Models for classification

Linear discriminant analysis

A quick recap

During the last lectures, we did...

- **Models for classification**.
  Introduce the classification problem
  Introduce logistic regression models

Today's goal

Today, we going to do...

- **Models for classification**: Linear discriminant analysis

⇝ Defining the terms: Review of probability theory fundamentals
⇝ Bayes' theorem for classification
⇝ Linear discriminant analysis for one predictor, $p = 1$
⇝ Linear discriminant analysis for $p > 1$

Reading list

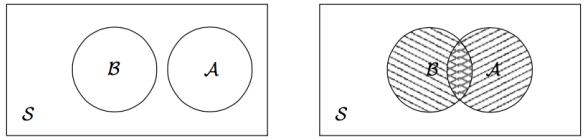
Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*, Springer (2017)

Review of probability theory fundamentals

To measure our degree of uncertainty about an experiment we assign a probability $P(A)$ to each event $A \subseteq S$, where S is the sample space of the experiment.

These probabilities must obey the following three axioms:

- **Axioms 1**: $P(S) = 1$.
- **Axioms 2**: For all $A \subseteq S$ it must hold that $P(A) \geq 0$.
- **Axioms 2**: If $A \subseteq S, B \subseteq S$ and $A \cap B = \varnothing$ then $P(A \cup B) = P(A) + P(B)$.



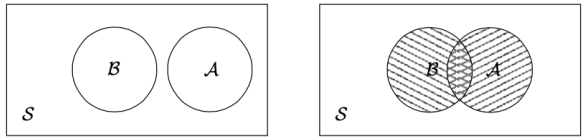
⇝ If two events A and B are **disjoint**, then the probability of the combined event is the sum of the probabilities for the two individual events.

Review of probability theory fundamentals

To measure our degree of uncertainty about an experiment we assign a probability $P(A)$ to each event $A \subseteq S$, where S is the sample space of the experiment.

These probabilities must obey the following three axioms:

- **Axioms 1**: $P(S) = 1$.
- **Axioms 2**: For all $A \subseteq S$ it must hold that $P(A) \geq 0$.
- **Axioms 2**: If $A \subseteq S, B \subseteq S$ and $A \cap B = \varnothing$ then $P(A \cup B) = P(A) + P(B)$.



$\rightsquigarrow$ If two events A and B are **not disjoint**, then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Review of probability theory fundamentals

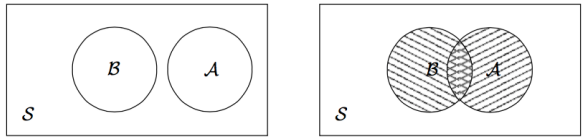
Conditional probabilities

The conditional probabilities are generally statements of the following kind:

- Given the event B, the probability of the event A is p. We write $P(A|B) = p$.

For two events A and B, with $P(B) > 0$, the **conditional probability** for A and B is:

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



$\rightsquigarrow$ If we know the event B, then all possible outcomes are elements of B, and the outcomes for which A can be true are $A \cap B$.
$\rightsquigarrow$ So, we look for the probability assignment for $A \cap B$ given that we know B.

Review of probability theory fundamentals

Probability calculus

The fundament rule of calculus: $P(A|B)P(B) = P(A \cap B)$

$\rightsquigarrow$ The fundamental rule tells us how to calculate the probability of seeing both A and B when we know the probability of A given B and the probability of B.

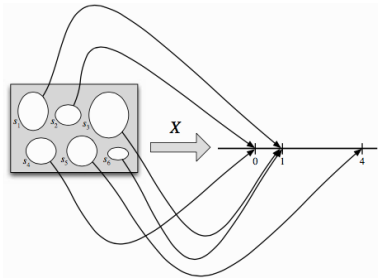
Bayes' theorem: $P(A|B) = \dfrac{P(B|A)P(A)}{P(B)}$

$\rightsquigarrow$ Bayes' theorem gives us a method for updating our beliefs about an event A given that we get information about another event B.
$\rightsquigarrow$ $P(A)$ is usually called **prior probability** of A.
$\rightsquigarrow$ $P(A|B)$ is called the **posterior distribution** of A given B.

Review of probability theory fundamentals

Random variables

Let S be a sample space. A random variable is a function assigning a real number to every possible outcome of an experiment on S; $X : S \to \mathbb{R}$.

