x_j - \mathbb{E}(x_j)(x_l - \mathbb{E}(x_l)) = \mathbb{E}(x_j X_l) - \mathbb{E}(x_j)\mathbb{E}(x_l) \quad (15)$$

$$\text{Cov}(x_j, X_j) = \mathbb{E}(x_j^2) - (\mathbb{E}(x_j))^2 = \mathbb{V}(x_j) \quad (16)$$

Multivariate Gaussian distributions: notation

If X is a p -dimensional normally distributed random variable, then:

$$X \sim \mathcal{N}(\mu, \Sigma) \quad (14)$$

where $\mu = \mathbb{E}(x)$ and $\Sigma = \text{Cov}(x)$ is the $p \times p$ covariance matrix of X .

$$\text{Cov}(x_j, X_l) = \mathbb{E}(x_j - \mathbb{E}(x_j)(x_l - \mathbb{E}(x_l)) = \mathbb{E}(x_j X_l) - \mathbb{E}(x_j)\mathbb{E}(x_l) \quad (15)$$

$$\text{Cov}(x_j, X_j) = \mathbb{E}(x_j^2) - (\mathbb{E}(x_j))^2 = \mathbb{V}(x_j) \quad (16)$$

Probability density function

Multivariate Gaussian distributions: notation

If X is a p -dimensional normally distributed random variable, then:

$$X \sim \mathcal{N}(\mu, \Sigma) \quad (14)$$

where $\mu = \mathbb{E}(x)$ and $\Sigma = \text{Cov}(x)$ is the $p \times p$ covariance matrix of X .

$$\text{Cov}(x_j, X_l) = \mathbb{E}(x_j - \mathbb{E}(x_j)(x_l - \mathbb{E}(x_l)) = \mathbb{E}(x_j X_l) - \mathbb{E}(x_j)\mathbb{E}(x_l) \quad (15)$$

$$\text{Cov}(x_j, X_j) = \mathbb{E}(x_j^2) - (\mathbb{E}(x_j))^2 = \mathbb{V}(x_j) \quad (16)$$

Probability density function

The multivariate Gaussian/normal probability density function (PDF) is given by:

Multivariate Gaussian distributions: notation

If X is a p -dimensional normally distributed random variable, then:

$$X \sim \mathcal{N}(\mu, \Sigma) \quad (14)$$

where $\mu = \mathbb{E}(x)$ and $\Sigma = \text{Cov}(x)$ is the $p \times p$ covariance matrix of X .

$$\text{Cov}(x_j, X_l) = \mathbb{E}(x_j - \mathbb{E}(x_j)(x_l - \mathbb{E}(x_l)) = \mathbb{E}(x_j X_l) - \mathbb{E}(x_j)\mathbb{E}(x_l) \quad (15)$$

$$\text{Cov}(x_j, X_j) = \mathbb{E}(x_j^2) - (\mathbb{E}(x_j))^2 = \mathbb{V}(x_j) \quad (16)$$

Probability density function

The multivariate Gaussian/normal probability density function (PDF) is given by:

$$f(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp \left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right) \quad (17)$$

LDA with multiple predictors

LDA with multiple predictors

- Assume that observations in c th class are drawn from multivariate Gaussian $\mathcal{N}(\mu_c, \Sigma)$ with a common covariance matrix

LDA with multiple predictors

- Assume that observations in c th class are drawn from multivariate Gaussian $\mathcal{N}(\mu_c, \Sigma)$ with a common covariance matrix
- Estimate distribution parameters $\hat{\mu}_c$, $\hat{\Sigma}$ and priors $\hat{\pi}_c$:

LDA with multiple predictors

- Assume that observations in c th class are drawn from multivariate Gaussian $\mathcal{N}(\mu_c, \Sigma)$ with a common covariance matrix
- Estimate distribution parameters $\hat{\mu}_c$, $\hat{\Sigma}$ and priors $\hat{\pi}_c$:

LDA with multiple predictors

- Assume that observations in c th class are drawn from multivariate Gaussian $\mathcal{N}(\mu_c, \Sigma)$ with a common covariance matrix
- Estimate distribution parameters $\hat{\mu}_c$, $\hat{\Sigma}$ and priors $\hat{\pi}_c$:

$$\hat{\Sigma} = \frac{1}{n - C} \sum_{c=1}^C \sum_{i:y_i=c} (\mathbf{x}_i - \hat{\mu}_c)(\mathbf{x}_i - \hat{\mu}_c)^T \quad (18)$$

- Compute the linear discriminant functions:

LDA with multiple predictors

- Assume that observations in c th class are drawn from multivariate Gaussian $\mathcal{N}(\mu_c, \Sigma)$ with a common covariance matrix
- Estimate distribution parameters $\hat{\mu}_c$, $\hat{\Sigma}$ and priors $\hat{\pi}_c$:

$$\hat{\Sigma} = \frac{1}{n - C} \sum_{c=1}^C \sum_{i:y_i=c} (\mathbf{x}_i - \hat{\mu}_c)(\mathbf{x}_i - \hat{\mu}_c)^T \quad (18)$$

- Compute the linear discriminant functions:

LDA with multiple predictors

- Assume that observations in c th class are drawn from multivariate Gaussian $\mathcal{N}(\mu_c, \Sigma)$ with a common covariance matrix
- Estimate distribution parameters $\hat{\mu}_c$, $\hat{\Sigma}$ and priors $\hat{\pi}_c$:

$$\hat{\Sigma} = \frac{1}{n - C} \sum_{c=1}^C \sum_{i:y_i=c} (\mathbf{x}_i - \hat{\mu}_c)(\mathbf{x}_i - \hat{\mu}_c)^T \quad (18)$$

- Compute the linear discriminant functions:

$$\delta_c(\mathbf{x}) = \mathbf{x}^T \Sigma^{-1} \mu_c - \frac{1}{2} \mu_c^T \Sigma^{-1} \mu_c + \log \pi_c \quad (19)$$

- Assign each observation x_i to the class c which maximizes the linear discriminant functions:

LDA with multiple predictors

- Assume that observations in c th class are drawn from multivariate Gaussian $\mathcal{N}(\mu_c, \Sigma)$ with a common covariance matrix
- Estimate distribution parameters $\hat{\mu}_c$, $\hat{\Sigma}$ and priors $\hat{\pi}_c$:

$$\hat{\Sigma} = \frac{1}{n - C} \sum_{c=1}^C \sum_{i:y_i=c} (\mathbf{x}_i - \hat{\mu}_c)(\mathbf{x}_i - \hat{\mu}_c)^T \quad (18)$$

- Compute the linear discriminant functions:

$$\delta_c(\mathbf{x}) = \mathbf{x}^T \Sigma^{-1} \mu_c - \frac{1}{2} \mu_c^T \Sigma^{-1} \mu_c + \log \pi_c \quad (19)$$

- Assign each observation x_i to the class c which maximizes the linear discriminant functions:

LDA with multiple predictors

- Assume that observations in c th class are drawn from multivariate Gaussian $\mathcal{N}(\mu_c, \Sigma)$ with a common covariance matrix
- Estimate distribution parameters $\hat{\mu}_c$, $\hat{\Sigma}$ and priors $\hat{\pi}_c$:

$$\hat{\Sigma} = \frac{1}{n - C} \sum_{c=1}^C \sum_{i:y_i=c} (\mathbf{x}_i - \hat{\mu}_c)(\mathbf{x}_i - \hat{\mu}_c)^T \quad (18)$$

