innovate achieve lead

BITS Pilani

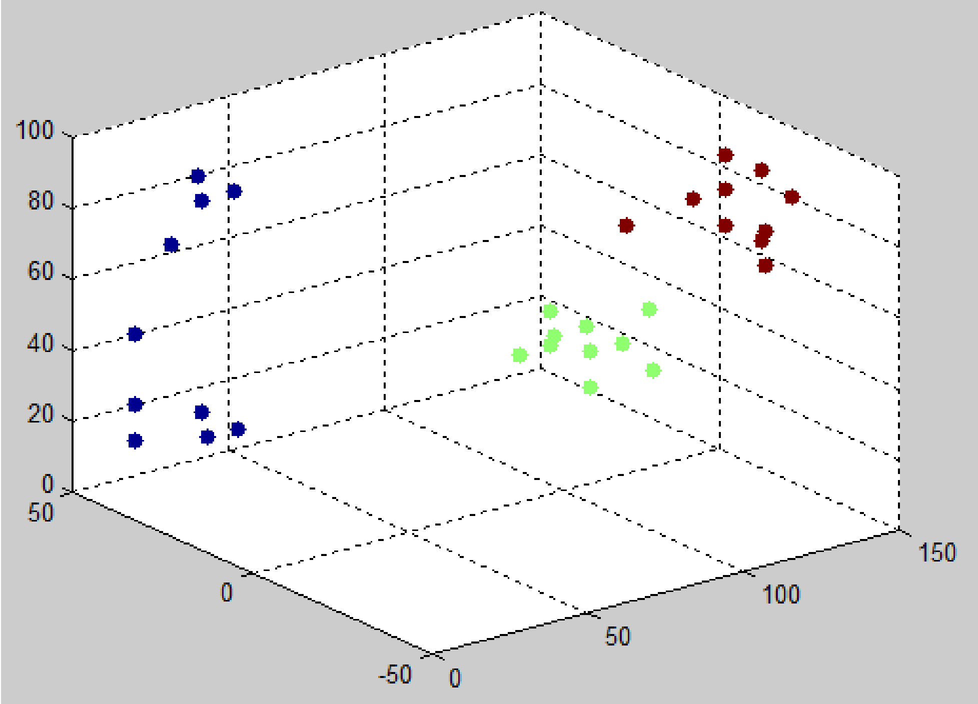
Machine Learning (IS ZC464) Session 7: Linear models for Classification

Classification

- The goal of classification is to take an input vector x and to assign it to one of K discrete classes C_k where $k = 1, 2, 3, \ldots, K$
- Examples
  - Email: Spam / Not Spam?
  - Online Transactions: Fraudulent (Yes / No)?
  - Tumor: Malignant / Benign ?

Decision Regions

- Training data is viewed to be plotted in a d-dimensional space where d is the number of features used.
- A test data is also viewed to be mapped in the same space.
- Similarity (or closeness) of the test data from the cluster of training classes is obtained.
- The nearest class is assigned to the test data



Binary Classification

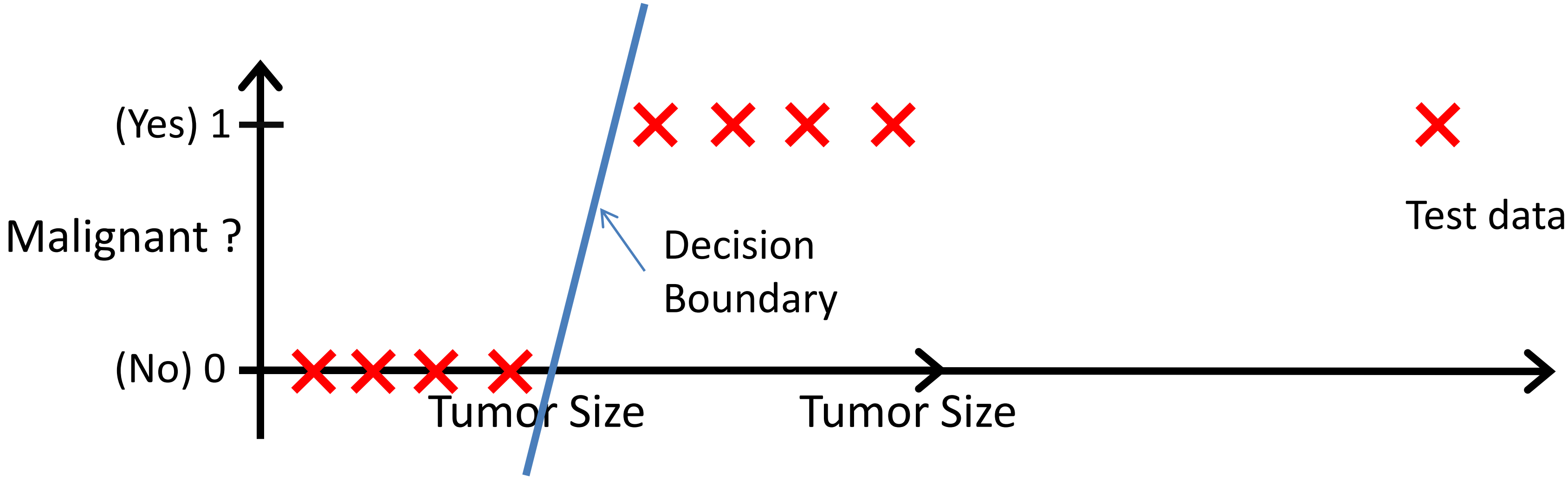
- Only two classes

$$y \in \{0, 1\}$$

0: “Negative Class” (e.g., benign tumor)

