# MATH 189
# Classification via Linear Discriminant Analysis

Wenxin Zhou
UC San Diego

Time: 2:00—3:20 & 3:30—4:50pm TueThur
Location: CENTR 115

# Outline

- In previous lectures, we considered hypothesis testing problems on population means in both univariate and multivariate settings..
  - Student's $t$-test
  - Hotelling's $T^2$-test
  - Multiple testing with FWER or FDR control

- Today we will start a new topic, namely classification, with a particular focus on the linear discriminant analysis method.
  - Classification problems
  - Bayes' rule in classification
  - Linear discriminant analysis

# Qualitative variables

- Qualitative or categorical variables are very common in many datasets. They naturally reflect the discrete properties of observations.
- When a qualitative variable contains only two outcomes, it is also called a binary variable.

For example:

- Gender, blood type, smoking history, eye color of patients.
- The political party that a voter might vote for.
- Four nucleobases (A,G,T,C) of a DNA double helix.
- Nationalities of students in a department.
- Authenticity of a bank note (Recall the swiss banknotes dataset).
- Location of pottery shards (Recall the pottery dataset).

# Classification

- The process to predict a qualitative variable is called classification. Or we can say we classify that observation into a class/group/category. The statistical tools to classify the observations are called classifiers.
- Examples:
  - A patient arrives at the emergency room with a set of symptoms that could possibly be attributed to one of the three medical conditions. Which one of the three does this patient have?
  - An online banking service must be able to determine whether or not a transaction being processed on the site is fraudulent, on the basis of the user's IP address, transaction history, and so forth.
  - On the basis of DNA sequence data for a number of patients with and without a given disease, a biologist would like to figure out which DNA mutations are dependent (disease-causing) and which are not.

# Classification versus Regression

## Classification

- Predict qualitative variables
- Rarely interested in population parameters
- With or without a response variable

- Many methods first predict the probability of each category. In this sense, they behave like regression methods

## Regression

- Predict quantitative variables
- Inference on population parameters
- Require a response variable to formulate the problem

- Response variable can be qualitative (e.g. logistic regression)

Classification problems are different but connected to regression problems!

# Two Types of Classification

- In general, there are two types of classification problems: supervised and un-supervised, depending on if we know (or partially know) the true values of the qualitative variables in the data.

## Supervised Classification:

- We observed a sample in which we know the true class labels. We usually call this sample as a training set.
- The true labels in the training set can "supervise" us to train a classifier.
- When we observe a sample without labelling, we can classify them using the classifier learned from the training set.

# Two Types of Classification

- In general, there are two types of classification problems: supervised and un-supervised, depending on if we know (or partially know) the true values of the qualitative variable.

## Un-supervised Classification:

- Sometimes, we observe a sample without knowing the true class labels. Sometimes, there are no true class labels at all. There are also cases in which we do not even know how many groups/categories the variables belong to.
- We design a cost function based on the our objective. We classify the observations in order to minimize the cost function.
- Often more challenging than supervised classification.

# Linear Discriminant Analysis

- Linear discriminant analysis (LDA) is a statistical tool with an objective to solve a classification problem when the groups are known as a priori.
- LDA was developed by British statistician Ronald Fisher. Therefore, sometime LDA is also called Fisher's Linear discriminant analysis.
- LDA is different from ANOVA as it is used to predict the group membership of an observation.
- LDA and ANOVA are closely related!

# Example: Iris Dataset

- In 1936, Fisher analyzed an iris flower dataset as an example of linear discriminant analysis. The dataset is collect by American botanist Edagar Anderson. So this dataset is usually called Fisher's Iris Data or Anderson's Iris Data.

- The iris dataset contains measurements for 150 iris flowers from three different species:

Iris-setosa (n=50)

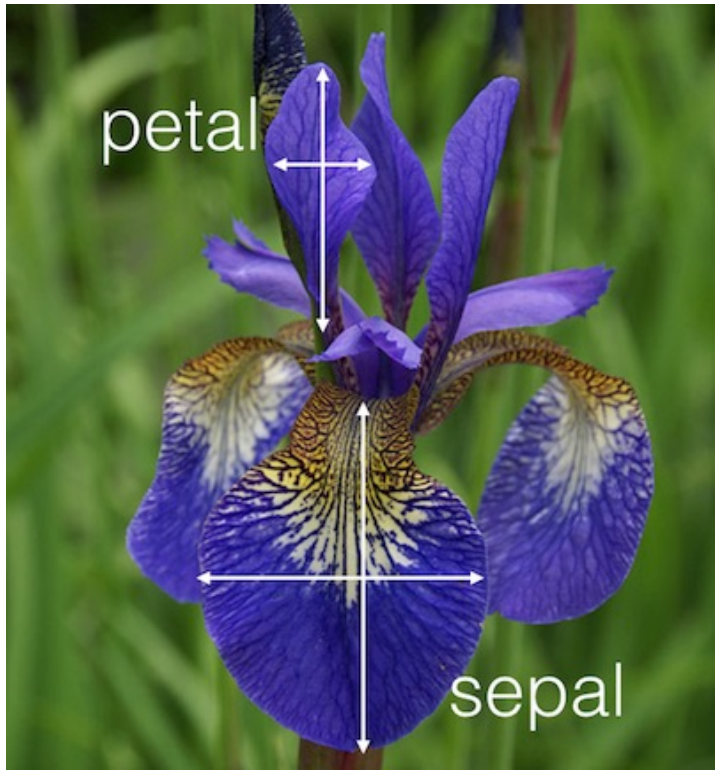
Iris-versicolor (n=50)

Iris-virginica (n=50)

# Example: Iris Dataset (cont.)

- Four variables were measured on each iris flower
  1. sepal length in cm
  2. sepal width in cm
  3. petal length in cm
  4. petal width in cm

Question:
Based on the combination of these four features, can we develop a classifier for iris flower which can classify newly observed iris flowers into one of these three species.



