LINEAR DISCRIMINANT ANALYSIS (LDA)

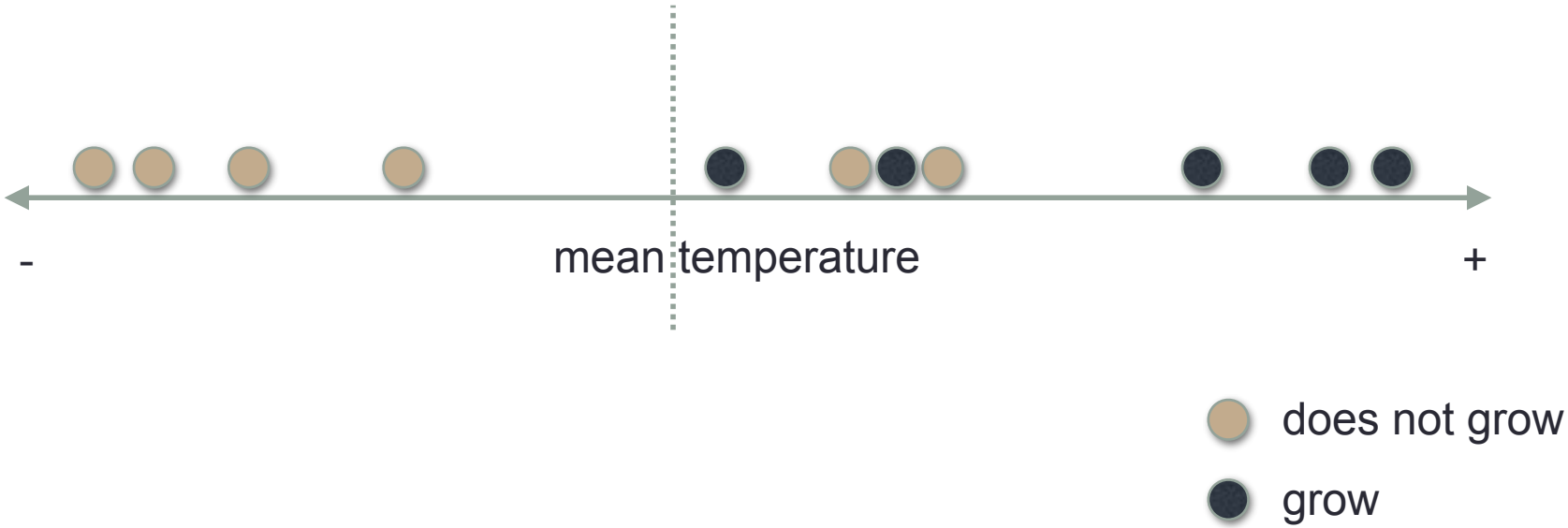
Introducción a la Ciencia de Datos

Linear Discriminant Analysis

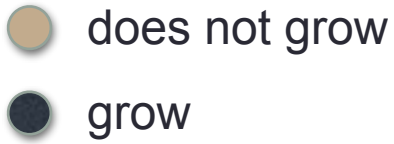
- LDA undertakes the same task as Logistic Regression.
- It classifies data based on categorical variables:
  - Making profit or not
  - Buy a product or not
  - Satisfied customer or not
  - Political party voting intention

Linear Discriminant Analysis

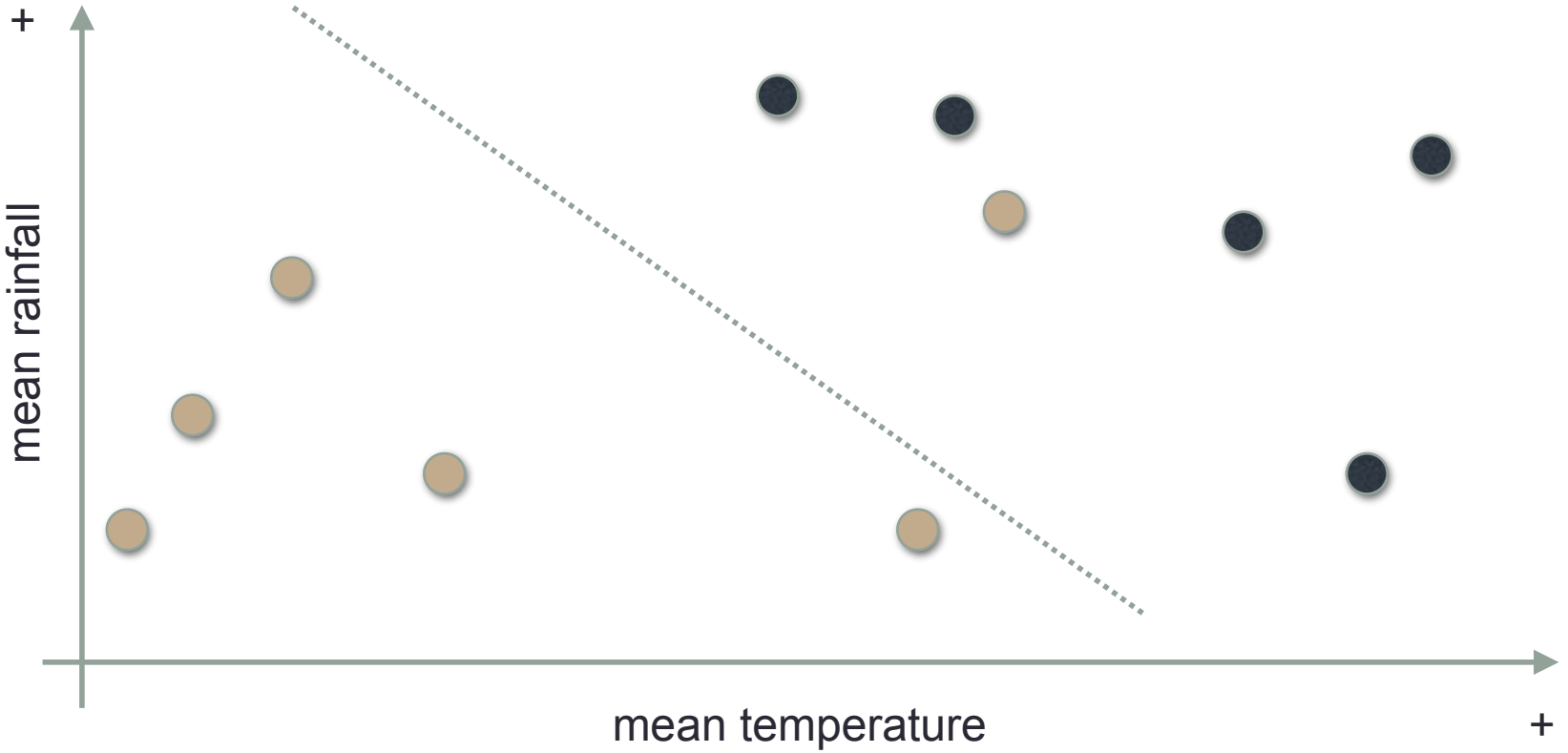
- Suppose we have a data set on plants with an independent variable (mean temperature) and another dependent variable (growth = grows / does not grow).