1: “Positive Class” (e.g., malignant tumor)

Example of a Decision Boundary



Threshold classifier output $h_\theta(x)$ at 0.5:

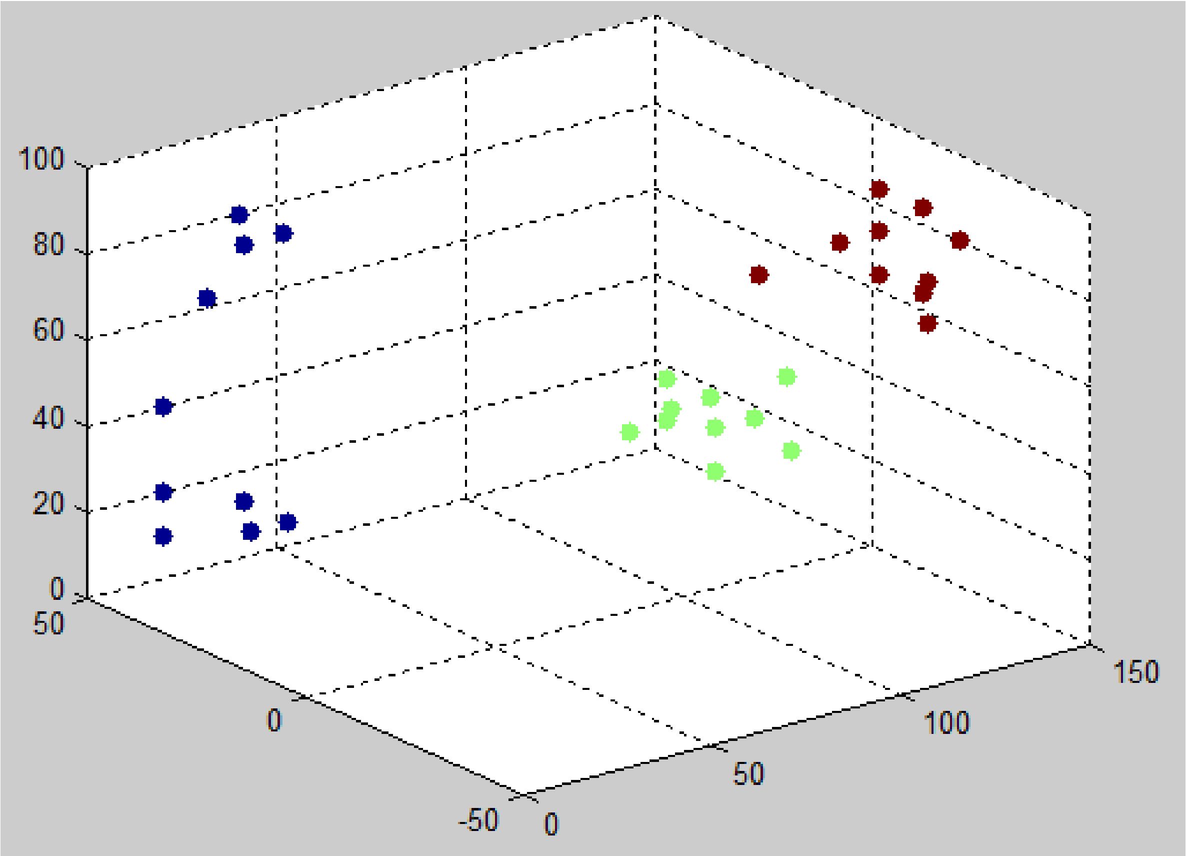
If $h_\theta(x) \geq 0.5$, predict "y = 1"

If $h_\theta(x) < 0.5$, predict "y = 0"