# A Peek at the Data

- Dataset contains 150 iris flowers from 3 species.
- For each flower, the label of species is given and four features are measured.

| Species | Sepal length | Sepal width | Petal length | Petal width |
|---|---|---|---|---|
| setosa | 5.1 | 3.5 | 1.4 | 0.2 |
| setosa | 4.9 | 3 | 1.4 | 0.2 |
| versicolor | 7 | 3.2 | 4.7 | 1.4 |
| versicolor | 6.4 | 3.2 | 4.5 | 1.5 |
| virginica | 6.3 | 3.3 | 6 | 2.5 |
| virginica | 5.8 | 2.7 | 5.1 | 1.9 |

# Bayes' Rule

- The idea of LDA is originated from a classic result in probability theory, called Bayes' rule.

- Consider any two events $A$ and $B$, the probability that *B* occurs given that *A* has occurred, Bayes' rule states the following:

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

- This says that the conditional probability $\mathbb{P}(B|A)$ is the probability that both *A* and *B* occur divided by the unconditional probability that *A* occurs.

- Some simple algebra yields

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

# Some Notations

- Suppose that we have $g$ populations (groups), $y$ is a univariate discrete random variable indicates the group membership, i.e. $y = 1, \ldots, g$.
- Let $p_{\mathrm{i}} = \mathbb{P}(y = i)$ be the probability mass function that a randomly selected observation is in population $i$. $p_{\mathrm{i}}$ is usually called the prior probability.
- Let $f(\boldsymbol{X} = \boldsymbol{x}|y = i)$ be the conditional probability density function of a multivariate random variable $\boldsymbol{X}$ with observed value $\boldsymbol{x}$, given that the observation came from population $i$.
- Let $p(y = i|\boldsymbol{X} = \boldsymbol{x})$ be the conditional probability mass function that an observation is a member of population $i$ given $\boldsymbol{x}$. $p(y = i|\boldsymbol{X} = \boldsymbol{x})$ is usually called the posterior probability.

# Bayes' Rule for Classification

- According to Bayes' rule, we have the following relationship between the prior probability and posterior probability:

$$p(y = i|\boldsymbol{X} = \boldsymbol{x}) = \frac{p(y=i,\boldsymbol{X}=\boldsymbol{x})}{f(\boldsymbol{X}=\boldsymbol{x})} = \frac{p_i f(\boldsymbol{X}=\boldsymbol{x}|y=i)}{\sum_{i=1}^{g} p_i f(\boldsymbol{X}=\boldsymbol{x}|y=i)}.$$

- The numerator gives the likelihood that a randomly selected observation $\boldsymbol{x}$ comes from population $i$.
- The denominator is the unconditional likelihood (over all populations) that we could observe $\boldsymbol{x}$.
- The posterior probability is the probability that an observation $\boldsymbol{x}$ belongs to the $i$-th group.

# Bayes' Classifier

- Given the relationship between the prior and posterior:

$$p(y = i|\boldsymbol{X} = \boldsymbol{x}) = \frac{p_i f(\boldsymbol{X}=\boldsymbol{x}|y=i)}{\sum_{i=1}^{g} p_i f(\boldsymbol{X}=\boldsymbol{x}|y=i)}.$$

- A natural classification rule is to classify the observation $\boldsymbol{x}$ into the group with highest posterior probability. This is called the Bayes' classifier.

- Note that the denominator term is the same for all groups and $\ln(\cdot)$ is a monotonic transformation, find highest posterior corresponds to find the group with highest log-likelihood:

$$\ln\{f(\boldsymbol{X} = \boldsymbol{x}|y = i)p_i\}.$$

# A Classification Rule

## Classification Rule:

- Find the group membership that maximizes a discriminant function:

$$d_i(\boldsymbol{x}) = \ln\{f(\boldsymbol{X} = \boldsymbol{x}|y = i)p_i\}, i = 1, \dots, g.$$

- Assumptions for the LAD are similar to those for MANOVA:
  1. The data from group $i$ has common **mean vector $\boldsymbol{\mu}_i$**, i.e. $\mathbb{E}(\boldsymbol{x}_{ij}) = \boldsymbol{\mu}_i$.
  2. *Homoskedasticity*: The data from all groups have common **variance-covariance matrix $\boldsymbol{\Sigma}$.**
  3. *Independence*: The observations are independently sampled.
  4. *Normality*: The data are **multivariate** normally distributed.