Linear Discriminant Analysis



- Now we have two independent variables: mean temperature and mean rainfall

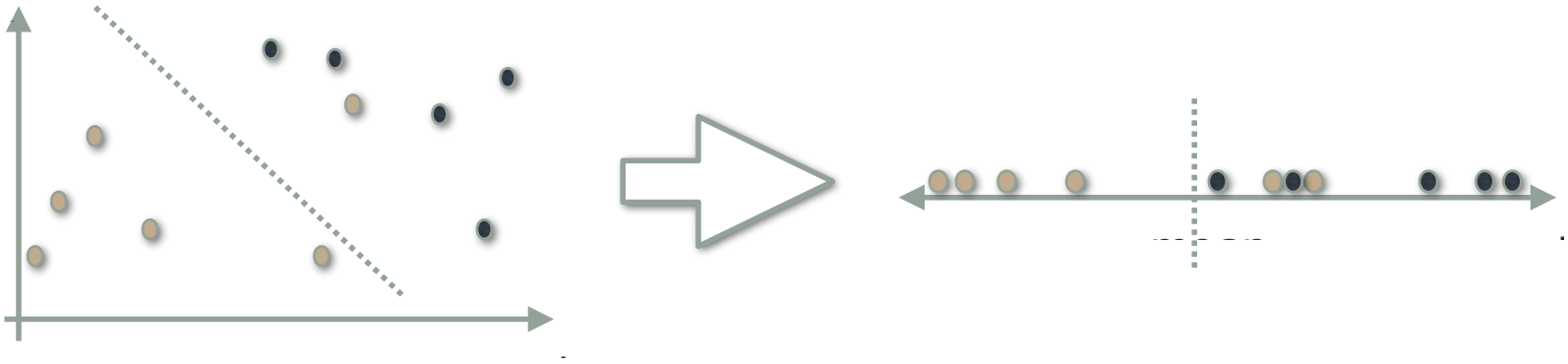


Principal Component Analysis

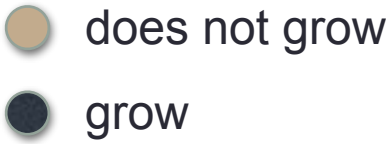
- Principal Component Analysis (PCA) is a technique for dimensionality reduction.
- It transforms the data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful properties of the original data.
- It performs a linear mapping of the data to a lower-dimensional space in such a way that the variance of the data in the low-dimensional representation is maximized.

Linear Discriminant Analysis

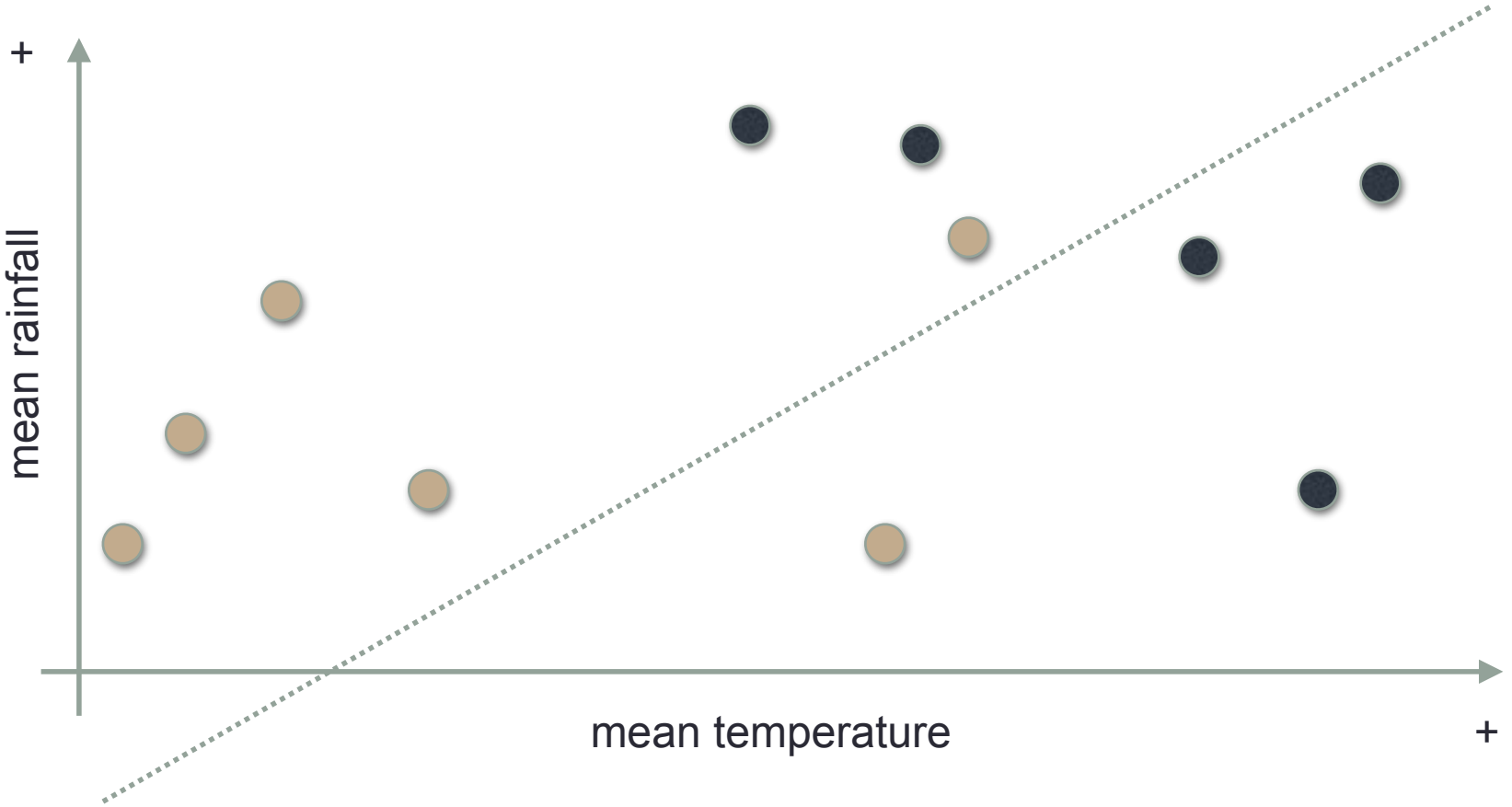
- LDA is like PCA but focuses on maximizing the separability among categories.



Linear Discriminant Analysis



- LDA creates a new axis and projects data to that axis

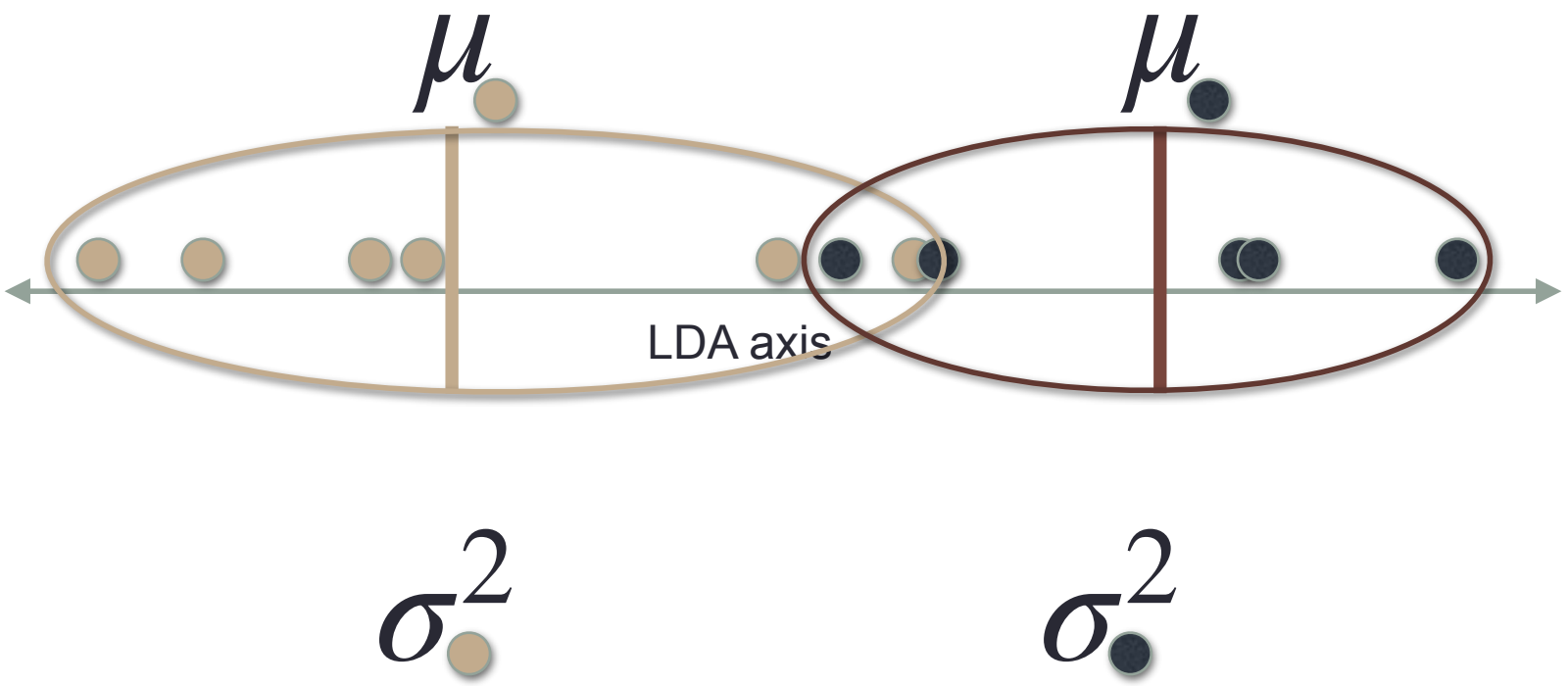


Linear Discriminant Analysis

does not grow

grow

- Maximize the function that represents the difference between the means, normalized by a measure of the within-class variability



Linear Discriminant Analysis

- LDA starts by assuming that:
  - Each independent variable has a normal distribution for each class.
  - Each of the classes has identical variance-covariance matrices.
  - The observations are a random sample.