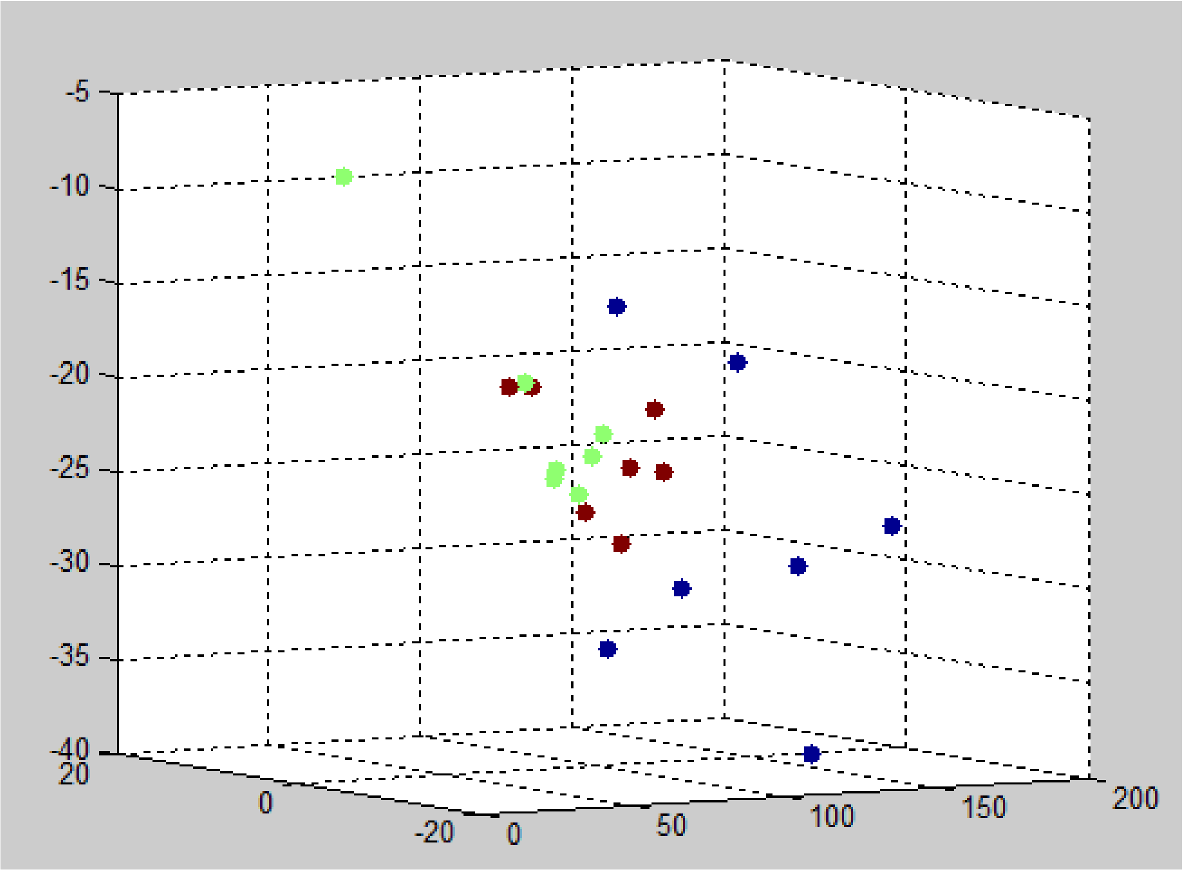
Solving Classification Problems

- Require the decision boundaries (or surfaces in hyper dimensional space) to be identified based on the training data.
- The decision boundary may be a line, a polynomial curve or a surface.
- The decision boundary can be represented as a hypothesis $h_\theta(x)$

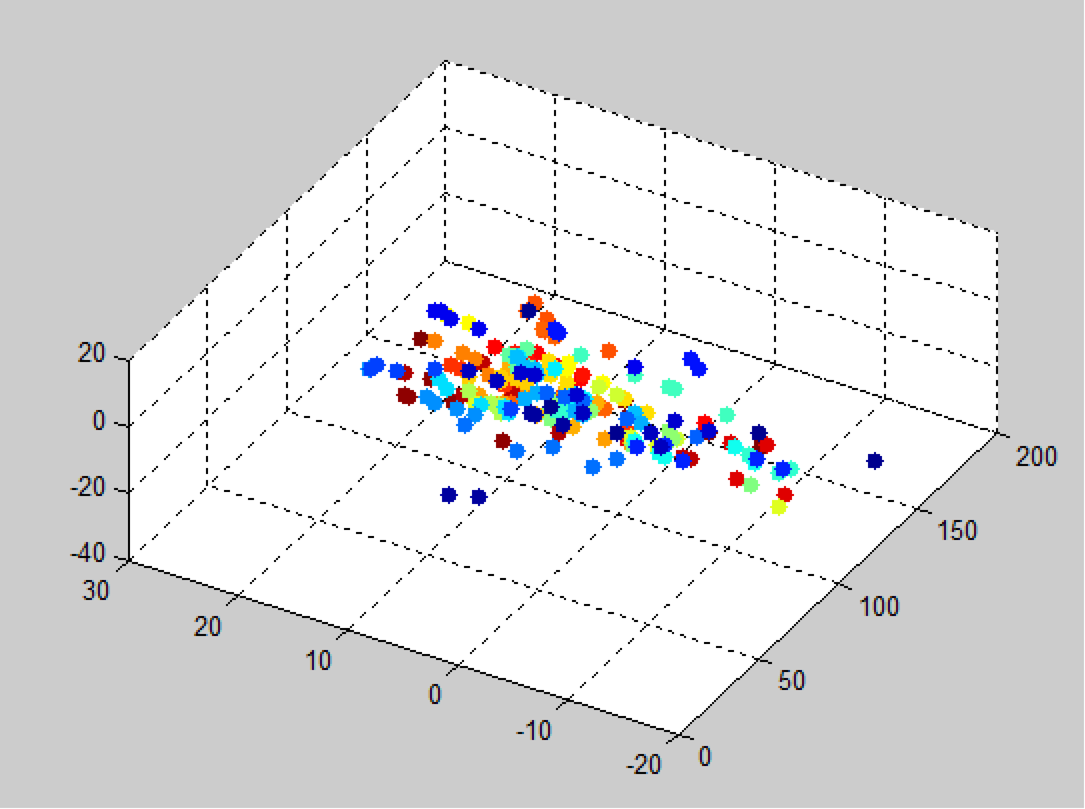
Linearly Separable Non-Face Data



Each face is a point in the n-dimensional space. (ORL face data for three persons)



The points in the n-dimensional space cannot be clustered (colorwise) by hyperplanes.