# Discriminant Analysis Procedure

- In general, the procedure of discriminant analysis can be summarized as follows:
  1. Collect training data
     Training data are data with known group memberships. For example, in the Swiss Bank Notes dataset, we know which notes are genuine and which are counterfeit.
  2. Choose prior probability
     We calculate or assign prior probability $p_{\mathrm{i}} = \mathbb{P}(y = i)$ for each group.
  3. Estimate the parameters of $f(\boldsymbol{X}=\boldsymbol{x} \mid y=i)$.
     Under the assumptions stated in the previous slide, we assume the conditional probability density function $f(\boldsymbol{X}=\boldsymbol{x} \mid y=i)$ is the density function of a multivariate normal distribution.
  4. Calculate the discriminant function
     Calculate $d_i(\boldsymbol{x}) = \ln\{f(\boldsymbol{X} = \boldsymbol{x}|y = i)p_i\}$ for each group, i.e. $i = 1, \dots, g$.
  5. Classify observation $\boldsymbol{x}$
     Classify observation $\boldsymbol{x}$ to a group with highest $d_i(\boldsymbol{x})$.

# Choose Prior Probability

The prior probability $p_i = \mathbb{P}(y = i)$ represents the expected portion of the community that belongs to group $i$. There are three common choices:

1. Equal priors:

$$p_i = 1/g.$$

Equal priors are useful if we believe that all of the population sizes are more or less the same.

2. Arbitrary priors:

Any combination that satisfies $p_1 + p_2 + \cdots + p_g = 1$.

Arbitrary priors are usually selected according to the investigators beliefs regarding the relative population sizes.

3. Training data priors:

$$p_i = n_i/N.$$

Training data priors are selected if we believe the relative sample sizes in the training data are close to the relative population sizes.

# Estimate the Parameters of $f(\boldsymbol{X}=\boldsymbol{x}\,|\,y=i)$.

We assume that in population $i$, the random variable $\boldsymbol{X}$ is multivariate normal with mean $\boldsymbol{\mu_i}$ and covariance matrix $\boldsymbol{\Sigma}$ (same for all populations). Then the conditional density function is

$$f(\boldsymbol{X}=\boldsymbol{x}\,|\,y=i)=\frac{1}{(2\pi)^{m/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right\}.$$

Both $\boldsymbol{\mu_i}$ and $\boldsymbol{\Sigma}$ are unknown parameters to us. We estimate them using the training data.

- Population means $\boldsymbol{\mu_i}$ are estimated by the sample means of $i$-th group $\bar{\boldsymbol{x}}_i$.

- Population covariance matrix $\boldsymbol{\Sigma}$ is estimated by the pooled sample covariance matrix

$$\mathbf{S}=\frac{\sum_{i=1}^{g}(n_i-1)\mathbf{S}_i}{\sum_{i=1}^{g}(n_i-1)}=\frac{1}{N-g}\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_i)',$$

where $\mathbf{S}_i$ is the sample covariance matrix of group $i$ using the training data.

# Calculate the Discriminant Function

- The discriminant function $d_i(\boldsymbol{x}) = \ln\{f(\boldsymbol{X} = \boldsymbol{x}|y = i)p_i\}$ can be calculated as

$$
\begin{aligned}
d_i(\boldsymbol{x}) &= \ln f(\boldsymbol{X} = \boldsymbol{x}|y = i) + \ln p_i \\
&= \ln \frac{1}{(2\pi)^{m/2}|\boldsymbol{\Sigma}|^{1/2}} - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i) + \ln p_i \\
&= \underbrace{\ln \frac{1}{(2\pi)^{m/2}|\boldsymbol{\Sigma}|^{1/2}} - \frac{1}{2}\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}}_{\text{group-independent}} \underbrace{- \frac{1}{2}\boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \ln p_i}_{\text{group-dependent}}
\end{aligned}
$$

- With some calculations, we decompose $d_i(\boldsymbol{x})$ into 5 terms. The first two terms do not dependent on the group membership $i$. Therefore, we can ignore them without changing the rank of $d_i(\boldsymbol{x})$.

# Linear Discriminant Function

- Motivated by the previous decomposition, we define a simplified discriminant function:

$$d_i^{\mathrm{L}}(\boldsymbol{x}) = -\frac{1}{2}\boldsymbol{\mu_i}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \boldsymbol{\mu_i}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \ln p_i$$

- We call $d_i^{\mathrm{L}}(\boldsymbol{x})$ the **linear** discriminant function as it is a linear function of $\boldsymbol{x}$:

$$d_i^{\mathrm{L}}(\boldsymbol{x}) = \alpha_i + \sum_{k=1}^{m}\beta_{i,k}x_k = \alpha_i + \boldsymbol{\beta}_i'\boldsymbol{x},$$

where $\alpha_i = -\frac{1}{2}\boldsymbol{\mu_i}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln p_i$, $\boldsymbol{\beta}_i = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i$ (or $\boldsymbol{\beta}_i' = \boldsymbol{\mu_i}'\boldsymbol{\Sigma}^{-1}$) and $\beta_{i,k}$ is the $k$-th entry of $\boldsymbol{\beta}_i$.

- Classify $\boldsymbol{x}$ into the group with highest linear discriminant function is equivalent to classify $\boldsymbol{x}$ into the group with highest posterior.

# Estimate Linear Discriminant Function

- Given a training dataset, we can estimate the linear discriminant function by:

$$\hat{d}_i^{\mathrm{L}}(\boldsymbol{x}) = -\frac{1}{2}\overline{\boldsymbol{x}}_i'\mathbf{S}^{-1}\overline{\boldsymbol{x}}_i + \overline{\boldsymbol{x}}_i'\mathbf{S}^{-1}\boldsymbol{x} + \ln p_i,$$

  where $\overline{\boldsymbol{x}}_i$ is the sample mean of group $i$, $\mathbf{S}$ is the pooled sample covariance matrix.