- Each observation is assigned to the class with the maximum probability among all k probabilities.

Linear Discriminant Analysis

- LDA has 2 distinct stages:
  - **Extraction**: latent variables, called *discriminants*, are formed as linear combinations of the independent variables. The coefficients in that linear combinations are called discriminant coefficients.
  - **Classification**: data points are assigned to classes by those discriminants, not by original variables.

Why Linear? Why Discriminant?

- LDA involves the determination of linear equation that will predict to which group the case belongs to:

$$D = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$$

  - D - discriminant function
  - X_i - variable
  - β_i - discriminant coefficient or weight for the variable X_i
  - β_0 - constant

How to extract the coefficients

- Between-class scatter matrix:

$$S_b = \sum_{i=1}^{K} N_i(\mu_i - \mu)(\mu_i - \mu)^T$$

$$\mu_i = \frac{1}{N_i} \sum_{j=1}^{N_i} x_{ij}$$

$$\mu = \frac{1}{N} \sum_{i=1}^{K} N_i \mu_i$$

- Within-class scatter matrix:

$$S_w = \sum_{i=1}^{K} \sum_{j=1}^{N_i} (x_{ij} - \mu_i)(x_{ij} - \mu_i)^T$$

How to extract the coefficients

- Construct the lower-dimensional space that maximizes between-class variance and minimizes within-class variance

$$P_{lda} = argmax_P \frac{|P^T S_b P|}{|P^T S_w P|}$$

Linear Discriminant Analysis

- LDA has 2 distinct stages:
  - **Extraction**: latent variables called discriminants are formed, as linear combinations of the input variables. The coefficients in that linear combinations are called discriminant coefficients.
  - **Classification**: data points are assigned to classes by those discriminants, not by original variables.

Estimating Bayes' Classifier

- The approach is to model the distribution of X in each of the classes separately, and the use Bayes' theorem to flip things around and obtain P(Y|X).

- We can compute:

$$P(Y = k \mid X = x) = \frac{\pi_k f_k(x)}{\sum_{i=1}^{K} \pi_i f_i(x)}$$

π_k - Probability of coming from class k (prior probability)

$f_k(x)$ - Density function for x given that x is an observation from class k

Normal density

- We use a multivariate normal/gaussian density for $f_k(x)$:

$$f_k(x) = \frac{1}{(2\pi)^{p/2} |\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)}$$

- Σ_k: population variance-covariance matrix
- μ_k: population mean vector

Estimating Bayes' Classifier

- In comparing two classes *a* and *b*, it is sufficient to look at the log-ratio, and see that is an equation linear in *x*:

$$ln\frac{P(Y=a\,|\,X=x)}{P(Y=b\,|\,X=x)} = ln\frac{f_a(x)}{f_b(x)} + ln\frac{\pi_a}{\pi_b} =$$

$$= ln\frac{\pi_a}{\pi_b} - \frac{1}{2}(\mu_a+\mu_b)^T\Sigma^{-1}(\mu_a-\mu_b) + x^T\Sigma^{-1}(\mu_a-\mu_b)$$

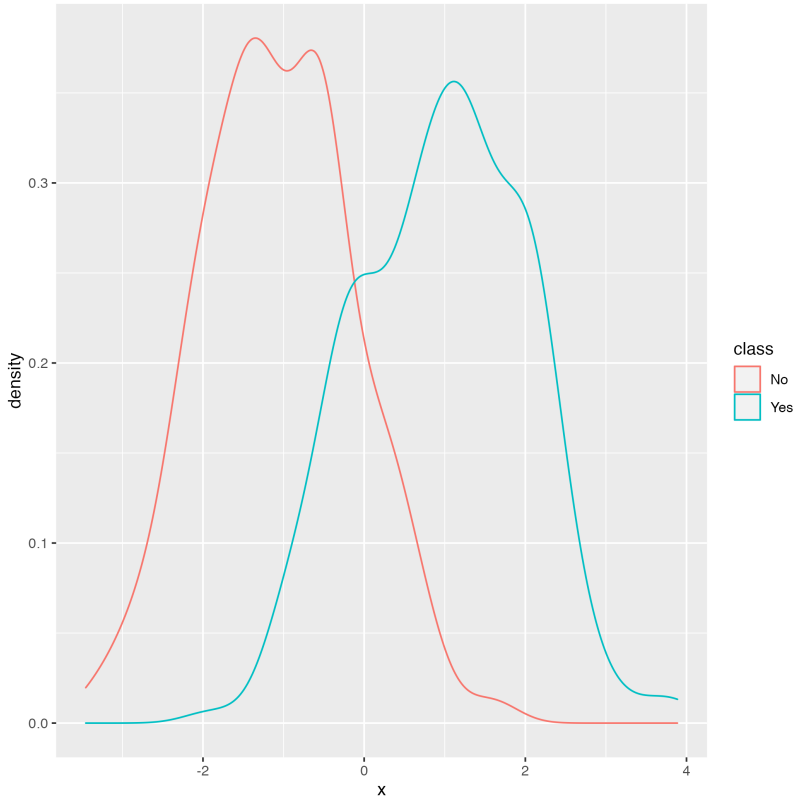
How to extract the coefficients

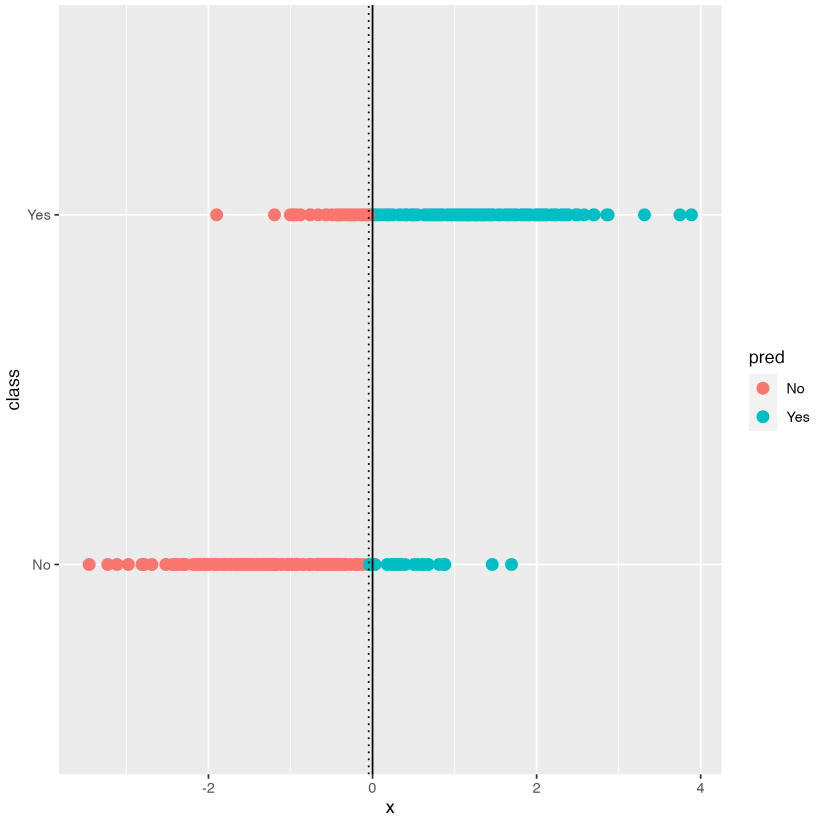
1. Compute the mean vectors for the different classes from the dataset
2. Compute the scatter matrices (between-class and within-class scatter matrix)
3. Compute the eigenvectors and corresponding eigenvalues for the scatter matrices
4. Sort the eigenvectors by decreasing eigenvalues and choose k eigenvectors with the largest eigenvalues to form a d×k dimensional matrix W (where every column represents an eigenvector)
5. Use this d×k eigenvector matrix to transform the samples onto the new subspace (Y=X×W)

https://www.apsl.net/blog/2017/07/18/using-linear-discriminant-analysis-lda-data-explore-step-step/

A Simple Example with One Predictor (p = 1)

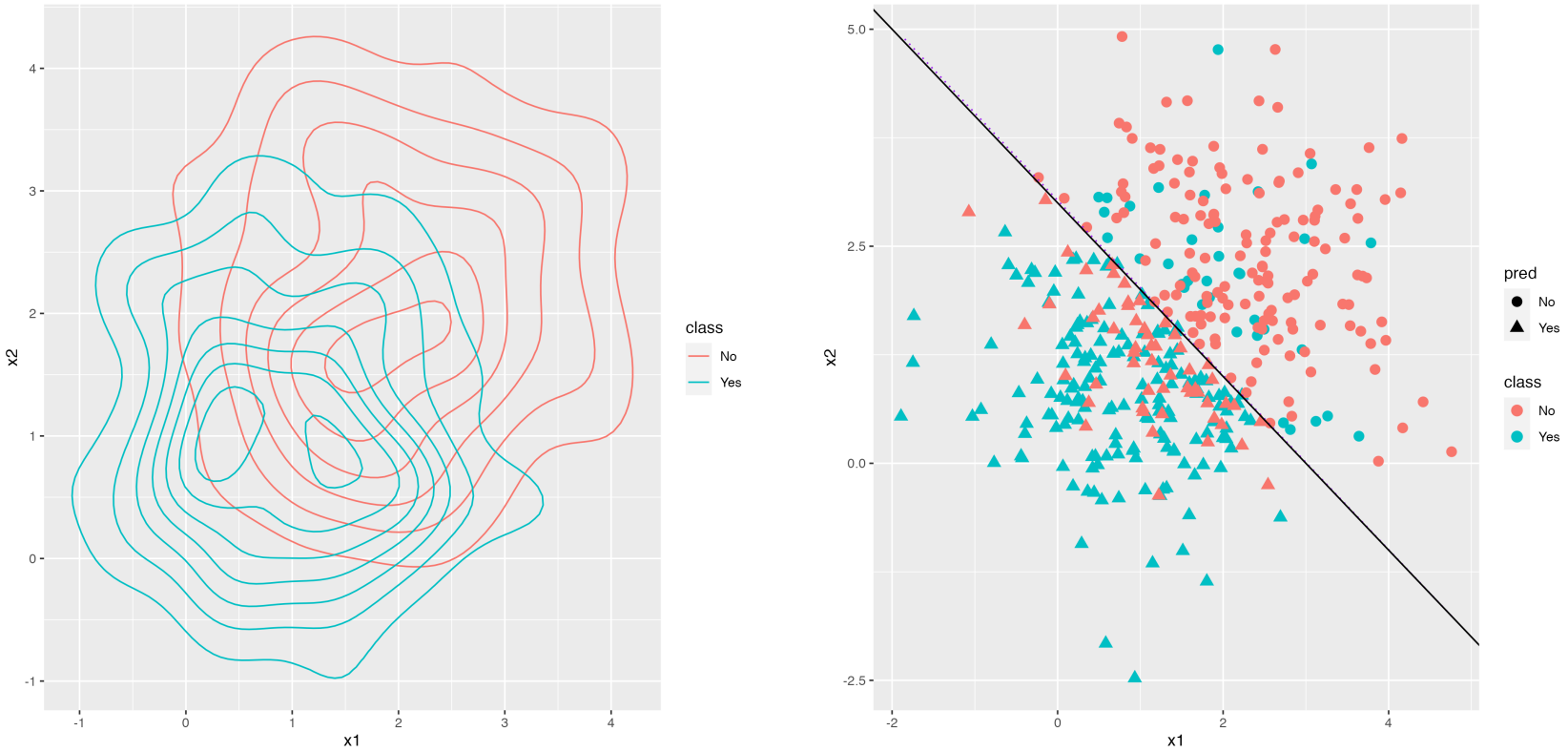
- Suppose we have only one predictor (p = 1)
- Two normal density function $f_1(x)$ and $f_2(x)$, represent two distinct classes





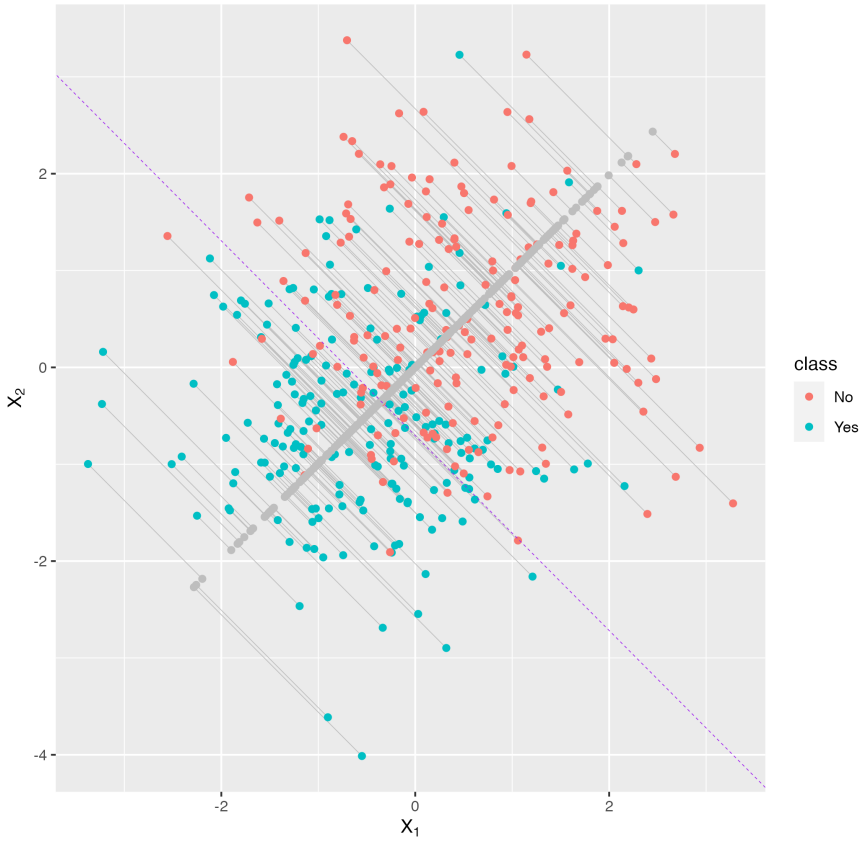
An Example When p > 1

- If X is multidimensional (p > 1), we use exactly the same approach except the density function f(x) is modeled using the multivariate normal density



An Example When p > 1

- If X is multidimensional (p > 1), we use exactly the same approach except the density function f(x) is modeled using the multivariate normal density



An example with > 2 classes

- When there are more than two groups we can estimate more than one discriminant function:

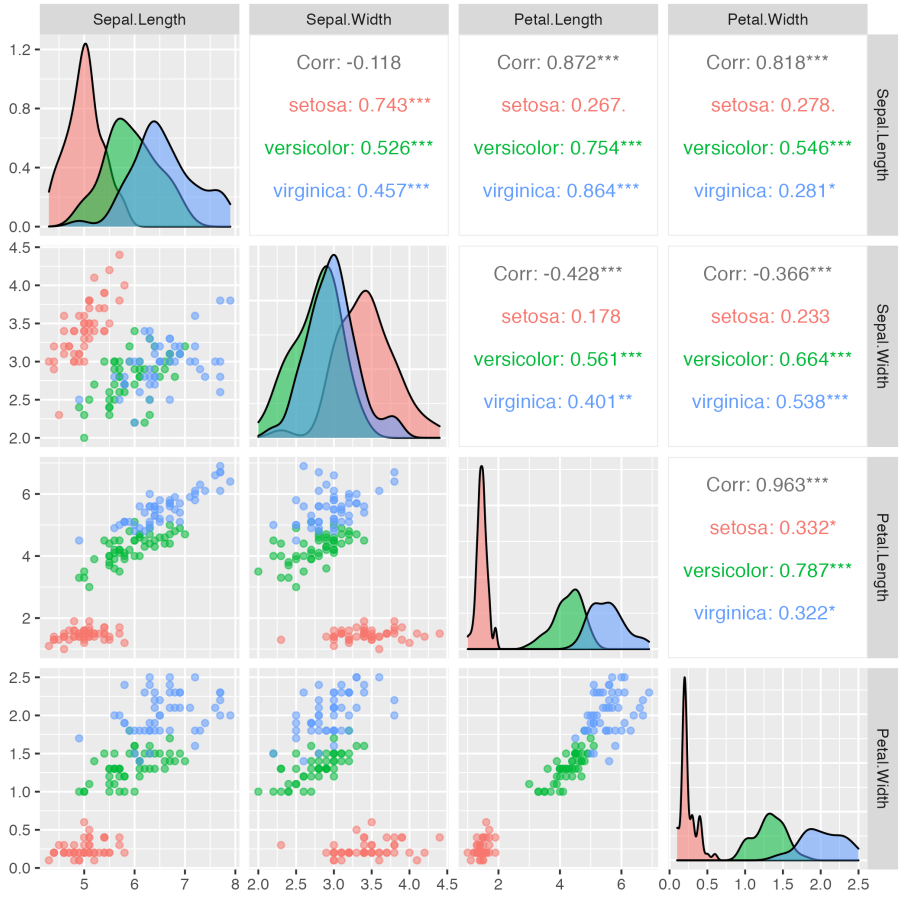
$$D_1 = v_1X_1 + v_2X_2 + \ldots + v_iX_i + a$$

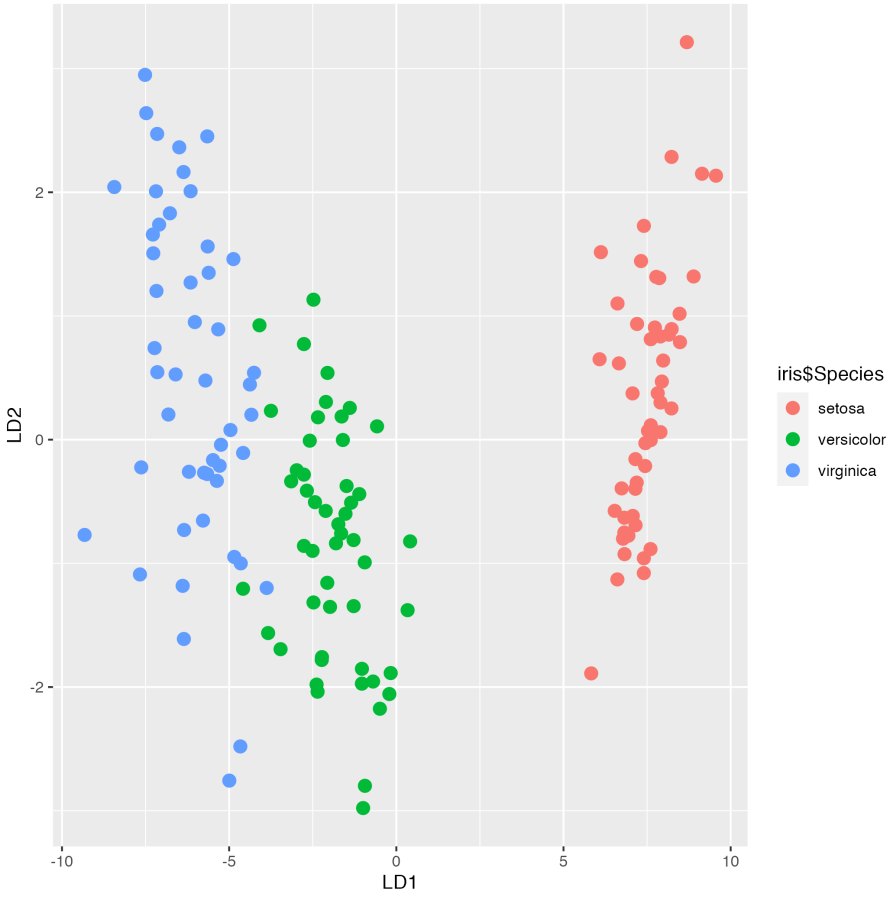
$$D_2 = w_1X_1 + w_2X_2 + \ldots + w_iX_i + b$$

- For example, when there are three groups, we could estimate (1) a function for discriminating between group 1 and groups 2 and 3 combined, and (2) another function for discriminating between group 2 and group 3.

An example with > 2 classes

- We have two predictors (p = 2) and three classes





Why not Logistic Regression?

- Logistic regression is unstable when the classes are well separated
- In the case where *n* is small, and the distribution of predictors X is approximately normal, then LDA is more stable than Logistic Regression
- LDA is more popular when we have more than two response classes

Recommendations

- The LDA solution depends on inverting a covariance matrix, thus a unique solution exists.
  - The data must contain more samples than predictors, and the predictors **must be independent**.
  - It is recommended that predictors be centered and scaled and that near-zero variance predictors be removed.
  - It is also recommended that LDA be used on data sets that have at least 5-10 times more samples than predictors.

Extensions

- Quadratic Discriminant Analysis (QDA):
  - Each class uses its own estimate of variance (or covariance when there are multiple input variables)
- Flexible Discriminant Analysis (FDA):
  - Where non-linear combinations of inputs is used such as splines
- Regularized Discriminant Analysis (RDA):
  - Introduces regularization into the estimate of the variance (actually covariance), moderating the influence of different variables on LDA

QUADRATIC DISCRIMINANT ANALYSIS (QDA)

Introducción a la Ciencia de Datos

Quadratic Discriminant Analysis (QDA)

- LDA assumed that every class has the same variance/ covariance
  - However, LDA may perform poorly if this assumption is far from true
- QDA works identically as LDA except that it estimates separate variances/covariance for each class
  - Now, the number of predictors must be less than the number of cases within each class.
  - Also, the predictors within each class must not have pathological levels of collinearity

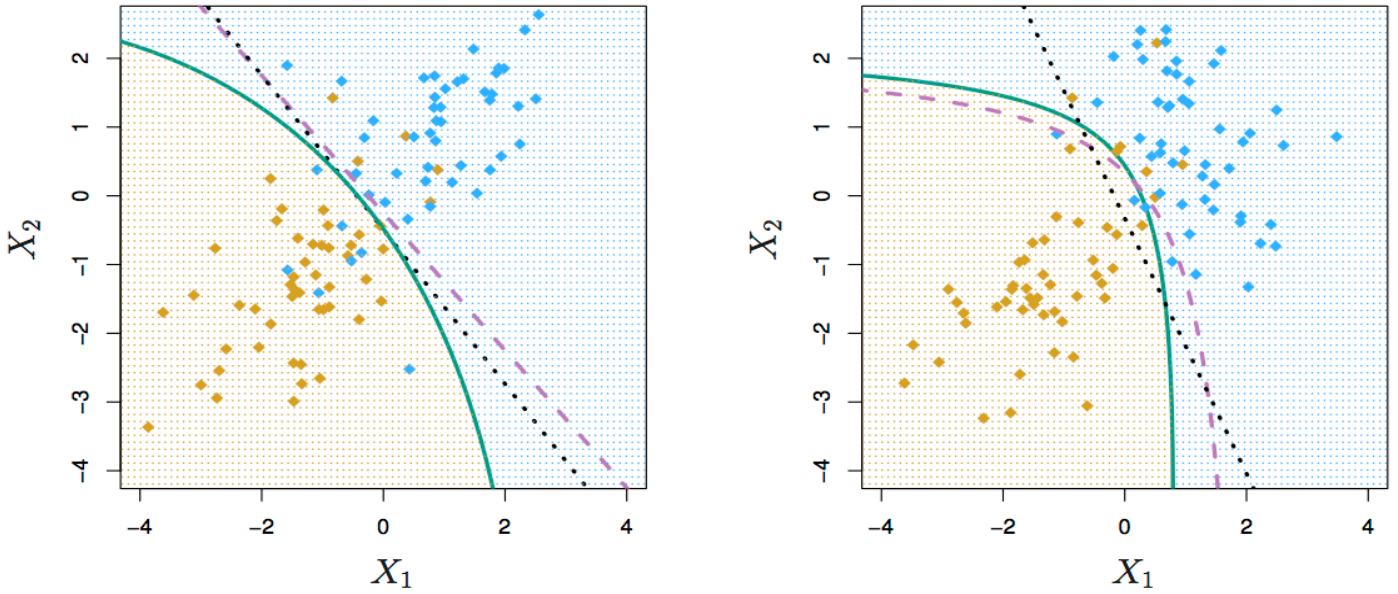
Which is better? LDA or QDA?

- Since QDA allows for different variances among classes, the resulting boundaries become quadratic!

- Which approach is better: LDA or QDA?
  - QDA will work best when the variances are very different between classes and we have enough observations to accurately estimate the variances
  - LDA will work best when the variances are similar among classes or we don't have enough data to accurately estimate the variances

Comparing LDA to QDA

- Black dotted: LDA boundary
- Purple dashed: Bayes' boundary
- Green solid: QDA boundary
- Left: variances of the classes are equal (LDA is better fit)
- Right: variances of the classes are not equal (QDA is better fit)



LDA & QDA

R session

COMPARISON OF CLASSIFICATION METHODS

Introducción a la Ciencia de Datos

Comparison of Classification Methods

- k-NN
- Logistic Regression
- LDA
- QDA

Logistic Regression vs. LDA

- Similarity: Both Logistic Regression and LDA produce linear boundaries
- Difference: LDA assumes that the observations are drawn from the normal distribution with common variance, while logistic regression does not have this assumption. LDA would do better than Logistic Regression if the assumption of normality holds, otherwise logistic regression can outperform LDA

k-NN vs. (LDA and Logistic Regression)

- k-NN takes a completely different approach
- k-NN is completely non-parametric: No assumptions are made about the shape of the decision boundary!
- Advantage of k-NN: We can expect k-NN to dominate both LDA and Logistic Regression when the decision boundary is highly non-linear
- Disadvantage of k-NN: k-NN does not tell us which predictors are important (no table of coefficients!)

QDA vs. (LDA, Logistic Regression, and k-NN)

- QDA is a compromise between non-parametric k-NN method and the linear LDA and logistic regression

- If the true decision boundary is:
  - Linear: LDA and Logistic outperforms
  - Moderately Non-linear: QDA outperforms
  - More complicated: k-NN is superior

Exercise 1

- Use Smarket data (without Today variable)
- Try with `lda` using all Lag variables.
- Repeat with `qda` and compare.

Bibliography

- DSO 530: Applied Modern Statistical Learning Techniques. Abbass Al Sharif. http://www.alsharif.info/#!iom530/c21o7
- An Introduction to Statistical Learning. Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. http://www-bcf.usc.edu/~gareth/ISL/index.html
- Applied Predictive Modeling. Max Kuhn and Kjell Johnson. 2013th Edition. Springer. http://appliedpredictivemodeling.com
- Linear Discriminant Analysis (LDA) clearly explained. https://statquest.org