Discriminant Functions

- Represent the decision boundary
- Discriminant functions are obtained by taking a linear function of the input vector (feature vector).
- Define $y(x) = w_0 + w_1x+w_2x+\ldots w_Dx$
- Take a simple case

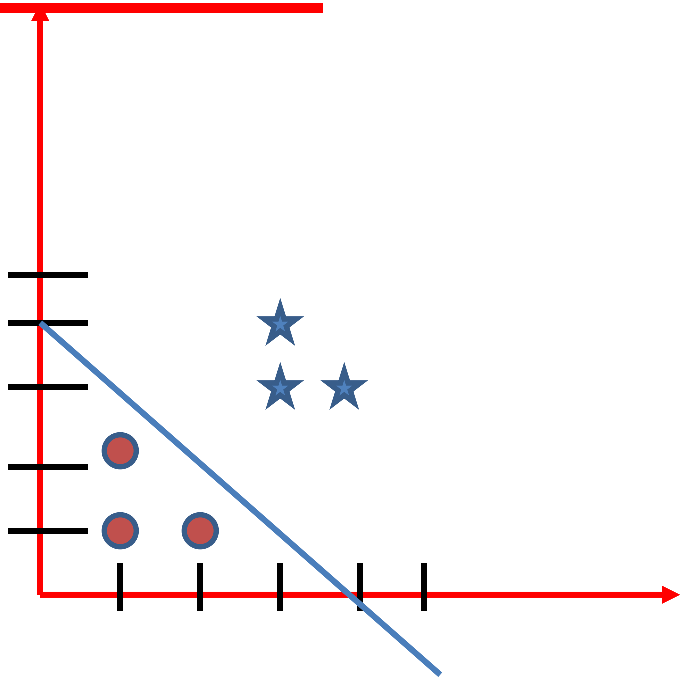
  $y(x) = w_0 + w_1x$
- This is the equation of line.
- How does this behave as a decision boundary

Example

- Consider the following training data
- Class 1: <1,2>, <1,1>, <2,1>
- Class 2: <3,3>, 3,4>, <4,3>
- Can view a decision boundary as a line separating two classes
- The equation of the line is

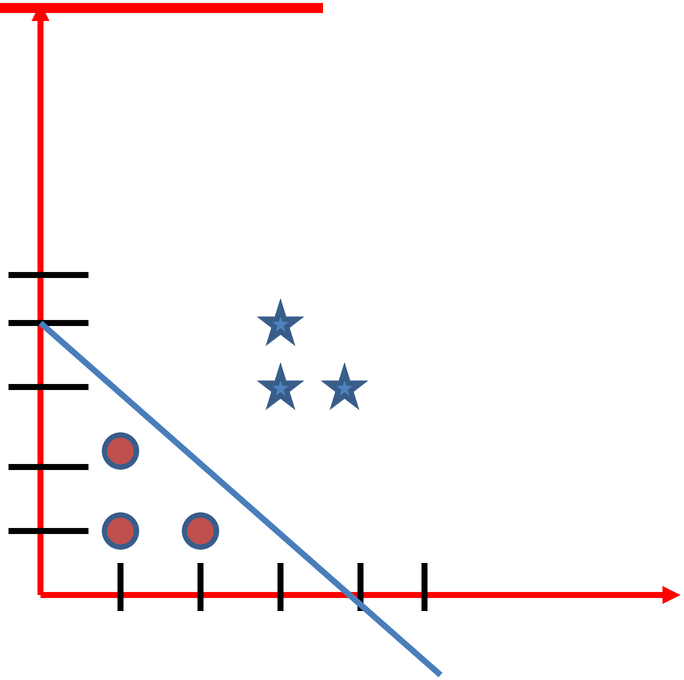
$$x_2 = -x_1 + 1$$

(not using y deliberately as used for target)



Example

- Test vector <4,4>
- Compute $h(x) = x_1+x_2-1$ as $4+4-1= 7$
- Since $h(x) > 4$, then the test data belongs to class 2
- Test vector <2,1.5>
- $h(x) = 2+1.5-1 = 2.5 <4$
- Then it belongs to class 1



Define the hypothesis in terms of vector product

$$x = \begin{bmatrix} x_0 \\ x_1 \\ .. \\ x_D \end{bmatrix} \qquad W = \begin{bmatrix} w_0 & w_1 & .. & w_D \end{bmatrix}$$