- This is a function of the sample mean vectors, the pooled covariance matrix, and prior probabilities for $g$ different populations.

- The estimated linear discriminant function can also be written as a linear form:

$$\hat{d}_i^{\mathrm{L}}(\boldsymbol{x}) = \hat{\alpha}_i + \widehat{\boldsymbol{\beta}}_i'\boldsymbol{x}$$

  with $\hat{\alpha}_i = -\frac{1}{2}\overline{\boldsymbol{x}}_i'\mathbf{S}^{-1}\overline{\boldsymbol{x}}_i + \ln p_i$ and $\widehat{\boldsymbol{\beta}}_i = \mathbf{S}^{-1}\overline{\boldsymbol{x}}_i$.

# Linear Discriminant Analysis: Procedure

- Given a training dataset which contains $g$ populations. Suppose we have a new observation $\boldsymbol{x}$ that to be classified into one of the $g$ populations.
- We can apply linear discriminant analysis as follows:
  1. Choose priors $p_i$ for $i = 1, \dots, g$.
  2. Calculate sample mean vectors $\bar{\boldsymbol{x}}_i$ for $i = 1, \dots, g$.
  3. Calculate pooled sample covariance matrix
     $\mathbf{S} = \frac{1}{N-g} \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)'$.
  4. Calculate coefficients $\hat{\alpha}_i = -\frac{1}{2} \bar{\boldsymbol{x}}_i' \mathbf{S}^{-1} \bar{\boldsymbol{x}}_i + \ln p_i$, for $i = 1, \dots, g$.
  5. Calculate coefficients $\hat{\boldsymbol{\beta}}_i = \mathbf{S}^{-1} \bar{\boldsymbol{x}}_i$.
  6. Calculate $\hat{d}_i^{\mathrm{L}}(\boldsymbol{x}) = \hat{\alpha}_i + \hat{\boldsymbol{\beta}}_i' \boldsymbol{x}$, for $i = 1, \dots, g$.
  7. Classify $\boldsymbol{x}$ into the group with highest $\hat{d}_i^{\mathrm{L}}(\boldsymbol{x})$.

# Example: LAD on Iris Dataset

1. Divide the iris dataset into a training data of 120 observations (40 from each species) and a prediction data of 30 observations (10 from each species). For the prediction set, we pretend we do not know the true species.

2. Choose the prior for each species according to the relative sample size in the training data:
$$p_1 = p_2 = p_3 = \frac{40}{120} = \frac{1}{3}.$$

3. Estimate the coefficients in linear discriminant functions using the training data.

4. Calculate the linear discriminant functions for each observation in the prediction data.

5. Classify each observation in the prediction dataset to the species with highest linear discriminant function value.

6. Compare the estimated species labels with the truth.

# Example: LAD on Iris Dataset ($\overline{\boldsymbol{x}}_i$ and $\mathbf{S}$)

- The sample mean vectors of each species are given in the following table:

| Species | Sepal length | Sepal width | Petal length | Petal width |
|---|---|---|---|---|
| setosa | 5.0375 | 3.4525 | 1.4600 | 0.2350 |
| versicolor | 6.0100 | 2.7800 | 4.3175 | 1.3500 |
| virginica | 6.6225 | 2.9600 | 5.6075 | 1.9900 |

- The pooled sample covariance matrix is provided in the following table:

| | Sepal length | Sepal width | Petal length | Petal width |
|---|---|---|---|---|
| Sepal length | 0.2909 | 0.0980 | 0.1810 | 0.0389 |
| Sepal width | 0.0980 | 0.1181 | 0.0547 | 0.0345 |
| Petal length | 0.1810 | 0.0547 | 0.1928 | 0.0460 |
| Petal width | 0.0389 | 0.0345 | 0.0460 | 0.0422 |

# Example: LDA on Iris Dataset ($\hat{\alpha}_i$ and $\widehat{\boldsymbol{\beta}}_i$)

- The intercepts of linear discriminant function $\hat{\alpha}_i$:

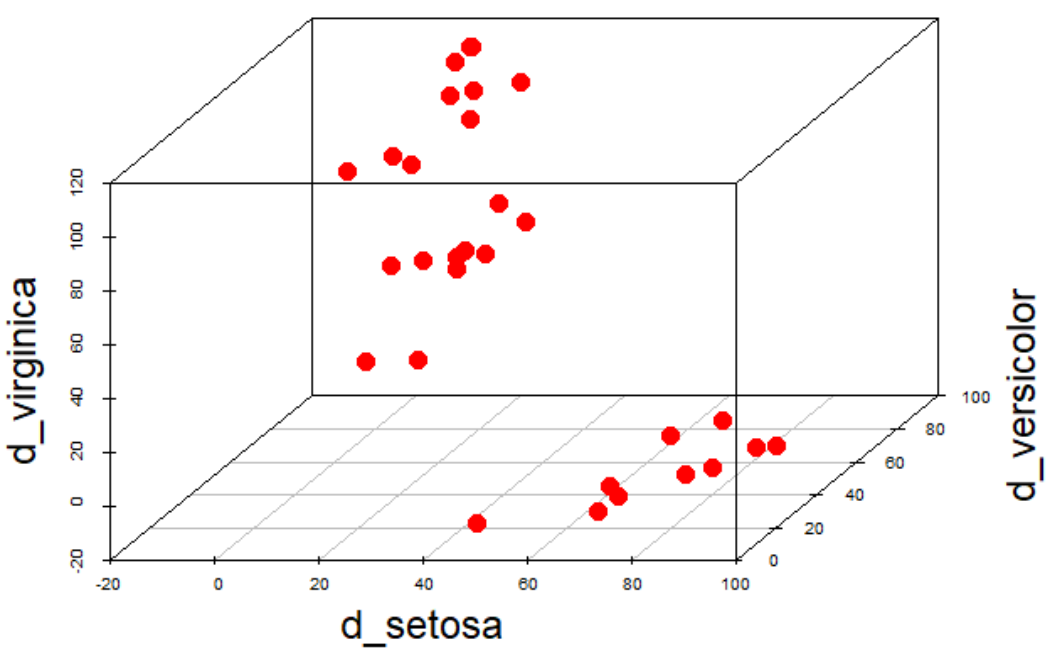
| Intercept | setosa | versicolor | virginica |
|---|---|---|---|
| $\hat{\alpha}_i$ | -80.3597 | -68.5151 | -97.7910 |

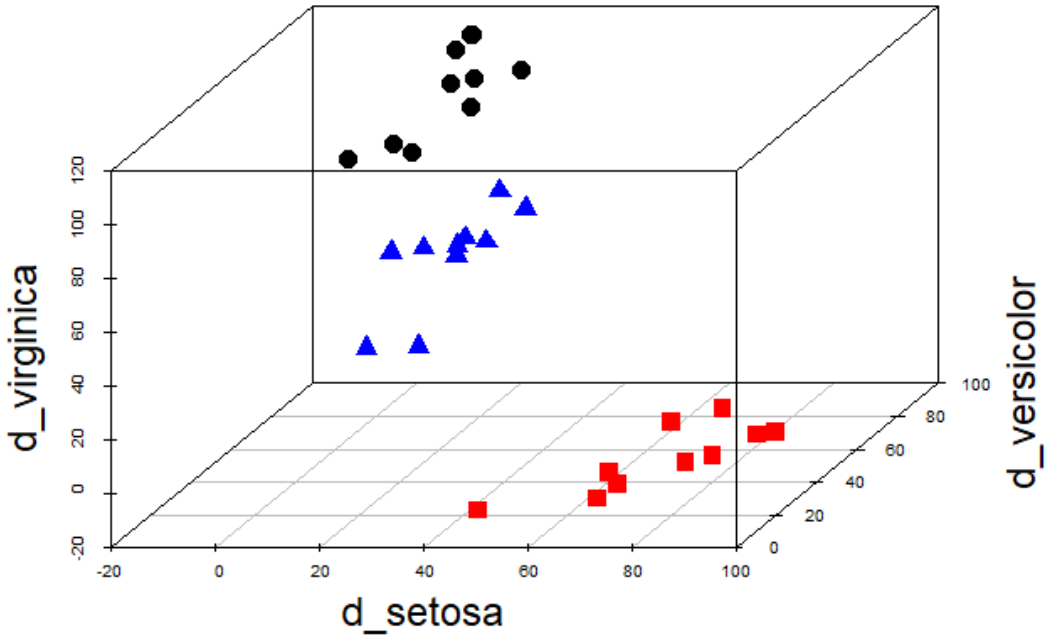
- The slope coefficients vector in linear discriminant function $\widehat{\boldsymbol{\beta}}_i$:

| Intercept | setosa | versicolor | virginica |
|---|---|---|---|
| $\widehat{\boldsymbol{\beta}}_i$ | 20.4884 | 13.4881 | 10.0227 |
| | 23.8916 | 7.5837 | 4.8309 |
| | -14.3348 | 5.9383 | 13.7926 |
| | -17.2257 | 6.8479 | 18.8765 |

# Example: LDA on Iris Dataset (LD Function Values)

- We calculate the LD function values for each observation in the test dataset. The results are presented in the following 3D scatter plot:




setosa obs. versicolor obs. virginica obs.

# Example: LDA on Iris Dataset (Classification Results)

- The classification results versus truth are given in the following table:

| | Setosa | versicolor | virginica |
|---|---|---|---|
| # Observations | 10 | 10 | 10 |
| # Correct Classification | **10** | **10** | **10** |
| # False Classification | 0 | 0 | 0 |

- All the observations in the test data are correctly classified into their true groups!

- Looks great, but:
  1. Large train, small test.
  2. Near oracle prior.
  3. Non-random and balanced partition of train and test.