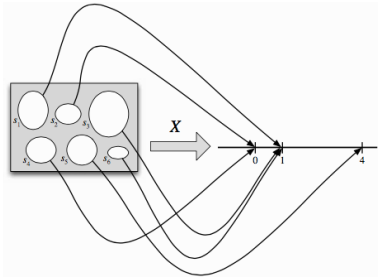


- X is a **discrete random variable**, if its range is countable.
- X is a **continuous random variable**, if it can take on any value in an interval, although the probability that equals any particular value is exactly 0.
- The distribution of a random variable X is a full specification of the probabilities for the events associated with X.

Review of probability theory fundamentals

Random variables

Let S be a sample space. A random variable is a function assigning a real number to every possible outcome of an experiment on S; $X : S \to \mathbb{R}$.



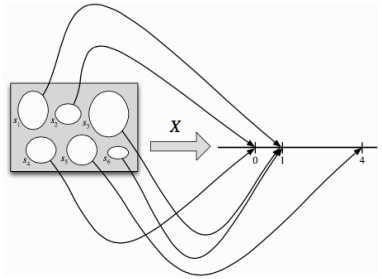
- The distribution of a discrete random variable X is defined by a PMF or a CDF.
- The **Probability Mass Function** of X is the function $P(X = x),\ x \in \mathbb{R}$.
- The **Cumulative Distribution Function** of X is the function $F(x) = P(X \leq x),\ x \in \mathbb{R}$.
- The **Expectation** of X is

$$E(X) = \sum_x xP(X = x)$$

Review of probability theory fundamentals

Random variables

Let S be a sample space. A random variable is a function assigning a real number to every possible outcome of an experiment on S; $X : S \to \mathbb{R}$.



- The distribution of a continuous random variable X is defined by a PDF or a CDF.
- The **Probability Density Function** of X is the derivative of the CDF.
- The **Cumulative Distribution Function** of X is $f(x) = P(X \leq x),\ x \in \mathbb{R}$. It is differentiable.
- The probability is given by the area under the PDF (it is not the value at a point).
- The **Expectation** of X is

$$E(X) = \int_{-\infty}^{\infty} xf(x)$$

Review of probability theory fundamentals

Normal distribution

The Normal (or Gaussian) distribution is a famous continuous distribution with a bell-shaped PDF.

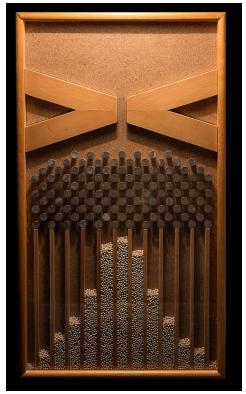
It is widely used in statistics because of the **central limit theorem**, which states that

- ⇝ Under very weak assumptions, the sum of a large number of **independent and identically distributed** random variables has an approximately Normal distribution, regardless of the distribution of the individual random variables.
- ⇝ We can start with independent random variable from almost any distribution, discrete or continuous, but once we add up a bunch of them, the **distribution of the resulting random variable looks like a Normal distribution**.

Review of probability theory fundamentals

Normal distribution

The Normal (or Gaussian) distribution is a famous continuous distribution with a bell-shaped PDF.



Source: Wiki

Galton board

Review of probability theory fundamentals

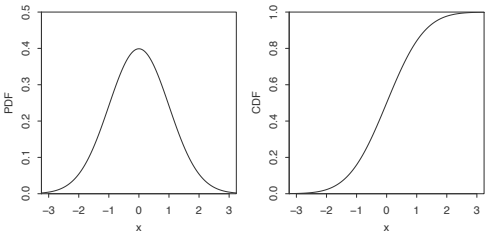
Normal distribution

The Normal (or Gaussian) distribution is a famous continuous distribution with a bell-shaped PDF.

Standard Normal distribution: A continuous random variable Z is said to have the standard normal distribution if its PDF is given by

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty$$

We write $Z \sim N(0,1)$ since it has mean 0 and variance 1.



Review of probability theory fundamentals

Normal distribution

The Normal (or Gaussian) distribution is a famous continuous distribution with a bell-shaped PDF.

Normal distribution: If $Z \sim N(0,1)$, then

$$X = \mu + \sigma Z$$

is said to be the Normal distribution with mean μ and variance σ^2, for any μ and σ^2 with $\sigma > 0$. We write $N \sim N(\mu, \sigma)$.