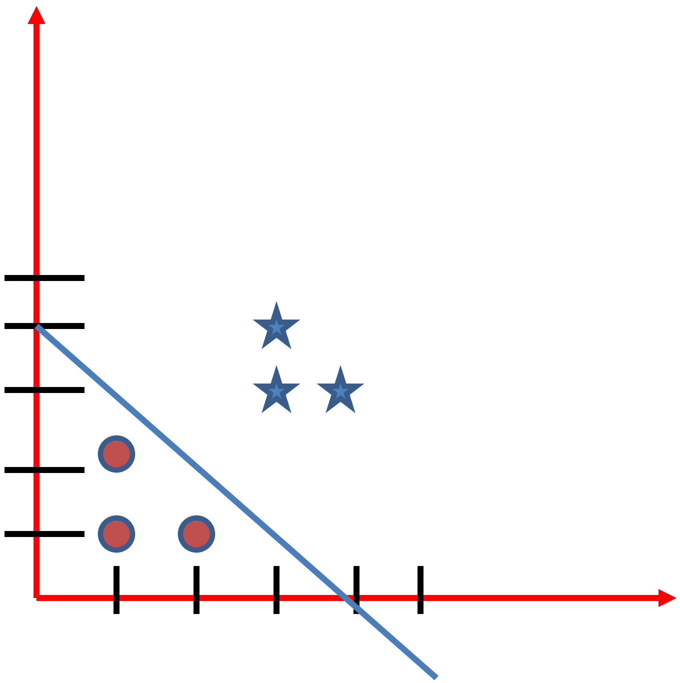
Since y (x) = $w_0 + w_1x + w_2x + \ldots w_Dx$

$$y = \mathbf{W}^T X$$

Classification

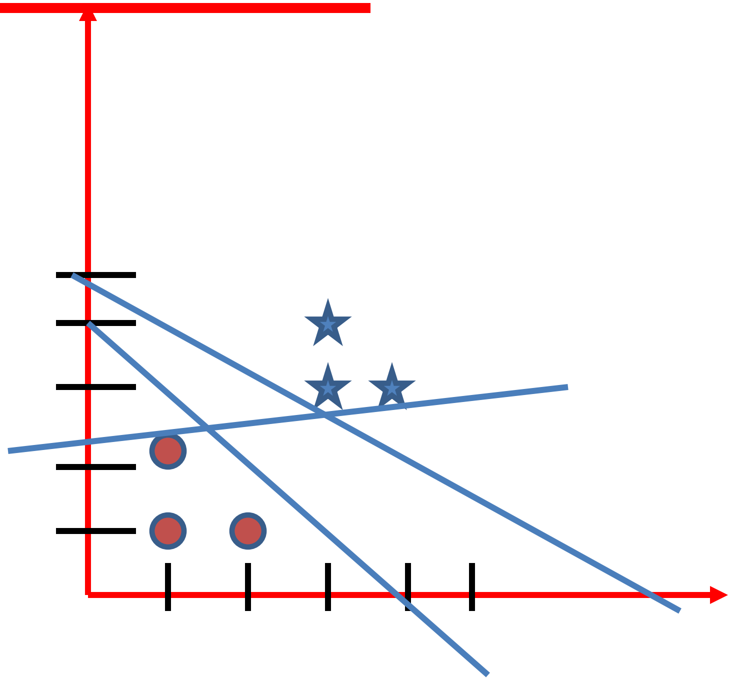
- If $W^T X \geq 0$, then the vector x belongs to class 2
- $W^T X < 0$, then the test vector belongs to class 1

- Class Assignment
- Classify
  - <5,1>
  - <4,1>
  - <3,1>



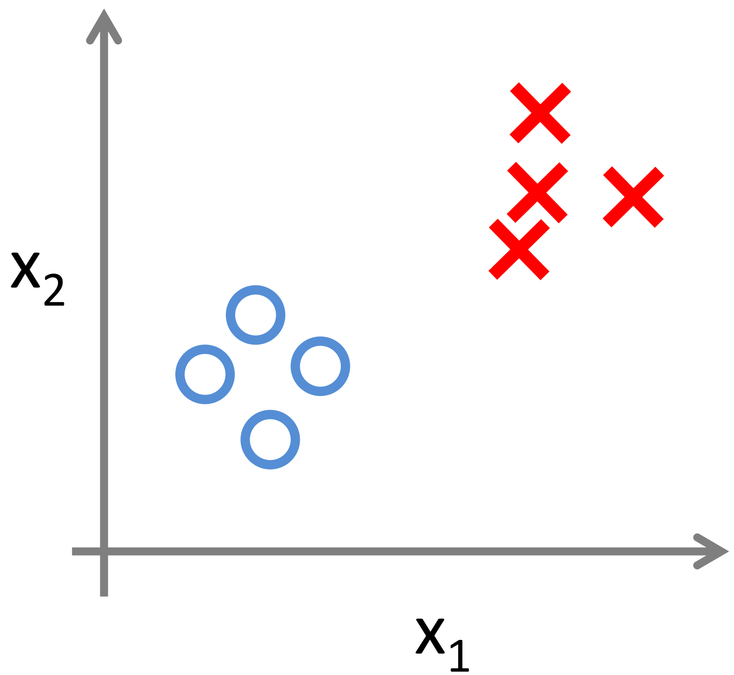
How to get the best decision boundary?