# A Binary Classification Example: Why LDA Works?

- Suppose there are only two groups in the training data, i.e. $y = 1, 2$.
- Classify an observation $\boldsymbol{x}$ into the first group is equivalent to

$$d_1^{\mathrm{L}}(\boldsymbol{x}) - d_2^{\mathrm{L}}(\boldsymbol{x}) > 0.$$

- Note that this difference can also be written as a linear function of $\boldsymbol{x}$:

$$d_1^{\mathrm{L}}(\boldsymbol{x}) - d_2^{\mathrm{L}}(\boldsymbol{x}) = -\frac{1}{2}\boldsymbol{\mu_1'}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_1} + \frac{1}{2}\boldsymbol{\mu_2'}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_2} + \ln p_1 - \ln p_2 + (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}.$$

- Denote $c = \frac{1}{2}(\boldsymbol{\mu_1'}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_1} - \boldsymbol{\mu_2'}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_2}) - (\ln p_1 - \ln p_2)$ and $\boldsymbol{\omega} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})$, the inequality at top is equivalent to

$$\boldsymbol{\omega}'\boldsymbol{x} > c.$$

# LDA for Binary Classification (cont.)

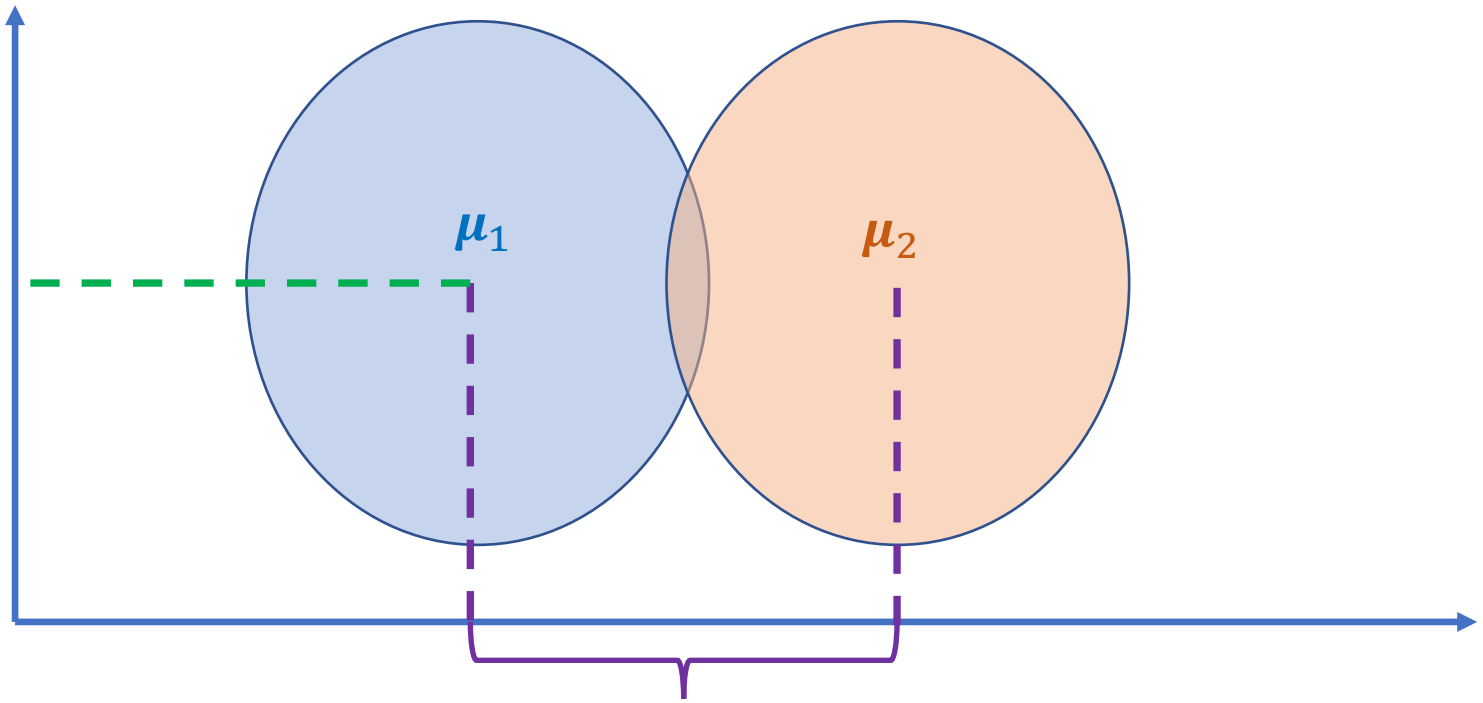
- For a given training data, we can estimate $\boldsymbol{\omega}$ and $c$ by $\hat{\boldsymbol{\omega}}$ and $\hat{c}$.

- The classification rule for an observation $\boldsymbol{x}$ becomes
  - Classify $\boldsymbol{x}$ to group 1 if $\boldsymbol{\omega}'\boldsymbol{x} > c$;
  - Classify $\boldsymbol{x}$ to group 2 if $\boldsymbol{\omega}'\boldsymbol{x} < c$.

- The vector $\boldsymbol{\omega} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})$ is a "direction" that makes the two groups (populations) most separable:
  - Suppose the covariates are independent, i.e. $\boldsymbol{\Sigma}$ is an identity matrix.
  - Then $\boldsymbol{\omega}$ is in the same direction as $(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})$.
  - $\boldsymbol{\omega}'\boldsymbol{x} = \|\boldsymbol{\omega}\| \cdot \|\boldsymbol{x}\| \cdot \cos(\theta)$ which is maximized when $\theta = 0$ or $\pi$.
  - This means $\boldsymbol{x}$ and $\boldsymbol{\omega}$ (and hence $\boldsymbol{\mu_1} - \boldsymbol{\mu_2}$) are in the the same "direction"!

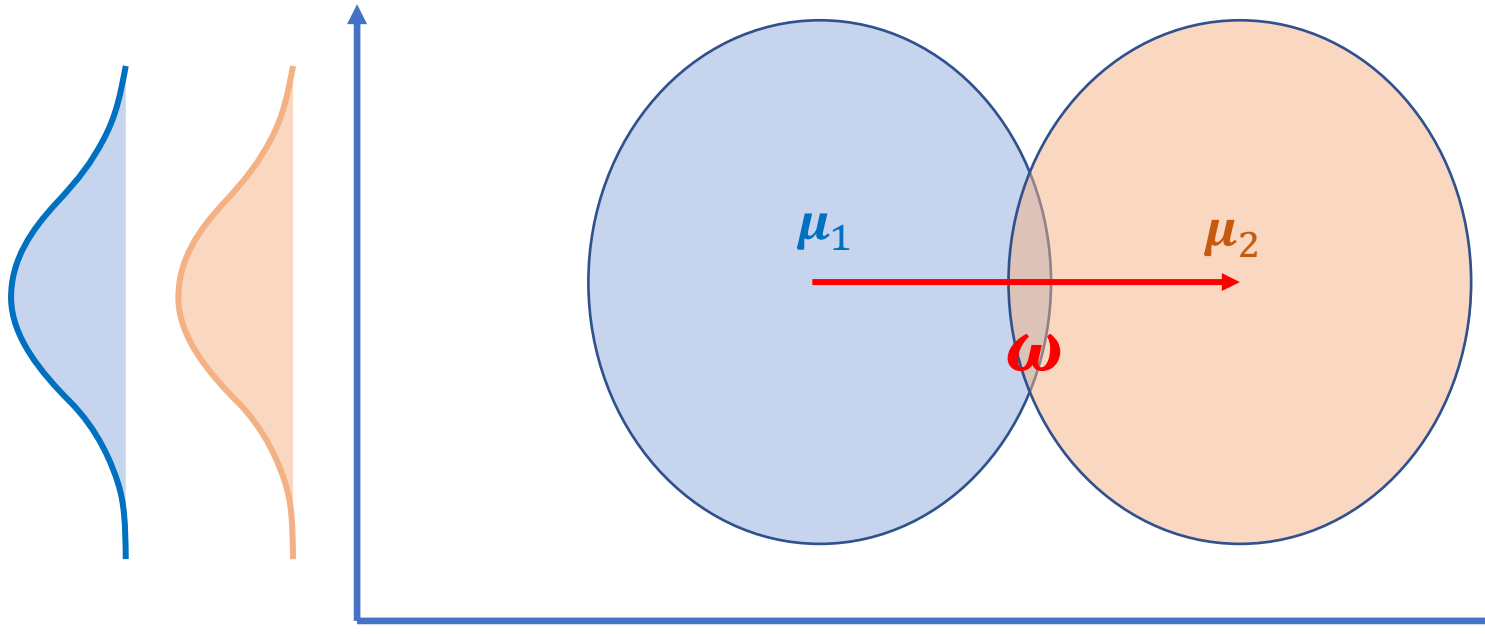
- Suppose we have two 2-dimensional populations. For illustration purpose, we assume their population means only differ at the first coordinate.

- $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are not separable in the second direction.




- $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are most separable along the first direction.

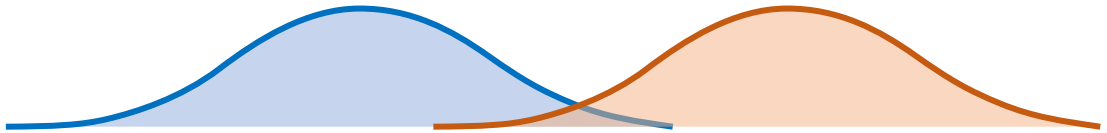
- If we draw the marginal density functions on two dimensions, two populations are more separable along the first direction than the second direction.





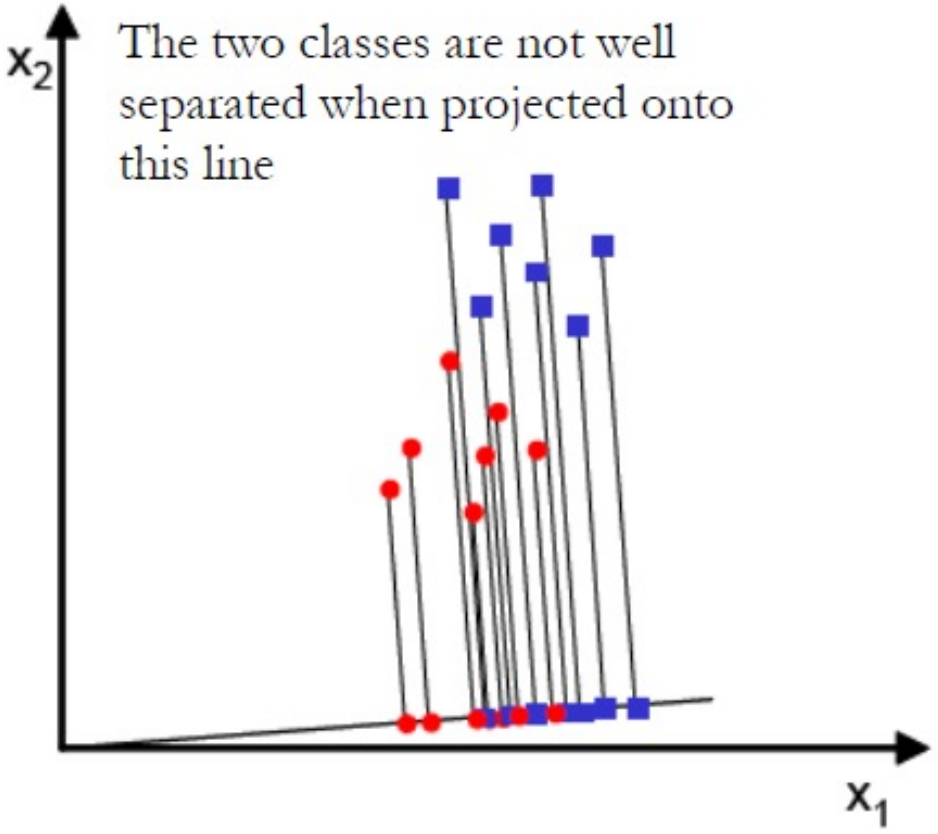

LDA chooses the direction that separates the two populations most!

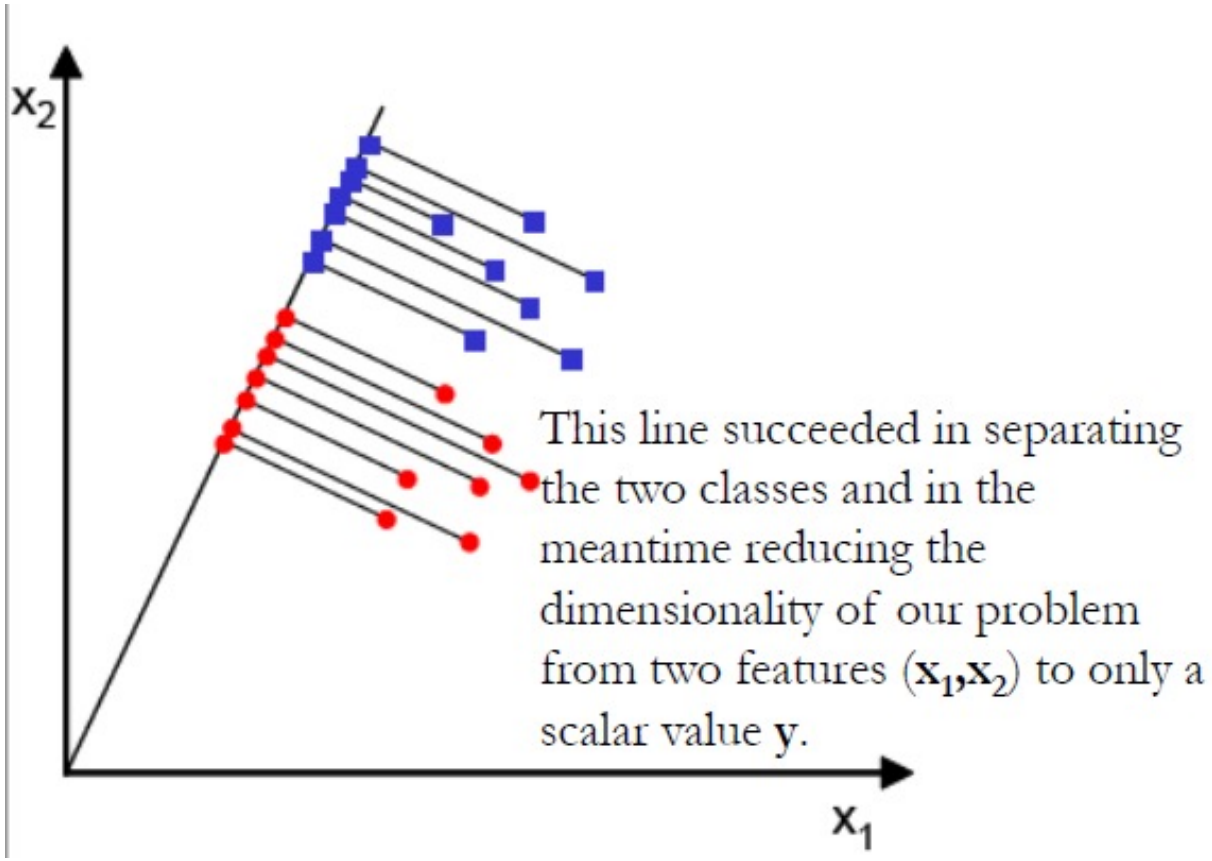
- Two populations are not separable in the second direction.



- Two populations are most separable along the first direction.

# Another Example



# LDA as a Dimension Reduction Tool

- As discussed in the binary classification example, LDA uses one number $(\boldsymbol{\omega}'\boldsymbol{x} - c)$ to classify an observation $\boldsymbol{x} \in \mathbb{R}^m$ that contains $m$ features.
- LDA reduced the dimensionality of a classifier from $m$ to 1.

- Now, let's discuss it in a multi-class classification setting:
  - Suppose we observe a training sample with $g$ classes.
  - For each class $i = 1, \dots, g$, we collect $n_i$ observations.
  - For each observation, we measure $m$ features.

  - To apply LDA, we calculate $g$ linear combinations of $m$ features.
  - When $g$ is much smaller than $m$, we reduce the dimensionality of our classifier from $m$ to $g$.

  - In addition, the $g$ linear combinations correspond to $g$ directions, along which the $g$ populations are most separable!