- If we can get from Z to X, then we can get from X back to Z.
- The process of getting a standard Normal from a non-standard Normal is called, appropriately enough, standardization.
- For $X \sim N(\mu, \sigma^2)$, the standardized version of X is $\frac{X-\mu}{\sigma} \sim N(0,1)$
- The PDF of the Normal distribution can be written as:

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

Linear discriminant analysis

We have seen that using logistic regression we can model the conditional distribution of the response Y, given the predictor(s) X, $P(Y = k|X = x)$, as:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_P X_P}}{1 + e^{\beta_0+\beta_1 X_1+\ldots+\beta_P X_P}}$$

Multiclass logistic regression is also known as **multinomial regression**.

There are some **issues with logistic regression**:

- The coefficients become unstable where there is collinearity.
  - ⇝ This affects the convergence of the fitting.
- When the classes are linearly separated (we can pass an hyperplane between them), the coefficients become unstable.

Linear discriminant analysis

Now, we consider an **alternative and less direct strategy** to estimate these probabilities.

Instead of estimating P(Y|X), we can:

- Model the distribution of the predictors X separately in each of the response classes (i.e. given Y).
- Use the Bayes' theorem to flip these estimates for $P(Y = k|X = x)$.
- Use normal distributions for each class
  - ⇝ This leads to linear or quadratic discriminant analysis.
- Generalize this approach to other distributions can be used as well.

Linear discriminant analysis

Now, we consider an **alternative and less direct strategy** to estimate these probabilities.

Instead of estimating P(Y|X), we can:

- Estimate $P(X|Y)$
  - ⇝ Given the response, what is the distribution of the inputs.
- ⇝ Estimate $P(X)$
  - How likely are each of the categories.
- Use the Bayes' theorem, given the estimates $P(X|Y)$ and $P(Y)$, to obtain:

$$P(Y = k|X = x) = \frac{P(X = x|Y = k)P(Y = k)}{P(X = x)} = \frac{P(X = x|Y = k)P(Y = k)}{\sum_{l=1}^{K} P(X = x|Y = y)P(Y = l)}$$

- For any input x_0, we predict the response as in a **Bayes classifier**:

$$y_0 = \underset{x}{\operatorname{argmax}}\, P(Y = y|X = x_0)$$

Linear discriminant analysis

Suppose that we wish to classify an observation into one of K classes, where $K \geq 2$.

⇝ Y can take on K possible distinct and unordered values.

We can write Bayes theorem in a slightly different way:

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

- π_k is the overall or **prior** probability that a randomly chosen observation comes from the kth category of the response variable Y.
- $f_k(x) \equiv P(X = x|Y = y)$ is the **density function** of X for an observation that comes from the kth class.

Linear discriminant analysis

Suppose that we wish to classify an observation into one of K classes, where $K \geq 2$.

⇝ The qualitative response variable Y can take on K possible distinct and unordered values.

We can write Bayes theorem in a slightly different way:

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

Many techniques are based on models for class densities:

- **Linear and quadratic discriminant analysis use Gaussian densities**.
- More flexible mixture of Gaussians allow for nonlinear decision boundaries.
- Nonparametric density functions for each class density allow the most flexibility.
- Naive Bayes models are variant of the previous case, and assume that each of the class densities are products of marginal densities (they assume that the inputs are conditionally independent in each class).

Linear discriminant analysis

Suppose that we wish to classify an observation into one of K classes, where $K \geq 2$.

⇝ The qualitative response variable Y can take on K possible distinct and unordered values.

We can write Bayes theorem in a slightly different way:

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

- We will use the abbreviation: $p_k(X) = P(Y = k|X)$ for the posterior probability that an observation $X = x$ belongs to the class kth class.
  - ⇝ It is the probability that the observation belongs to the kth class, given the predictor value for that observation.
- Instead of directly computing $p_k(X)$, we can simply estimate π_k and $f_k(X)$.

Linear discriminant analysis

Suppose that we wish to classify an observation into one of K classes, where $K \geq 2$.

$\rightsquigarrow$ The qualitative response variable Y can take on K possible distinct and unordered values.