- Based of experience using training data
- We try to optimize the fitting of the decision boundary.
- If the training data s inappropriate, the classifier is likely to misclassify data.

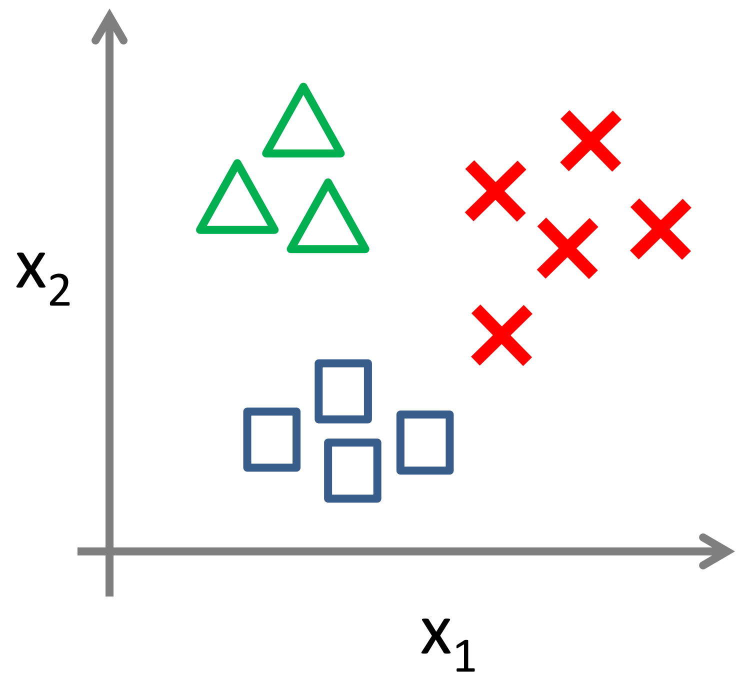


Binary versus Multi class classification

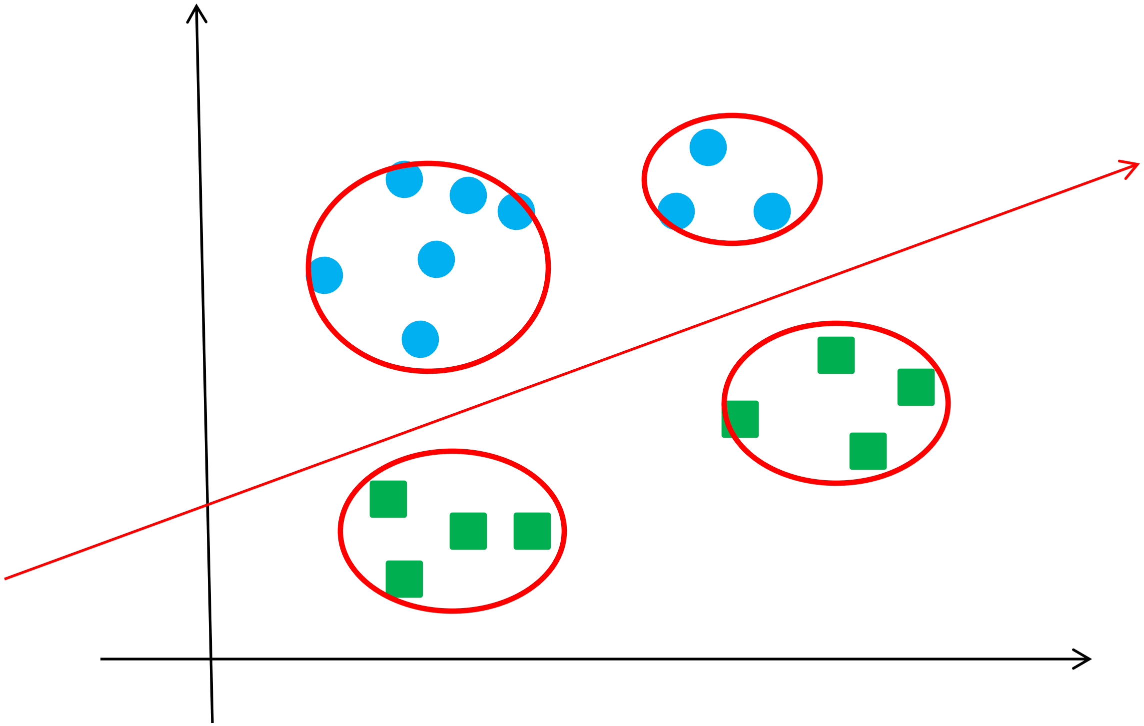
Binary classification:



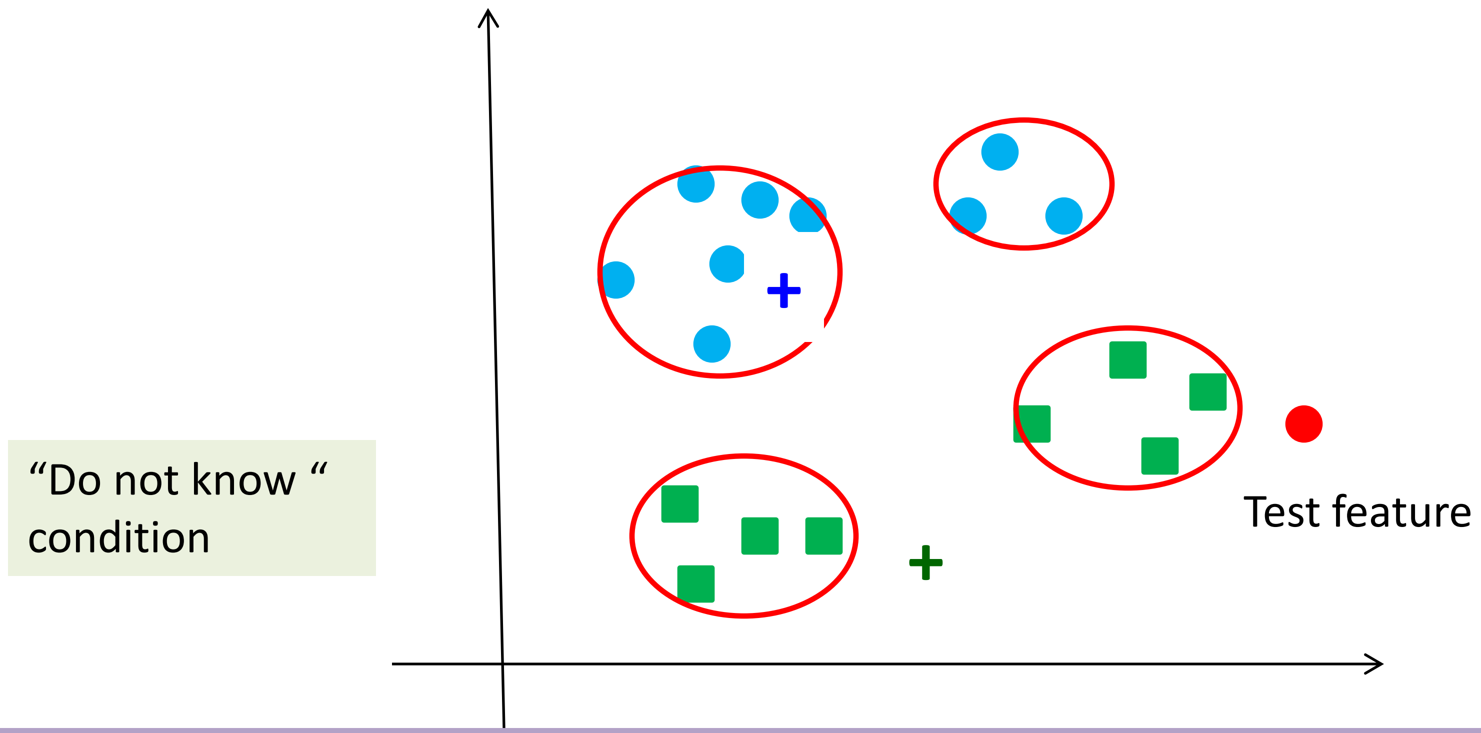
Multi-class classification:



Linear versus circular boundaries

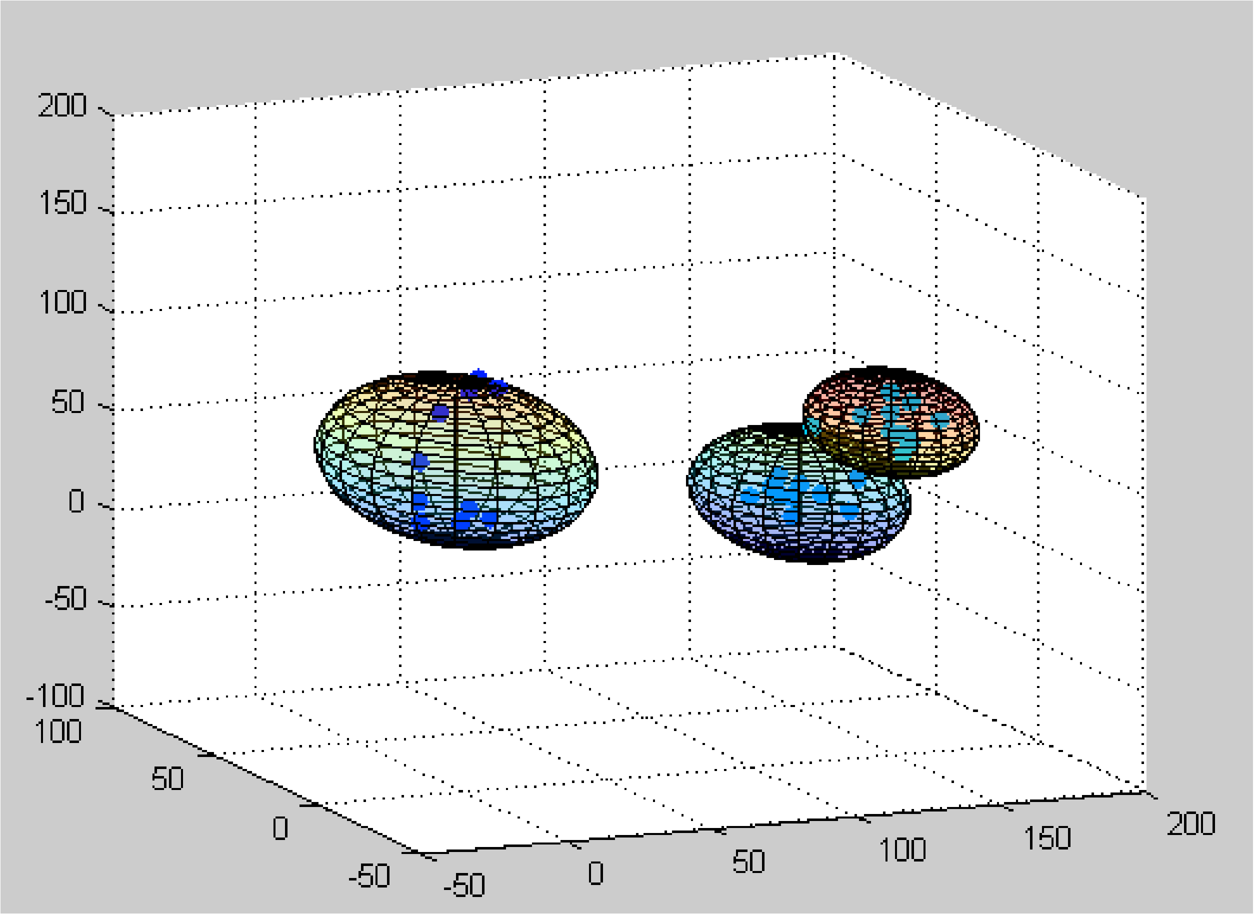


Nearest Neighbor Classification



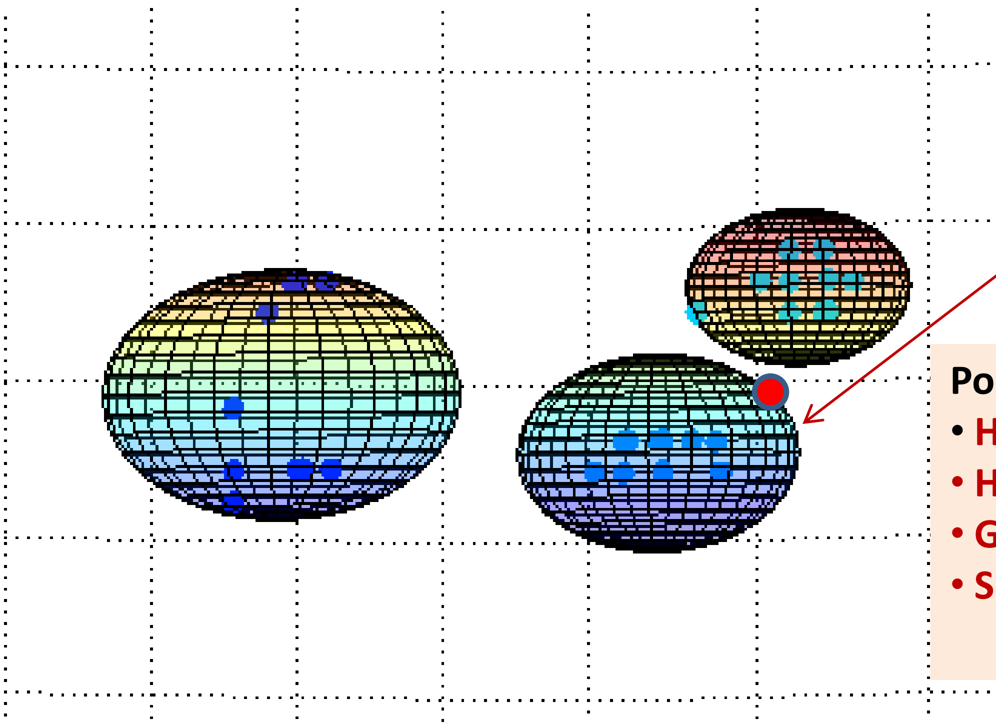
Nearest neighbor: Shortest distance to the mean of the cluster

Face data is nonlinearly separable (Hyper-Surfaces can create boundaries between clusters)



Classification Problem

Given Training Data



Closest cluster to the n-dimensional test feature vector is computed

Possible Decision Boundaries

- **Hyper Plane**
- **Hyper Sphere**
- **Gaussian Surface**
- **Support Vectors**

Challenge: Design of Decision Boundary

What to optimize ?

- Given $y(x) = w_0 + w_1x+w_2x+\dots w_Dx$
- Objective 1: Obtain W that gives minimum error of classification OR
- Objective 2: Obtain W that maximizes the separation of the classes
- Visualize the error surface discussed earlier with respect to classification error and find the parameters W that give the least error.

Multiclass classification

- Different targets
- Different thresholds
- Different boundaries

Multi-class classification: