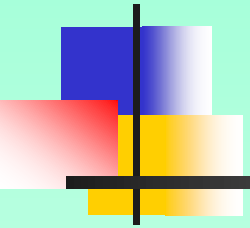
CLASSIFIERS



Classification Methods

- Classification is the process of categorizing the data into some known class-labels.
- It is a supervised process since the class labels are known in advance
- Some major classification algorithms are as follows:
  - Decision Tree classifier
  - Nearest Neighbor Classifier
  - Naïve Bayes Classifier
  - Artificial Neural Network (ANN) Based Classifier
  - Support Vector Machine (SVM)
  - Ensemble Based Classifiers

Measures for Performance Evaluation

Accuracy, sensitivity and specificity

Confusion matrix

		Predicted Class	
		Yes	**No**
Actual Class	**Yes**	TP	FN
	No	FP	TN

The true positives (TP) and true negatives (TN) are the correct classifications

A false positive (FP) is when a ‘no’ sample of a class is incorrectly classified as a ‘yes’ sample

A false negative (FN) is when a ‘yes’ sample of a class is classified as ‘no’ sample.

Measures for Performance Evaluation

We define the following measures:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

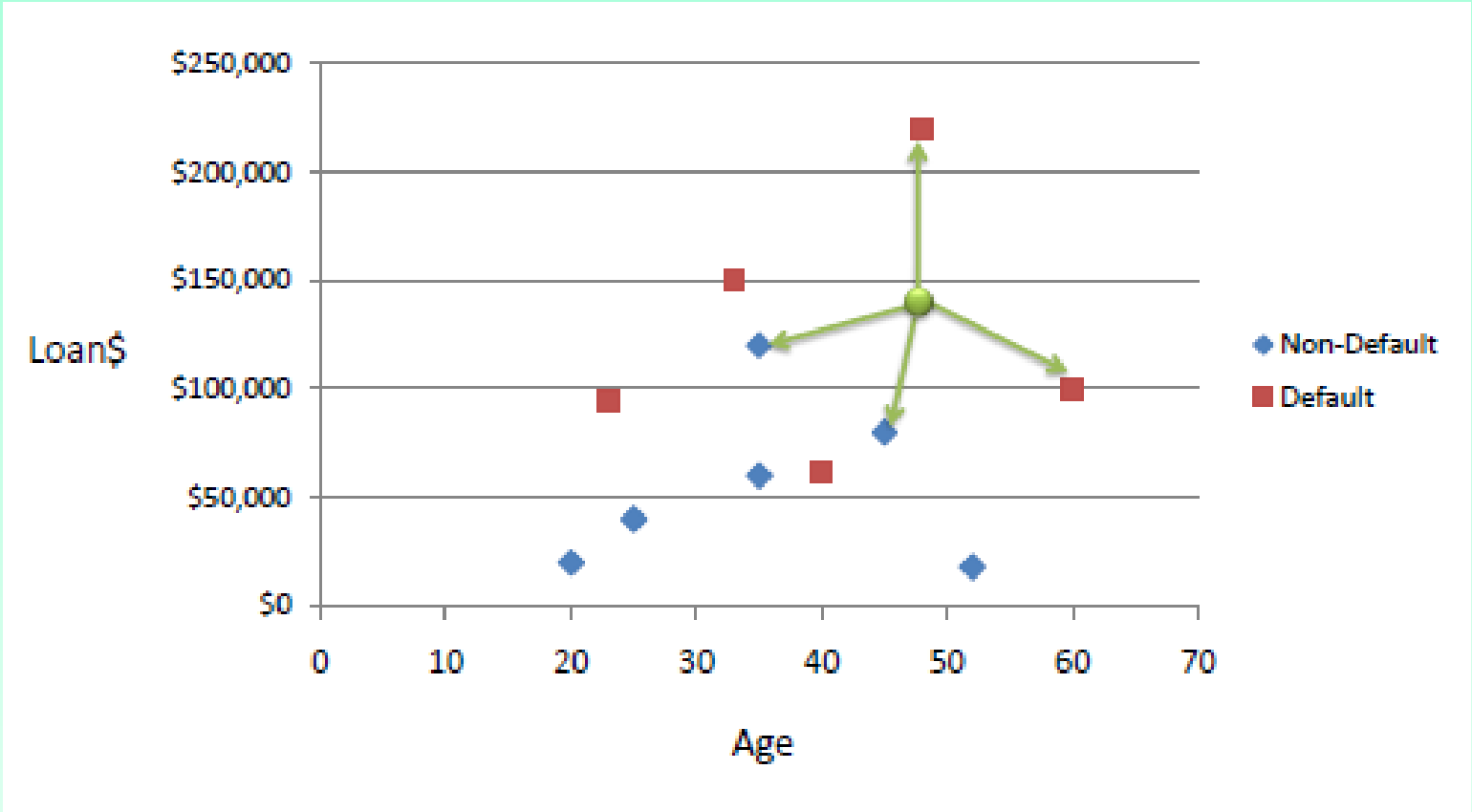
$$Specificity = \frac{TN}{TN + FP}$$

K-Nearest Neighbor Classifier

- This approach requires proximity measure and a classification function that gives the predicted class based on the proximity
- The class label of a test example is computed on the bases of its proximity to the data points in the training dataset
- The K-Nearest Neighbors are chosen for this purpose

The K-Nearest Neighbor Example

Consider the following data concerning credit default. Age and Loan are two numerical variables (predictors) and Default is the target.



The K-Nearest Neighbor Example

We can now use the training set to classify an unknown case (Age=48 and Loan=$142,000) using Euclidean distance. If K=1 then the nearest neighbor is the last case in the training set with Default=Y.

D = Sqrt[(48-33)^2 + (142000-150000)^2] = 8000.01 >> Default=Y

Age	Loan	Default	Distance	
25	$40,000	N	102000	
35	$60,000	N	82000	
45	$80,000	N	62000	
20	$20,000	N	122000	
35	$120,000	N	22000	2
52	$18,000	N	124000	
23	$95,000	Y	47000	
40	$62,000	Y	80000	
60	$100,000	Y	42000	3
48	$220,000	Y	78000	
33	$150,000	Y	8000	1
48	**$142,000**	**?**		

Euclidean Distance

$$D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

The K-Nearest Neighbor Example

With K=3, there are two Default=Y and one Default=N out of three closest neighbors. The prediction for the unknown case is again Default=Y.

Age	Loan	Default	Distance	
25	$40,000	N	102000	
35	$60,000	N	82000	
45	$80,000	N	62000	
20	$20,000	N	122000	
35	$120,000	N	22000	2
52	$18,000	N	124000	
23	$95,000	Y	47000	
40	$62,000	Y	80000	
60	$100,000	Y	42000	3
48	$220,000	Y	78000	
33	$150,000	Y	8000	1
48	**$142,000**	**?**		

Euclidean Distance

$$D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

Pros and Cons

- **Pros:**
  - Simple to implement
  - Flexible to feature / distance choices
  - Naturally handles multi-class cases
  - Can do well in practice with enough representative data
- **Cons:**
  - Large search problem to find nearest neighbours
  - We must have a meaningful distance function