We can write Bayes theorem in a slightly different way:

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

- In general, estimating π_k is easy if we have a random sample of Ys from the population.
- Estimating $f_k(X)$ tends to be more challenging, unless we assume some simple forms for these densities.
- We know the Bayes classifier, which classifies an observation to the class for which $p_k(X)$ is largest, has the lowest possible error rate out of all classifiers.
- Therefore, if we can find a way to estimate $f_k(X)$, then we can develop a classifier that approximates the Bayes classifier.

Linear discriminant analysis for $p = 1$

We assume for now that we only have one predictor: $p = 1$.

We would like to obtain an estimate for $f_k(x)$ that we can use in order to estimate $p_k(x)$.

We will then classify an observation to the class for which $p_k(x)$ is greatest. In order to estimate $f_k(x)$, we will first make some assumptions about its form.

We assume that $f_k(x)$ is Normal or Gaussian:

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right)$$

For now, let us further assume that $\sigma_1^2 = \ldots = \sigma_K^2$: that is, there is a shared variance term across all K classes, which for simplicity we can denote by σ^2:

$$p_k(X) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_l)^2\right)}$$

Linear discriminant analysis for $p = 1$

The Bayes classifier involves assigning an observation $(X = x)$ to the class for which $p_k(x)$ is larger.

- Taking the log of the expression for $p_k(x)$ and rearranging the terms, it can be shown that this is equivalent to assigning the observation to a class for which the **linear discriminant function** $\delta_k(x)$ is larger, where

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

- This gives the **decision boundary** for the Bayes classifier.

  $\rightsquigarrow$ It is the set of points in which 2 classes are equally probable.

Linear discriminant analysis for $p = 1$

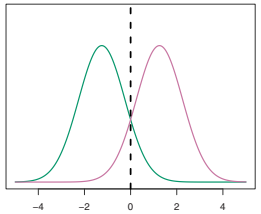
The Bayes classifier involves assigning an observation $(X = x)$ to the class for which $p_k(x)$ is larger.

- If $K = 2$ and $\pi_1 = \pi_2$, the Bayes classifier assigns an observation to Class 1 if

$$2x(\mu_1 - \mu_2) > \mu_1^2 - \mu_2^2$$

  and to class 2 otherwise.

- Here, the Bayes decision boundary corresponds to the point where $x = \frac{\mu_1 + \mu_2}{2}$



- The two normal density functions $f_1(x)$ and $f_2(x)$, represent the two classes.
- We have $\mu_1 = -1.24, \mu_2 = 1.25$ and $\sigma_1 = \sigma_2 = 1$.
- $f_1(x)$ and $f_2(x)$ overlap and given that $X = x$, there is some uncertainty about the class to which the observation belongs.

Linear discriminant analysis for $p = 1$

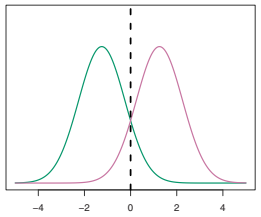
The Bayes classifier involves assigning an observation $(X = x)$ to the class for which $p_k(x)$ is larger.

- If $K = 2$ and $\pi_1 = \pi_2$, the Bayes classifier assigns an observation to Class 1 if

$$2x(\mu_1 - \mu_2) > \mu_1^2 - \mu_2^2$$

and to class 2 otherwise.

- Here, the Bayes decision boundary corresponds to the point where $x = \frac{\mu_1 + \mu_2}{2}$



- Assuming that an observation is equally likely to come from either class ($\pi_1 = \pi_2 = 0.5$), then the Bayes classifier assigns the observation to Class 1 if $x < 0$ and Class 2 otherwise.
- Here we can compute the Bayes classifier because we know that X has a Gaussian distribution within each class and the parameters are known.
- This is not the case in real-life situations.

Linear discriminant analysis for $p = 1$

In practice, even if we are quite certain of our assumption that X is drawn from a Gaussian distribution within each class, we still need to **estimate the parameters** $\mu_1, \mu_2, \ldots, \mu_K, \pi_1, \pi_2, \ldots, \pi_K$ and σ^2.

The **linear discriminant analysis** (LDA) method approximate the Bayes classifier by plugging the estimates for μ_k, σ^2 and π_k.

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i = k} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n - K} \sum_{k=1}^{K} \sum_{i:y_i = k} (x_i - \mu_k)^2$$

n is the total number of training observations.

n_k is the number of training observation in the kth class.

$\hat{\mu}_k$ is the average of all the training observations from the kth class.

$\hat{\sigma}^2$ is a weighted average of the sample variances for each of the K classes.

Linear discriminant analysis for $p = 1$

In practice, even if we are quite certain of our assumption that X is drawn from a Gaussian distribution within each class, we still need to **estimate the parameters** $\mu_1, \mu_2, \ldots, \mu_K, \pi_1, \pi_2, \ldots, \pi_K$ and σ^2.

The **linear discriminant analysis** (LDA) method approximate the Bayes classifier by plugging the estimates for μ_k, σ^2 and π_k.

If we know the class membership probabilities $\pi_1, \ldots, \pi_K$, we can be used it directly.

In the absence of any additional information, LDA estimates π_k using the proportion of the training observations that belong to the kth class.

$$\hat{\pi}_k = \frac{n_k}{n}$$

Linear discriminant analysis for $p = 1$

In practice, even if we are quite certain of our assumption that X is drawn from a Gaussian distribution within each class, we still need to **estimate the parameters** $\mu_1, \mu_2, \ldots, \mu_K, \pi_1, \pi_2, \ldots, \pi_K$ and σ^2.

The **linear discriminant analysis** (LDA) method approximate the Bayes classifier by plugging the estimates for μ_k, σ^2 and π_k.

LDA classifier plugs the estimates for μ_k, σ^2 and π_k into the discriminant functions $\delta_k(x)$ and assigns an observation $X = x$ to the class for which

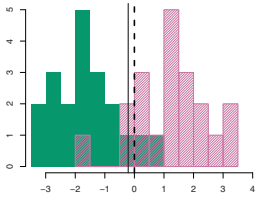
$$\hat{\delta}_k(x) = x \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k) \quad \text{is largest.}$$

$\rightsquigarrow$ The word linear in the classifier's name stems from the fact that the discriminant functions $\hat{\delta}_k(x)$ are linear functions of x.

Linear discriminant analysis for $p = 1$

In practice, even if we are quite certain of our assumption that X is drawn from a Gaussian distribution within each class, we still need to **estimate the parameters** $\mu_1, \mu_2, \ldots, \mu_K, \pi_1, \pi_2, \ldots, \pi_K$ and σ^2.

The **linear discriminant analysis** (LDA) method approximate the Bayes classifier by plugging the estimates for μ_k, σ^2 and π_k.



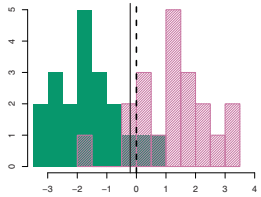
Consider a histogram of a random sample of 20 observations from each class.

- We begin by estimating μ_k, σ^2 and π_k.
- We then computed the decision boundary, shown as a black solid line, that results from assigning an observation to the class for which $\hat{\delta}_k$ is largest.
- All points to the left of this line will be assigned to the green class, while points to the right of this line are assigned to the purple class.

Linear discriminant analysis for $p = 1$

In practice, even if we are quite certain of our assumption that X is drawn from a Gaussian distribution within each class, we still need to **estimate the parameters** $\mu_1, \mu_2, \ldots, \mu_K, \pi_1, \pi_2, \ldots, \pi_K$ and σ^2.

The **linear discriminant analysis** (LDA) method approximate the Bayes classifier by plugging the estimates for μ_k, σ^2 and π_k.



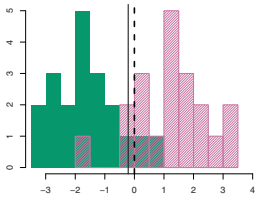
Consider a histogram of a random sample of 20 observations from each class.

- Here, $n_1 = n_2 = 20$ and we have $\hat{\pi}_1 = \hat{\pi}_2$.
- As a result, the decision boundary corresponds to the midpoint between the sample means for the two classes: $(\hat{\mu}_1 + \hat{\mu}_2)/2$.
- The LDA decision boundary is slightly to the left of the optimal Bayes decision boundary $((\mu_1 + \mu_2)/2 = 0)$

Linear discriminant analysis for $p = 1$

In practice, even if we are quite certain of our assumption that X is drawn from a Gaussian distribution within each class, we still need to **estimate the parameters** $\mu_1, \mu_2, \ldots, \mu_K, \pi_1, \pi_2, \ldots, \pi_K$ and σ^2.

The **linear discriminant analysis** (LDA) method approximate the Bayes classifier by plugging the estimates for μ_k, σ^2 and π_k.



How well does the LDA classifier perform on these data?

- These are simulated data, we can generate a large number of test observations in order to compute the **Bayes error rate** and the **LDA test error rate**:

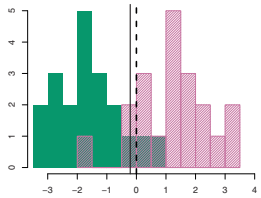
$$1 - E\left(\max_x P(Y = k|X)\right)$$

where the expectation averages the probabilities over all possible values of X.

Linear discriminant analysis for $p = 1$

In practice, even if we are quite certain of our assumption that X is drawn from a Gaussian distribution within each class, we still need to **estimate the parameters** $\mu_1, \mu_2, \ldots, \mu_K, \pi_1, \pi_2, \ldots, \pi_K$ and σ^2.

The **linear discriminant analysis** (LDA) method approximate the Bayes classifier by plugging the estimates for μ_k, σ^2 and π_k.



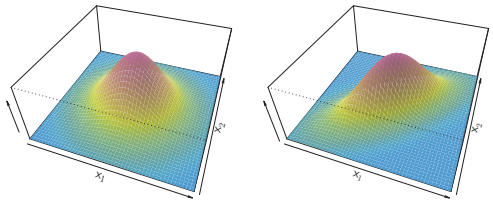
How well does the LDA classifier perform on these data?

- Here, we have that the Bayes and LDA errors are 10.6% and 11.1%, respectively.
- The LDA classifier's error rate is only 0.5% above the smallest possible error rate.
- LDA is performing pretty well on this data set.

Linear discriminant analysis for $p > 1$

To extend the LDA classifier to the case of multiple predictors, we assume that $X = (X_1, X_2, \ldots, X_p)$ is drawn from a **multivariate Gaussian distribution**, with a class-specific mean vector and a common covariance matrix.

- We assume that each individual predictor is follows a one-dimensional distribution with some correlation between each pair of predictors.



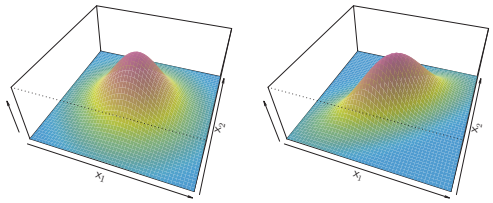
Consider two examples of multivariate Gaussian distributions with $p = 2$.

- In both figures, the height of the surface at any particular point represents the probability that both X_1 and X_2 fall in a small region around that point.

Linear discriminant analysis for $p > 1$

To extend the LDA classifier to the case of multiple predictors, we assume that $X = (X_1, X_2, \ldots, X_p)$ is drawn from a **multivariate Gaussian distribution**, with a class-specific mean vector and a common covariance matrix.

- We assume that each individual predictor is follows a one-dimensional distribution with some correlation between each pair of predictors.



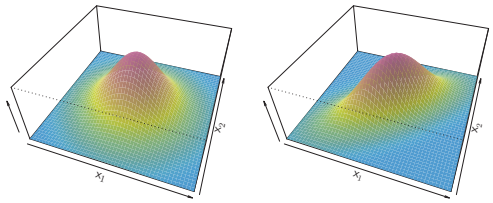
Consider two examples of multivariate Gaussian distributions with $p = 2$.

- **Left example**: $Var(X_1) = Var(X_2)$ and $Cor(X_1, X_2) = 0$.
- This surface has a characteristic bell shape.

Linear discriminant analysis for $p > 1$

To extend the LDA classifier to the case of multiple predictors, we assume that $X = (X_1, X_2, \ldots, X_p)$ is drawn from a **multivariate Gaussian distribution**, with a class-specific mean vector and a common covariance matrix.

- We assume that each individual predictor is follows a one-dimensional distribution with some correlation between each pair of predictors.



Consider two examples of multivariate Gaussian distributions with $p = 2$.

- **Right example**: The bell shape is distorted because the predictors are correlated or have unequal variances.
- The base of the bell will have an elliptical, rather than circular,

Linear discriminant analysis for $p > 1$

To extend the LDA classifier to the case of multiple predictors, we assume that $X = (X_1, X_2, \ldots, X_p)$ is drawn from a **multivariate Gaussian distribution**, with a class-specific mean vector and a common covariance matrix.

- We assume that each individual predictor is follows a one-dimensional distribution with some correlation between each pair of predictors.
- To indicate that a $p-$dimensional random variable X has multivariate Gaussian distribution, we write $X \sim N(\mu, \mathbf{\Sigma})$.

  $E(X) = \mu$ is the mean of X (a vector with p components).

  $Cov(X) = \mathbf{\Sigma}$ is the $p \times p$ covariance matrix of X.
- The multivariate Gaussian density is:

$$f(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\Big(-\frac{1}{2}(x-\mu)^T \mathbf{\Sigma}^{-1}(x-\mu)\Big)$$

Linear discriminant analysis for $p > 1$

To extend the LDA classifier to the case of multiple predictors, we assume that $X = (X_1, X_2, \ldots, X_p)$ is drawn from a **multivariate Gaussian distribution**, with a class-specific mean vector and a common covariance matrix.

- We assume that each individual predictor is follows a one-dimensional distribution with some correlation between each pair of predictors.
- In the case of $p > 1$ predictors, the LDA classifier assumes that the observations in the kth class are drawn from a multivariate Gaussian distribution $N(\mu_k, \mathbf{\Sigma})$

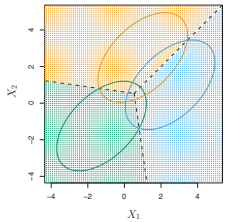
  μ_k is a class-specific mean vector.

  $\mathbf{\Sigma}$ is a covariance matrix that is **common to all K classes**.
- The Bayes classifier assigns an observation $X = x$ to the class for which

$$\delta_k(x) = x^T \mathbf{\Sigma}^{-1} \mu_k - \frac{1}{2} \mu_k^T \mathbf{\Sigma}^{-1} \mu_k + \log(\pi_k) \quad \text{is largest}$$

Linear discriminant analysis for $p > 1$

Consider two examples of multivariate Gaussian distributions with $p = 2$.



- We have three equally-sized Gaussian classes with class-specific mean vectors and a common covariance matrix.
- The three ellipses represent regions that contain 95% of the probability for each of the three classes.
- The dashed lines are the Bayes decision boundaries.

The Bayes decision boundaries represent the set of values x for which $\delta_k(x) = \delta_l(x)$:

$$x^T \mathbf{\Sigma}^{-1} \mu_k - \frac{1}{2} \mu_k^T \mathbf{\Sigma}^{-1} \mu_k = x^T \mathbf{\Sigma}^{-1} \mu_l - \frac{1}{2} \mu_l^T \mathbf{\Sigma}^{-1} \mu_l \quad \text{for } k \neq l$$

- The $\log(\pi_k)$ term has disappeared because each of the three classes has the same number of training observations; i.e. π_k is the same for each class.

Linear discriminant analysis for $p > 1$

Consider two examples of multivariate Gaussian distributions with $p = 2$.

We need to estimate the unknown parameters $\mu_1, \ldots, \mu_K, \pi_1, \ldots, \pi_K$, and $\mathbf{\Sigma}$.

The formulas are similar to those used in the one-dimensional case.

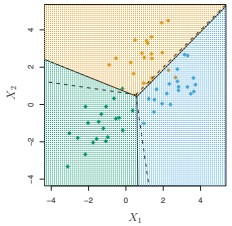
- To assign a new observation $X = x$, LDA plugs the estimates in $\delta_k(x)$ and classifies to the class for which $\hat{\delta}_k(x)$ is largest.
- Again, $\delta_k(x)$ is a linear function of x.

  ⇝ This means that the LDA decision rule depends on x only through a linear combination of its elements.

  Again, this is the reason for the word linear in LDA.

Linear discriminant analysis for $p > 1$

Consider two examples of multivariate Gaussian distributions with $p = 2$.



- 20 observations are generated for each of the three classes.
- The resulting LDA decision boundaries are shown as solid black lines.
- The LDA decision boundaries are pretty close to the Bayes decision boundaries, shown again as dashed lines.

The test error rates for the Bayes and LDA classifiers are 0.0746 and 0.0770, respectively.

⇝ LDA is performing well on this data.