- Compute the linear discriminant functions:

$$\delta_c(\mathbf{x}) = \mathbf{x}^T \Sigma^{-1} \mu_c - \frac{1}{2} \mu_c^T \Sigma^{-1} \mu_c + \log \pi_c \quad (19)$$

- Assign each observation x_i to the class c which maximizes the linear discriminant functions:

$$\hat{y}_i = \arg \max_c \delta_c(\mathbf{x}) \quad (20)$$

LDA with multiple predictors: illustration

LDA with multiple predictors: illustration

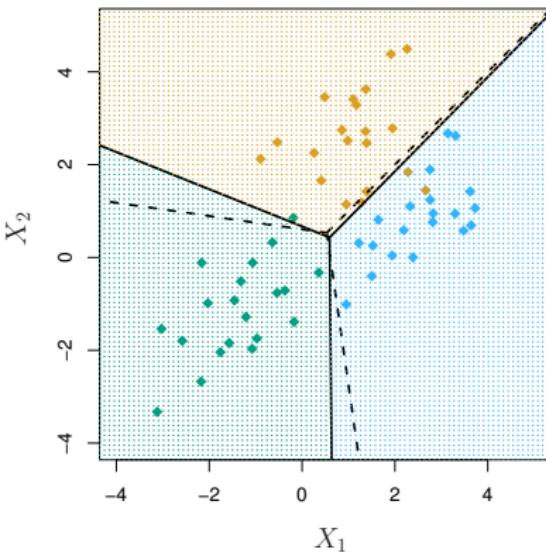
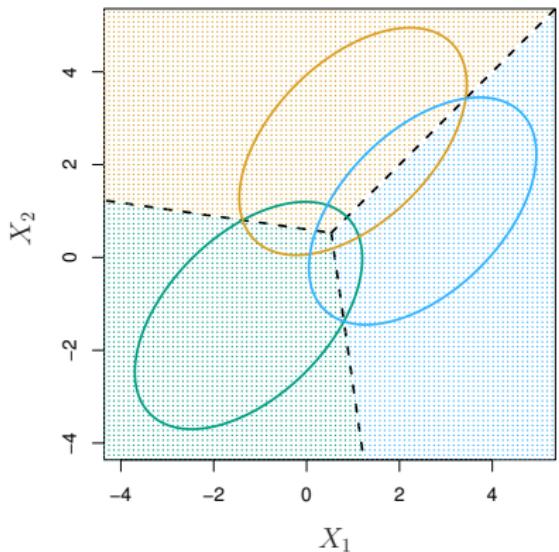


Figure: LDA for $p = 2$, $C = 3$. (L) Conditional class distributions of X (95% probability ellipses). (R) LDA decision boundaries (solid black) and Bayes decision boundaries (dashed lines).

Bayes' theorem for classification

Bayes' theorem for classification

Define:

$C = \text{number of classes}$

$\pi_c = \text{prior probability that random observation is in } c\text{th class}$

$p_c(\mathbf{x}) = \Pr(\mathbf{x} = \mathbf{x} | Y = c)$ (conditional probability distribution of X)

Bayes' theorem for classification

Define:

$C = \text{number of classes}$

$\pi_c = \text{prior probability that random observation is in } c\text{th class}$

$p_c(x) = \Pr(x = x | Y = c)$ (conditional probability distribution of X)

According to Bayes' theorem, the **posterior probability** that an observation is in class c given x is:

Bayes' theorem for classification

Define:

C = number of classes

π_c = prior probability that random observation is in c th class

$p_c(x) = \Pr(x = x | Y = c)$ (conditional probability distribution of X)

According to Bayes' theorem, the **posterior probability** that an observation is in class c given x is:

$$\Pr(Y = c | X = x) = \frac{\pi_c p_c(x)}{\sum_{l=1}^C \pi_l f_l(x)} \quad (21)$$

Bayes' theorem for classification

Define:

C = number of classes

π_c = prior probability that random observation is in c th class

$p_c(x) = \Pr(x = x | Y = c)$ (conditional probability distribution of X)

According to Bayes' theorem, the **posterior probability** that an observation is in class c given x is:

$$\Pr(Y = c | X = x) = \frac{\pi_c p_c(x)}{\sum_{l=1}^C \pi_l f_l(x)} \quad (21)$$

We can also express the posterior probability as:

Bayes' theorem for classification

Define:

C = number of classes

π_c = prior probability that random observation is in c th class

$p_c(\mathbf{x}) = \Pr(\mathbf{x} = \mathbf{x} | Y = c)$ (conditional probability distribution of X)

According to Bayes' theorem, the **posterior probability** that an observation is in class c given x is:

$$\Pr(Y = c | X = \mathbf{x}) = \frac{\pi_c p_c(\mathbf{x})}{\sum_{l=1}^C \pi_l f_l(\mathbf{x})} \quad (21)$$

We can also express the posterior probability as:

$$p_c(\mathbf{x}) = \Pr(Y = c | X = \mathbf{x}) \quad (22)$$

Bayes' theorem for classification (cont.)

Steps:

- ① Determine the probability density function of X : $p_c(x)$

Bayes' theorem for classification (cont.)

Steps:

- ① Determine the probability density function of X : $p_c(x)$
- ② Determine the prior probability π_c

Bayes' theorem for classification (cont.)

Steps:

- ① Determine the probability density function of X : $p_c(x)$
- ② Determine the prior probability π_c
- ③ Compute the posterior probability $p_c(x)$ using Bayes' theorem

Bayes' theorem for classification (cont.)

Steps:

- ① Determine the probability density function of X : $p_c(\mathbf{x})$
- ② Determine the prior probability π_c
- ③ Compute the posterior probability $p_c(\mathbf{x})$ using Bayes' theorem
- ④ Classify the observation based on the maximum probability:

$$\hat{y}_i = c^* = \arg \max_c p_c(\mathbf{x}_i) \quad (23)$$

Bayes' theorem for classification (cont.)

Steps:

- ① Determine the probability density function of X : $p_c(\mathbf{x})$
- ② Determine the prior probability π_c
- ③ Compute the posterior probability $p_c(\mathbf{x})$ using Bayes' theorem
- ④ Classify the observation based on the maximum probability:

$$\hat{y}_i = c^* = \arg \max_c p_c(\mathbf{x}_i) \quad (23)$$

Bayes' theorem for classification (cont.)

Steps:

- ① Determine the probability density function of X : $p_c(x)$
- ② Determine the prior probability π_c
- ③ Compute the posterior probability $p_c(x)$ using Bayes' theorem
- ④ Classify the observation based on the maximum probability:

$$\hat{y}_i = c^* = \arg \max_c p_c(x_i) \quad (23)$$

Eq. (23) is called the **decision rule**.

Assumptions

Assumption 1

X is normally distributed:

$$p_c(\mathbf{x}) = \frac{1}{\sqrt{2\pi}\sigma_c} \exp\left(-\frac{1}{2\sigma_c^2}(\mathbf{x} - \boldsymbol{\mu}_c)^2\right) \quad (24)$$

where $\boldsymbol{\mu}_c$ and σ_c^2 are the mean and variance of the observations in the c th class.

Assumptions

Assumption 1

X is normally distributed:

$$p_c(\mathbf{x}) = \frac{1}{\sqrt{2\pi}\sigma_c} \exp\left(-\frac{1}{2\sigma_c^2}(\mathbf{x} - \boldsymbol{\mu}_c)^2\right) \quad (24)$$

where $\boldsymbol{\mu}_c$ and σ_c^2 are the mean and variance of the observations in the c th class.

Assumption 2

There is a common variance across all C classes:

Assumptions

Assumption 1

X is normally distributed:

$$p_c(\mathbf{x}) = \frac{1}{\sqrt{2\pi}\sigma_c} \exp\left(-\frac{1}{2\sigma_c^2}(\mathbf{x} - \boldsymbol{\mu}_c)^2\right) \quad (24)$$

where $\boldsymbol{\mu}_c$ and σ_c^2 are the mean and variance of the observations in the c th class.

Assumption 2

There is a common variance across all C classes:

$$\sigma_1^2 = \dots =$$

Assumptions

Assumption 1

X is normally distributed:

$$p_c(\mathbf{x}) = \frac{1}{\sqrt{2\pi}\sigma_c} \exp\left(-\frac{1}{2\sigma_c^2}(\mathbf{x} - \boldsymbol{\mu}_c)^2\right) \quad (24)$$

where $\boldsymbol{\mu}_c$ and σ_c^2 are the mean and variance of the observations in the c th class.

Assumption 2

There is a common variance across all C classes:

$$\sigma_1^2 = \dots = \sigma_C^2 \quad (25)$$

Bayes classifier

Bayes classifier

Given the assumptions of normally distributed X and constant variance, then the **posterior probability** according to Bayes is:

Bayes classifier

Given the assumptions of normally distributed X and constant variance, then the **posterior probability** according to Bayes is:

$$p_c(\mathbf{x}) = \frac{\pi_c \frac{1}{\sqrt{2\pi}\sigma_c} \exp\left(-\frac{1}{2\sigma_c^2}(\mathbf{x} - \boldsymbol{\mu}_c)^2\right)}{\sum_{l=1}^C \pi_l \frac{1}{\sqrt{2\pi}\sigma_l} \exp\left(-\frac{1}{2\sigma_l^2}(\mathbf{x} - \boldsymbol{\mu}_l)^2\right)} \quad (26)$$

Bayes classifier

Given the assumptions of normally distributed X and constant variance, then the **posterior probability** according to Bayes is:

$$p_c(\mathbf{x}) = \frac{\pi_c \frac{1}{\sqrt{2\pi}\sigma_c} \exp\left(-\frac{1}{2\sigma_c^2}(\mathbf{x} - \boldsymbol{\mu}_c)^2\right)}{\sum_{l=1}^C \pi_l \frac{1}{\sqrt{2\pi}\sigma_l} \exp\left(-\frac{1}{2\sigma_l^2}(\mathbf{x} - \boldsymbol{\mu}_l)^2\right)} \quad (26)$$

The Bayes classifier assigns an observation to the class for which $p_c(\mathbf{x})$ is the largest.

Bayes classifier

Given the assumptions of normally distributed X and constant variance, then the **posterior probability** according to Bayes is:

$$p_c(\mathbf{x}) = \frac{\pi_c \frac{1}{\sqrt{2\pi}\sigma_c} \exp\left(-\frac{1}{2\sigma_c^2}(\mathbf{x} - \boldsymbol{\mu}_c)^2\right)}{\sum_{l=1}^C \pi_l \frac{1}{\sqrt{2\pi}\sigma_l} \exp\left(-\frac{1}{2\sigma_l^2}(\mathbf{x} - \boldsymbol{\mu}_l)^2\right)} \quad (26)$$

The Bayes classifier assigns an observation to the class for which $p_c(\mathbf{x})$ is the largest.

For the i th observation, this assignment can be written as:

Bayes classifier

Given the assumptions of normally distributed X and constant variance, then the **posterior probability** according to Bayes is:

$$p_c(\mathbf{x}) = \frac{\pi_c \frac{1}{\sqrt{2\pi}\sigma_c} \exp\left(-\frac{1}{2\sigma_c^2}(\mathbf{x} - \boldsymbol{\mu}_c)^2\right)}{\sum_{l=1}^C \pi_l \frac{1}{\sqrt{2\pi}\sigma_l} \exp\left(-\frac{1}{2\sigma_l^2}(\mathbf{x} - \boldsymbol{\mu}_l)^2\right)} \quad (26)$$

The Bayes classifier assigns an observation to the class for which $p_c(\mathbf{x})$ is the largest.

For the i th observation, this assignment can be written as:

$$y_i = c^* = \arg \max_c p_c(\mathbf{x}) =$$

Bayes classifier

Given the assumptions of normally distributed X and constant variance, then the **posterior probability** according to Bayes is:

$$p_c(\mathbf{x}) = \frac{\pi_c \frac{1}{\sqrt{2\pi}\sigma_c} \exp\left(-\frac{1}{2\sigma_c^2}(\mathbf{x} - \boldsymbol{\mu}_c)^2\right)}{\sum_{l=1}^C \pi_l \frac{1}{\sqrt{2\pi}\sigma_l} \exp\left(-\frac{1}{2\sigma_l^2}(\mathbf{x} - \boldsymbol{\mu}_l)^2\right)} \quad (26)$$

The Bayes classifier assigns an observation to the class for which $p_c(\mathbf{x})$ is the largest.

For the i th observation, this assignment can be written as:

$$y_i = c^* = \arg \max_c p_c(\mathbf{x}) = \arg \max_c \log(p_c(\mathbf{x})) =$$

Bayes classifier

Given the assumptions of normally distributed X and constant variance, then the **posterior probability** according to Bayes is:

$$p_c(\mathbf{x}) = \frac{\pi_c \frac{1}{\sqrt{2\pi}\sigma_c} \exp\left(-\frac{1}{2\sigma_c^2}(\mathbf{x} - \boldsymbol{\mu}_c)^2\right)}{\sum_{l=1}^C \pi_l \frac{1}{\sqrt{2\pi}\sigma_l} \exp\left(-\frac{1}{2\sigma_l^2}(\mathbf{x} - \boldsymbol{\mu}_l)^2\right)} \quad (26)$$

The Bayes classifier assigns an observation to the class for which $p_c(\mathbf{x})$ is the largest.

For the i th observation, this assignment can be written as:

$$y_i = c^* = \arg \max_c p_c(\mathbf{x}) = \arg \max_c \log(p_c(\mathbf{x})) = \arg \max_c \delta_c(\mathbf{x}) \quad (27)$$

The term $\delta_c(\mathbf{x})$ denotes the **linear discriminant functions**.

Linear discriminant functions

Linear discriminant functions

To derive, take the log of the posterior probability:

Linear discriminant functions

To derive, take the log of the posterior probability:

$$\begin{aligned}\log(p_c(\mathbf{x})) &= \log\left(\frac{\pi_c}{\sqrt{2\pi}\sigma_c}\right) - \frac{(\mathbf{x} - \boldsymbol{\mu}_c)^2}{2\sigma_c^2} - \sum_{l=1}^C \left[\log\left(\frac{\pi_l}{\sqrt{2\pi}\sigma_l}\right) - \frac{(\mathbf{x} - \boldsymbol{\mu}_l)^2}{2\sigma_l^2} \right] \\ &= \log\left(\frac{\pi_c}{\sqrt{2\pi}\sigma}\right) - \frac{(\mathbf{x} - \boldsymbol{\mu}_c)^2}{2\sigma^2} - \sum_{l=1}^C \left[\log\left(\frac{\pi_l}{\sqrt{2\pi}\sigma}\right) - \frac{(\mathbf{x} - \boldsymbol{\mu}_l)^2}{2\sigma^2} \right] \\ &= \log(\pi_c) - \log\left(\sqrt{2\pi}\sigma\right) - \frac{\mathbf{x}^2}{2\sigma^2} + \frac{\mathbf{x}\boldsymbol{\mu}_c}{2\sigma^2} - \frac{\boldsymbol{\mu}_c^2}{2\sigma^2} \\ &\quad - \sum_{l=1}^C \left[\log\left(\frac{\pi_l}{\sqrt{2\pi}\sigma}\right) - \frac{(\mathbf{x} - \boldsymbol{\mu}_l)^2}{2\sigma^2} \right]\end{aligned}$$

To find the maximizing value of $\log(p_c(\mathbf{x}))$, we discard the terms that are constant in c and thus obtain the **linear discriminant functions**:

$$\delta_c(\mathbf{x}) = \log(\pi_c) + \mathbf{x} \frac{\boldsymbol{\mu}_c}{\sigma^2} - \frac{\boldsymbol{\mu}_c^2}{2\sigma^2} \tag{28}$$

Linear discriminant functions (cont.)

Linear discriminant functions (cont.)

Given C classes each with priors π_c , means μ_c and common variance σ , we express the discriminant function as:

Linear discriminant functions (cont.)

Given C classes each with priors π_c , means μ_c and common variance σ^2 , we express the discriminant function as:

$$\delta_1(\mathbf{x}) = \log(\pi_1) + x \frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2}$$

$$\delta_2(\mathbf{x}) = \log(\pi_2) + x \frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2}$$

$$\vdots \qquad \vdots$$

$$\delta_C(\mathbf{x}) = \log(\pi_C) + x \frac{\mu_c}{\sigma^2} - \frac{\mu_c^2}{2\sigma^2}$$

For an observation with $X = x$, the *decision rule* assigns Y to the class c for which $\delta_c(\mathbf{x})$ is the greatest.

Linear discriminant functions (cont.)

Given C classes each with priors π_c , means μ_c and common variance σ^2 , we express the discriminant function as:

$$\delta_1(\mathbf{x}) = \log(\pi_1) + \mathbf{x} \frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2}$$

⋮ ⋮

$$\delta_C(\mathbf{x}) = \log(\pi_C) + \mathbf{x} \frac{\mu_C}{\sigma^2} - \frac{\mu_C^2}{2\sigma^2}$$

For an observation with $X = x$, the *decision rule* assigns Y to the class c for which $\delta_c(\mathbf{x})$ is the greatest.

Linear discriminant functions (cont.)

Given C classes each with priors π_c , means μ_c and common variance σ^2 , we express the discriminant function as:

$$\delta_1(\mathbf{x}) = \log(\pi_1) + \mathbf{x} \frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2}$$

$$\delta_C(\mathbf{x}) = \log(\pi_C) + \mathbf{x} \frac{\mu_C}{\sigma^2} - \frac{\mu_C^2}{2\sigma^2}$$

For an observation with $X = x$, the *decision rule* assigns Y to the class c for which $\delta_c(\mathbf{x})$ is the greatest.

Linear discriminant functions (cont.)

Given C classes each with priors π_c , means μ_c and common variance σ^2 , we express the discriminant function as:

$$\delta_1(\mathbf{x}) = \log(\pi_1) + x \frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2}$$

$$\delta_2(\mathbf{x}) = \log(\pi_2) + x \frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2}$$

 \vdots \vdots

$$\delta_C(\mathbf{x}) = \log(\pi_C) + x \frac{\mu_c}{\sigma^2} - \frac{\mu_c^2}{2\sigma^2}$$

Linear discriminant functions (cont.)

Given C classes each with priors π_c , means μ_c and common variance σ^2 , we express the discriminant function as:

$$\delta_1(\mathbf{x}) = \log(\pi_1) + x \frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2}$$

$$\delta_2(\mathbf{x}) = \log(\pi_2) + x \frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2}$$

$$\vdots \qquad \vdots$$

$$\delta_C(\mathbf{x}) = \log(\pi_C) + x \frac{\mu_c}{\sigma^2} - \frac{\mu_c^2}{2\sigma^2}$$

For an observation with $X = x$, the *decision rule* assigns Y to the class c for which $\delta_c(\mathbf{x})$ is the greatest.

Bayes decision boundary: binary case

Bayes decision boundary: binary case

In the binary case, $C = 2$.

Bayes decision boundary: binary case

In the binary case, $C = 2$. The discriminant functions are:

Bayes decision boundary: binary case

In the binary case, $C = 2$. The discriminant functions are:

$$\delta_1(\mathbf{x}) = \log(\pi_1) + x \frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2}$$
$$\delta_2(\mathbf{x}) = \log(\pi_2) + x \frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2}$$

The decision rule assigns an observation to Class 1 if $\delta_1(\mathbf{x}) > \delta_2(\mathbf{x})$.

Now, consider the situation where the *priors are equal*:

$$\pi_1 = \pi_2$$

implying that an observation is *equally likely* to come from Class 1 or Class 2. The decision rule would then **assign to Class 1** if:

$$x \frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} > x \frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2}$$
$$\Rightarrow x(\mu_1 - \mu_2) > \frac{\mu_1^2 - \mu_2^2}{2}$$

Bayes decision boundary: binary case

In the binary case, $C = 2$. The discriminant functions are:

$$\delta_1(\mathbf{x}) = \log(\pi_1) + x \frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2}$$
$$\delta_2(\mathbf{x}) = \log(\pi_2) + x \frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2}$$

Bayes decision boundary: binary case

In the binary case, $C = 2$. The discriminant functions are:

$$\delta_1(\mathbf{x}) = \log(\pi_1) + x \frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2}$$
$$\delta_2(\mathbf{x}) = \log(\pi_2) + x \frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2}$$

The decision rule assigns an observation to Class 1 if $\delta_1(\mathbf{x}) > \delta_2(\mathbf{x})$.

Bayes decision boundary: binary case

In the binary case, $C = 2$. The discriminant functions are:

$$\delta_1(\mathbf{x}) = \log(\pi_1) + x \frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2}$$
$$\delta_2(\mathbf{x}) = \log(\pi_2) + x \frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2}$$

The decision rule assigns an observation to Class 1 if $\delta_1(\mathbf{x}) > \delta_2(\mathbf{x})$.

Now, consider the situation where the *priors are equal*:

Bayes decision boundary: binary case

In the binary case, $C = 2$. The discriminant functions are:

$$\delta_1(\mathbf{x}) = \log(\pi_1) + x \frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2}$$
$$\delta_2(\mathbf{x}) = \log(\pi_2) + x \frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2}$$

The decision rule assigns an observation to Class 1 if $\delta_1(\mathbf{x}) > \delta_2(\mathbf{x})$.

Now, consider the situation where the *priors are equal*:

$$\pi_1 = \pi_2$$

Bayes decision boundary: binary case

In the binary case, $C = 2$. The discriminant functions are:

$$\delta_1(\mathbf{x}) = \log(\pi_1) + x \frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2}$$
$$\delta_2(\mathbf{x}) = \log(\pi_2) + x \frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2}$$

The decision rule assigns an observation to Class 1 if $\delta_1(\mathbf{x}) > \delta_2(\mathbf{x})$.

Now, consider the situation where the *priors are equal*:

$$\pi_1 = \pi_2$$

implying that an observation is *equally likely* to come from Class 1 or Class 2.

Bayes decision boundary: binary case

In the binary case, $C = 2$. The discriminant functions are:

$$\delta_1(\mathbf{x}) = \log(\pi_1) + x \frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2}$$
$$\delta_2(\mathbf{x}) = \log(\pi_2) + x \frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2}$$

The decision rule assigns an observation to Class 1 if $\delta_1(\mathbf{x}) > \delta_2(\mathbf{x})$.

Now, consider the situation where the *priors are equal*:

$$\pi_1 = \pi_2$$

implying that an observation is *equally likely* to come from Class 1 or Class 2. The decision rule would then **assign to Class 1** if:

Bayes decision boundary: binary case

In the binary case, $C = 2$. The discriminant functions are:

$$\delta_1(\mathbf{x}) = \log(\pi_1) + x \frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2}$$
$$\delta_2(\mathbf{x}) = \log(\pi_2) + x \frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2}$$

The decision rule assigns an observation to Class 1 if $\delta_1(\mathbf{x}) > \delta_2(\mathbf{x})$.

Now, consider the situation where the *priors are equal*:

$$\pi_1 = \pi_2$$

implying that an observation is *equally likely* to come from Class 1 or Class 2. The decision rule would then **assign to Class 1** if:

$$x \frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} > x \frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2}$$
$$\Rightarrow x(\mu_1 - \mu_2) > \frac{\mu_1^2 - \mu_2^2}{2}$$

Bayes decision boundary: binary case

In the binary case, $C = 2$. The discriminant functions are:

$$\delta_1(\mathbf{x}) = \log(\pi_1) + x \frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2}$$
$$\delta_2(\mathbf{x}) = \log(\pi_2) + x \frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2}$$

The decision rule assigns an observation to Class 1 if $\delta_1(\mathbf{x}) > \delta_2(\mathbf{x})$.

Now, consider the situation where the *priors are equal*:

$$\pi_1 = \pi_2$$

implying that an observation is *equally likely* to come from Class 1 or Class 2. The decision rule would then **assign to Class 1** if:

$$x \frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} > x \frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2}$$
$$\Rightarrow x(\mu_1 - \mu_2) > \frac{\mu_1^2 - \mu_2^2}{2}$$

Bayes decision boundary (cont.)

At the Bayes decision boundary, the $\delta_1 = \delta_2$, thus:

Bayes decision boundary (cont.)

At the Bayes decision boundary, the $\delta_1 = \delta_2$, thus:

$$x(\mu_1 - \mu_2) = \frac{\mu_1^2 - \mu_2^2}{2} = \frac{(\mu_1 - \mu_2)(\mu_1 + \mu_2)}{2}$$
$$x = \frac{(\mu_1 + \mu_2)}{2}$$

Thus, given a univariate predictor X whose two classes have equal priors $\pi_1 = \pi_2$, the decision boundary lies midway between the means μ_1 and μ_2 .

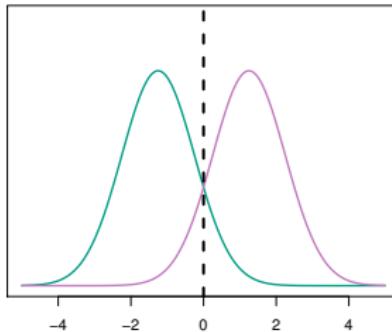


Figure: Bayes decision boundary (dashed line) shown for the priors of the distributions of 2 classes: Class 1 (green) and Class 2 (purple)

Bayes decision boundary (cont.)

At the Bayes decision boundary, the $\delta_1 = \delta_2$, thus:

$$\begin{aligned}x(\mu_1 - \mu_2) &= \frac{\mu_1^2 - \mu_2^2}{2} = \frac{(\mu_1 - \mu_2)(\mu_1 + \mu_2)}{2} \\x &= \frac{(\mu_1 + \mu_2)}{2}\end{aligned}$$

Bayes decision boundary (cont.)

At the Bayes decision boundary, the $\delta_1 = \delta_2$, thus:

$$\begin{aligned}x(\mu_1 - \mu_2) &= \frac{\mu_1^2 - \mu_2^2}{2} = \frac{(\mu_1 - \mu_2)(\mu_1 + \mu_2)}{2} \\x &= \frac{(\mu_1 + \mu_2)}{2}\end{aligned}$$

Thus, given a univariate predictor X whose two classes have equal priors $\pi_1 = \pi_2$, the decision boundary lies midway between the means μ_1 and μ_2 .

Bayes decision boundary (cont.)

At the Bayes decision boundary, the $\delta_1 = \delta_2$, thus:

$$x(\mu_1 - \mu_2) = \frac{\mu_1^2 - \mu_2^2}{2} = \frac{(\mu_1 - \mu_2)(\mu_1 + \mu_2)}{2}$$
$$x = \frac{(\mu_1 + \mu_2)}{2}$$

Thus, given a univariate predictor X whose two classes have equal priors $\pi_1 = \pi_2$, the decision boundary lies midway between the means μ_1 and μ_2 .

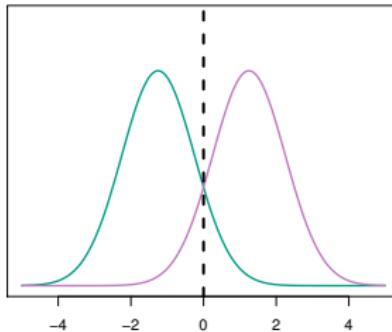


Figure: Bayes decision boundary (dashed line) shown for the priors of the distributions of 2 classes: Class 1 (green) and Class 2 (purple)

Quadratic discriminant analysis (QDA)

- A key assumption in LDA is that all classes share a common covariance structure: $\Sigma_1 = \dots = \Sigma_C = \Sigma$.

Quadratic discriminant analysis (QDA)

- A key assumption in LDA is that all classes share a common covariance structure: $\Sigma_1 = \dots = \Sigma_C = \Sigma$.
- If we discard this assumption, then:

Quadratic discriminant analysis (QDA)

- A key assumption in LDA is that all classes share a common covariance structure: $\Sigma_1 = \dots = \Sigma_C = \Sigma$.
- If we discard this assumption, then:

Quadratic discriminant analysis (QDA)

- A key assumption in LDA is that all classes share a common covariance structure: $\Sigma_1 = \dots = \Sigma_C = \Sigma$.
- If we discard this assumption, then:

$$X \sim \mathcal{N}(\mu_c, \Sigma_{\textcolor{red}{c}}) \quad (29)$$

Quadratic discriminant analysis (QDA)

- A key assumption in LDA is that all classes share a common covariance structure: $\Sigma_1 = \dots = \Sigma_C = \Sigma$.
- If we discard this assumption, then:

$$X \sim \mathcal{N}(\mu_c, \Sigma_c) \quad (29)$$

and we can no longer ignore the $\hat{\Sigma}_c$ terms in posterior probabilities.

- This results in discriminant functions that are **quadratic** in x :

Quadratic discriminant analysis (QDA)

- A key assumption in LDA is that all classes share a common covariance structure: $\Sigma_1 = \dots = \Sigma_C = \Sigma$.
- If we discard this assumption, then:

$$X \sim \mathcal{N}(\mu_c, \Sigma_c) \quad (29)$$

and we can no longer ignore the $\hat{\Sigma}_c$ terms in posterior probabilities.

- This results in discriminant functions that are **quadratic** in x :

Quadratic discriminant analysis (QDA)

- A key assumption in LDA is that all classes share a common covariance structure: $\Sigma_1 = \dots = \Sigma_C = \Sigma$.
- If we discard this assumption, then:

$$X \sim \mathcal{N}(\mu_c, \Sigma_c) \quad (29)$$

and we can no longer ignore the $\hat{\Sigma}_c$ terms in posterior probabilities.

- This results in discriminant functions that are **quadratic** in x :

$$\delta_c(x) = -\frac{1}{2}(x - \hat{\mu}_c)^T \Sigma_c^{-1} (x - \hat{\mu}_c) - \frac{1}{2} \log |\Sigma_c| + \log \pi_c \quad (30)$$

Quadratic discriminant analysis (QDA)

- A key assumption in LDA is that all classes share a common covariance structure: $\Sigma_1 = \dots = \Sigma_C = \Sigma$.
- If we discard this assumption, then:

$$X \sim \mathcal{N}(\mu_c, \Sigma_c) \quad (29)$$

and we can no longer ignore the $\hat{\Sigma}_c$ terms in posterior probabilities.

- This results in discriminant functions that are **quadratic** in x :

$$\begin{aligned}\delta_c(x) &= -\frac{1}{2}(x - \hat{\mu}_c)^T \Sigma_c^{-1} (x - \hat{\mu}_c) - \frac{1}{2} \log |\Sigma_c| + \log \pi_c \\ &= -\frac{1}{2} x^T \Sigma_c^{-1} x + x^T \Sigma_c^{-1} \mu_c - \frac{1}{2} \hat{\mu}_c^T \Sigma_c^{-1} \hat{\mu}_c - \frac{1}{2} \log |\Sigma_c| + \log \pi_c\end{aligned} \quad (30)$$

- Under this assumption of class-specific covariance, we perform **quadratic discriminant analysis**.

QDA considerations: bias-variance trade-off

QDA considerations: bias-variance trade-off

Number of parameters:

QDA considerations: bias-variance trade-off

Number of parameters:

- To find $\hat{\Sigma}$ in LDA, $D(D + 1)/2$ parameters must be estimated

QDA considerations: bias-variance trade-off

Number of parameters:

- To find $\hat{\Sigma}$ in LDA, $D(D + 1)/2$ parameters must be estimated
- In QDA, since there are C covariance matrices, $CD(D + 1)/2$ must be estimated

QDA considerations: bias-variance trade-off

Number of parameters:

- To find $\hat{\Sigma}$ in LDA, $D(D + 1)/2$ parameters must be estimated
- In QDA, since there are C covariance matrices, $CD(D + 1)/2$ must be estimated

QDA considerations: bias-variance trade-off

Number of parameters:

- To find $\hat{\Sigma}$ in LDA, $D(D + 1)/2$ parameters must be estimated
- In QDA, since there are C covariance matrices, $CD(D + 1)/2$ must be estimated

Bias-variance:

QDA considerations: bias-variance trade-off

Number of parameters:

- To find $\hat{\Sigma}$ in LDA, $D(D + 1)/2$ parameters must be estimated
- In QDA, since there are C covariance matrices, $CD(D + 1)/2$ must be estimated

Bias-variance:

- Thus QDA is more flexible (lower bias, but potentially higher variance)

QDA considerations: bias-variance trade-off

Number of parameters:

- To find $\hat{\Sigma}$ in LDA, $D(D + 1)/2$ parameters must be estimated
- In QDA, since there are C covariance matrices, $CD(D + 1)/2$ must be estimated

Bias-variance:

- Thus QDA is more flexible (lower bias, but potentially higher variance)
- LDA might be more stable (lower variance, but potentially higher bias) if the constant Σ assumption does not hold for the data.

QDA considerations: bias-variance trade-off

Number of parameters:

- To find $\hat{\Sigma}$ in LDA, $D(D + 1)/2$ parameters must be estimated
- In QDA, since there are C covariance matrices, $CD(D + 1)/2$ must be estimated

Bias-variance:

- Thus QDA is more flexible (lower bias, but potentially higher variance)
- LDA might be more stable (lower variance, but potentially higher bias) if the constant Σ assumption does not hold for the data.
- Generally, with fewer observations, LDA is preferred

Introduction
ooooo

LDA model
oooooooo

LD derivation
oooooooo

QDA
ooo○

Summary
○

Appendix: GLMs
ooooo

QDA vs. LDA

QDA vs. LDA

QDA produces a quadratic decision boundary which performs better in classifying the observations if the covariance matrices are different for each class.

QDA vs. LDA

QDA produces a quadratic decision boundary which performs better in classifying the observations if the covariance matrices are different for each class.

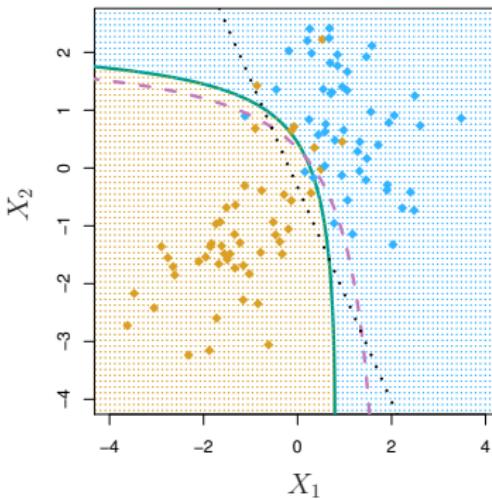
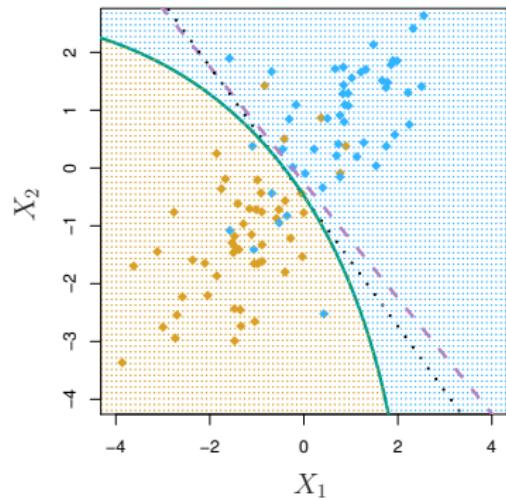


Figure: (L): shared covariance across classes (linear boundary). (R): different covariance in each class (quadratic boundary). Bayes (purple dashed); LDA (black dotted); QDA (green solid)

Approximating a quadratic decision boundary

Approximating a quadratic decision boundary

QDA can be approximated by LDA by including second-order terms in the predictor space, i.e.

Approximating a quadratic decision boundary

QDA can be approximated by LDA by including second-order terms in the predictor space, i.e.

$$(X_1, X_2) \rightarrow (X_1, X_2, X_1 X_2, X_1^2, X_2^2)$$

Approximating a quadratic decision boundary

QDA can be approximated by LDA by including second-order terms in the predictor space, i.e.

$$(X_1, X_2) \rightarrow (X_1, X_2, X_1 X_2, X_1^2, X_2^2)$$

The LDA approximation is may be less accurate but more stable, as fewer parameters are required.

Approximating a quadratic decision boundary

QDA can be approximated by LDA by including second-order terms in the predictor space, i.e.

$$(X_1, X_2) \rightarrow (X_1, X_2, X_1 X_2, X_1^2, X_2^2)$$

The LDA approximation is may be less accurate but more stable, as fewer parameters are required.

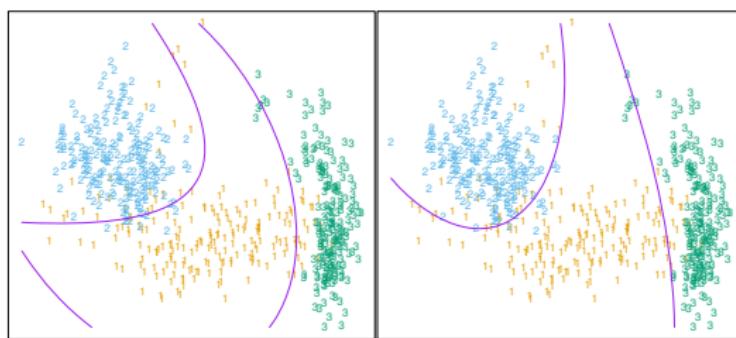


Figure: (L) LDA approximation of decision boundary. (R) QDA decision boundary.
 $D = 2, C = 3$.

Outlook

Outlook

- Next lecture: L2b: Logistic regression

Outlook

- Next lecture: L2b: Logistic regression
- Reading for today's lecture:

Outlook

- Next lecture: L2b: Logistic regression
- Reading for today's lecture: **PMLI** 9.1-2; **ESL** 4.3

Outlook

- Next lecture: L2b: Logistic regression
- Reading for today's lecture: **PMLI** 9.1-2; **ESL** 4.3
- Optional: Naive Bayes classifier (NBC) **PMLI** 9.3

The generalized linear model (GLM)

The generalized linear model (GLM)

- Conventional linear models have the form:

$$y_i \sim N(\mathbf{x}_i^T \boldsymbol{\beta}, \sigma^2) \quad (32)$$

The generalized linear model (GLM)

- Conventional linear models have the form:

$$y_i \sim N(\mathbf{x}_i^T \boldsymbol{\beta}, \sigma^2) \quad (32)$$

where

The generalized linear model (GLM)

- Conventional linear models have the form:

$$y_i \sim N(\mathbf{x}_i^T \boldsymbol{\beta}, \sigma^2) \quad (32)$$

where

- y_i is a continuous response
- \mathbf{x}_i is a vector of quantitative and/or qualitative explanatory variables

The generalized linear model (GLM)

- Conventional linear models have the form:

$$y_i \sim N(\mathbf{x}_i^T \boldsymbol{\beta}, \sigma^2) \quad (32)$$

where

- y_i is a continuous response
- \mathbf{x}_i is a vector of quantitative and/or qualitative explanatory variables
- Generalized linear models (GLMs) were introduced to extend this framework to allow y_i to be modeled by other exponential family distributions besides the normal/Gaussian, e.g.
 - exponential
 - binomial/multinomial (with fixed number of trials)
 - Poisson

The generalized linear model (GLM)

- Conventional linear models have the form:

$$y_i \sim N(\mathbf{x}_i^T \boldsymbol{\beta}, \sigma^2) \quad (32)$$

where

- y_i is a continuous response
- \mathbf{x}_i is a vector of quantitative and/or qualitative explanatory variables
- Generalized linear models (GLMs) were introduced to extend this framework to allow y_i to be modeled by other exponential family distributions besides the normal/Gaussian, e.g.
 - exponential
 - binomial/multinomial (with fixed number of trials)
 - Poisson
- In the GLM framework:

The generalized linear model (GLM)

- Conventional linear models have the form:

$$y_i \sim N(\mathbf{x}_i^T \boldsymbol{\beta}, \sigma^2) \quad (32)$$

where

- y_i is a continuous response
- \mathbf{x}_i is a vector of quantitative and/or qualitative explanatory variables
- Generalized linear models (GLMs) were introduced to extend this framework to allow y_i to be modeled by other exponential family distributions besides the normal/Gaussian, e.g.
 - exponential
 - binomial/multinomial (with fixed number of trials)
 - Poisson
- In the GLM framework:
 - The mean of y_i is given by μ_i
 - μ_i can be specified by a nonlinear function of $\mathbf{x}_i^T \boldsymbol{\beta}$
 - Note that the simple linear regression is a special case of GLM in which $\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$ and y_i follows a Gaussian distribution

The generalized linear model (GLM)

- Conventional linear models have the form:

$$y_i \sim N(\mathbf{x}_i^T \boldsymbol{\beta}, \sigma^2) \quad (32)$$

where

- y_i is a continuous response
- \mathbf{x}_i is a vector of quantitative and/or qualitative explanatory variables
- Generalized linear models (GLMs) were introduced to extend this framework to allow y_i to be modeled by other exponential family distributions besides the normal/Gaussian, e.g.
 - exponential
 - binomial/multinomial (with fixed number of trials)
 - Poisson
- In the GLM framework:
 - The mean of y_i is given by μ_i
 - μ_i can be specified by a nonlinear function of $\mathbf{x}_i^T \boldsymbol{\beta}$
 - Note that the simple linear regression is a special case of GLM in which $\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$ and y_i follows a Gaussian distribution

GLM components

GLM components

A GLM consists of three parts:

GLM components

A GLM consists of three parts:

- **Random component:** this is the probability distribution of the response variable

GLM components

A GLM consists of three parts:

- **Random component:** this is the probability distribution of the response variable
- **Systematic component:** specifies the explanatory variables within the linear combination of their coefficients ($\mathbf{X}\boldsymbol{\beta}$)

GLM components

A GLM consists of three parts:

- **Random component:** this is the probability distribution of the response variable
- **Systematic component:** specifies the explanatory variables within the linear combination of their coefficients ($\mathbf{X}\boldsymbol{\beta}$)
- **Link function $g(\mu)$:** defines the relationship between the random and systematic components:

GLM components

A GLM consists of three parts:

- **Random component:** this is the probability distribution of the response variable
- **Systematic component:** specifies the explanatory variables within the linear combination of their coefficients ($\mathbf{X}\boldsymbol{\beta}$)
- **Link function $g(\mu)$:** defines the relationship between the random and systematic components:
 - Simple linear regression (identity link function):

GLM components

A GLM consists of three parts:

- **Random component:** this is the probability distribution of the response variable
- **Systematic component:** specifies the explanatory variables within the linear combination of their coefficients ($\mathbf{X}\boldsymbol{\beta}$)
- **Link function $g(\mu)$:** defines the relationship between the random and systematic components:
 - Simple linear regression (identity link function):

$$g(\mu_i) = g(\mathbb{E}(y_i)) = \mathbf{x}_i^T \boldsymbol{\beta} \quad (33)$$

GLM components

A GLM consists of three parts:

- **Random component:** this is the probability distribution of the response variable
- **Systematic component:** specifies the explanatory variables within the linear combination of their coefficients ($\mathbf{X}\boldsymbol{\beta}$)
- **Link function $g(\mu)$:** defines the relationship between the random and systematic components:
 - Simple linear regression (identity link function):

$$g(\mu_i) = g(\mathbb{E}(y_i)) = \mathbf{x}_i^T \boldsymbol{\beta} \quad (33)$$

- Binary logistic regression (logit link function):

GLM components

A GLM consists of three parts:

- **Random component:** this is the probability distribution of the response variable
- **Systematic component:** specifies the explanatory variables within the linear combination of their coefficients ($\mathbf{X}\boldsymbol{\beta}$)
- **Link function $g(\mu)$:** defines the relationship between the random and systematic components:
 - Simple linear regression (identity link function):

$$g(\mu_i) = g(\mathbb{E}(y_i)) = \mathbf{x}_i^T \boldsymbol{\beta} \quad (33)$$

- Binary logistic regression (logit link function):

$$g(\mu_i) = g(p(\mathbf{x}_i)) =$$

GLM components

A GLM consists of three parts:

- **Random component:** this is the probability distribution of the response variable
- **Systematic component:** specifies the explanatory variables within the linear combination of their coefficients ($\mathbf{X}\boldsymbol{\beta}$)
- **Link function $g(\mu)$:** defines the relationship between the random and systematic components:
 - Simple linear regression (identity link function):

$$g(\mu_i) = g(\mathbb{E}(y_i)) = \mathbf{x}_i^T \boldsymbol{\beta} \quad (33)$$

- Binary logistic regression (logit link function):

$$g(\mu_i) = g(p(\mathbf{x}_i)) = \text{logit}(p(\mathbf{x}_i)) =$$

GLM components

A GLM consists of three parts:

- **Random component:** this is the probability distribution of the response variable
- **Systematic component:** specifies the explanatory variables within the linear combination of their coefficients ($\mathbf{X}\boldsymbol{\beta}$)
- **Link function $g(\mu)$:** defines the relationship between the random and systematic components:
 - Simple linear regression (identity link function):

$$g(\mu_i) = g(\mathbb{E}(y_i)) = \mathbf{x}_i^T \boldsymbol{\beta} \quad (33)$$

- Binary logistic regression (logit link function):

$$g(\mu_i) = g(p(\mathbf{x}_i)) = \text{logit}(p(\mathbf{x}_i)) = \ln \left(\frac{p(\mathbf{x}_i)}{1 - p(\mathbf{x}_i)} \right) =$$

GLM components

A GLM consists of three parts:

- **Random component:** this is the probability distribution of the response variable
- **Systematic component:** specifies the explanatory variables within the linear combination of their coefficients ($\mathbf{X}\boldsymbol{\beta}$)
- **Link function $g(\mu)$:** defines the relationship between the random and systematic components:
 - Simple linear regression (identity link function):

$$g(\mu_i) = g(\mathbb{E}(y_i)) = \mathbf{x}_i^T \boldsymbol{\beta} \quad (33)$$

- Binary logistic regression (logit link function):

$$g(\mu_i) = g(p(\mathbf{x}_i)) = \text{logit}(p(\mathbf{x}_i)) = \ln \left(\frac{p(\mathbf{x}_i)}{1 - p(\mathbf{x}_i)} \right) = \mathbf{x}_i^T \boldsymbol{\beta} \quad (34)$$

Assumptions of GLM

Assumptions of GLM

- The observations of the response variable y are i.i.d.

Assumptions of GLM

- The observations of the response variable y are i.i.d.
- Response variable y_i is typically exponentially distributed (not restricted to being normally distributed)

Assumptions of GLM

- The observations of the response variable y are i.i.d.
- Response variable y_i is typically exponentially distributed (not restricted to being normally distributed)
 - Implies that errors need not be normally distributed (but should be independent)
- Link function is linear with respect to the coefficients (β_j)

Assumptions of GLM

- The observations of the response variable y are i.i.d.
- Response variable y_i is typically exponentially distributed (not restricted to being normally distributed)
 - Implies that errors need not be normally distributed (but should be independent)
- Link function is linear with respect to the coefficients (β_j)
 - Relationship between response and explanatory variables does not have to be linear

Assumptions of GLM

- The observations of the response variable y are i.i.d.
- Response variable y_i is typically exponentially distributed (not restricted to being normally distributed)
 - Implies that errors need not be normally distributed (but should be independent)
- Link function is linear with respect to the coefficients (β_j)
 - Relationship between response and explanatory variables does not have to be linear
 - Explanatory variables can be nonlinear transformations of original values (as in simple linear regression)

Assumptions of GLM

- The observations of the response variable y are i.i.d.
- Response variable y_i is typically exponentially distributed (not restricted to being normally distributed)
 - Implies that errors need not be normally distributed (but should be independent)
- Link function is linear with respect to the coefficients (β_j)
 - Relationship between response and explanatory variables does not have to be linear
 - Explanatory variables can be nonlinear transformations of original values (as in simple linear regression)
- Variance may not homogeneous (i.e. homoscedasticity is not a requirement)

Assumptions of GLM

- The observations of the response variable y are i.i.d.
- Response variable y_i is typically exponentially distributed (not restricted to being normally distributed)
 - Implies that errors need not be normally distributed (but should be independent)
- Link function is linear with respect to the coefficients (β_j)
 - Relationship between response and explanatory variables does not have to be linear
 - Explanatory variables can be nonlinear transformations of original values (as in simple linear regression)
- Variance may not homogeneous (i.e. homoscedasticity is not a requirement)
- Parameters are estimated via MLE

Commonly used GLM models and their components

Commonly used GLM models and their components

Model	Random component	Link function
Linear regression	Gaussian	Identity: $g(\mu_i) = \mu_i = \beta^T x_i$
Binary logistic regression	Bernoulli	Logit: $g(\mu_i) = \ln\left(\frac{\mu_i}{1-\mu_i}\right)$
Probit regression	Bernoulli	Probit: $g(\mu_i) = \Phi^{-1}(\mu_i)$
Multinomial logit/logistic	Categorical	Multinomial logit: $g(\mu_{ic}) = \ln\left(\frac{\mu_{ic}}{\mu_{iC}}\right)$
Poisson regression	Poisson	Log: $g(\mu_i) = \ln(\mu_i)$

Commonly used GLM models and their components

Model	Random component	Link function
Linear regression	Gaussian	Identity: $g(\mu_i) = \mu_i = \beta^T x_i$
Binary logistic regression	Bernoulli	Logit: $g(\mu_i) = \ln\left(\frac{\mu_i}{1-\mu_i}\right)$
Probit regression	Bernoulli	Probit: $g(\mu_i) = \Phi^{-1}(\mu_i)$
Multinomial logit/logistic	Categorical	Multinomial logit: $g(\mu_{ic}) = \ln\left(\frac{\mu_{ic}}{\mu_{iC}}\right)$
Poisson regression	Poisson	Log: $g(\mu_i) = \ln(\mu_i)$

Note that in all cases, the link function always results in:

$$g(\mu_i) = \beta^T x_i \tag{35}$$

Commonly used GLM models and their components

Model	Random component	Link function
Linear regression	Gaussian	Identity: $g(\mu_i) = \mu_i = \beta^T x_i$
Binary logistic regression	Bernoulli	Logit: $g(\mu_i) = \ln\left(\frac{\mu_i}{1-\mu_i}\right)$
Probit regression	Bernoulli	Probit: $g(\mu_i) = \Phi^{-1}(\mu_i)$
Multinomial logit/logistic	Categorical	Multinomial logit: $g(\mu_{ic}) = \ln\left(\frac{\mu_{ic}}{\mu_{iC}}\right)$
Poisson regression	Poisson	Log: $g(\mu_i) = \ln(\mu_i)$

Note that in all cases, the link function always results in:

$$g(\mu_i) = \beta^T x_i \tag{35}$$

Its job is to “link” the response to the systematic component via a suitable transformation that results in a linear function of the β 's.

Further reading on GLMs

Further reading on GLMs

- German Rodriguez's lecture notes on GLMs:
<https://data.princeton.edu/wws509/notes/>
- Penn State: <https://online.stat.psu.edu/stat504/lesson/6/6.1>
(Including more on logistic and multinomial logistic)