Classification Model

Agenda

- **Introduction to Classification**
- Decision Tree
- Random Forest
- Bayesian
- Lazy Learner (kNN)
- Support Vector Machine
- Model Evaluation and Selection

Classification
Definition

- In machine learning and statistics, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.
- Examples are assigning a given email to the "spam" or "non-spam" class, and assigning a diagnosis to a given patient based on observed characteristics of the patient (sex, blood pressure, presence or absence of certain symptoms, etc.)

Source : Wikipedia.com

Examples of
Classification Application

- **Handwriting recognition :** used to interpret intelligible handwritten input from sources such as paper documents, photographs, touch-screens and other devices
- **Web search engine :** used to classify information on World Wide Web
- **Speech recognition :** used for recognition and translation of spoken language into text by computers.
- **Biological classification :** used for classifying biological organism on the basis of shared characteristics (taxonomy)
- **Credit scores :** used to determine who qualifies for a loan, at what interest rate, and what credit limits.

Agenda

- Introduction to Classification
- **Decision Tree Method**
- Random Forest Method
- Bayesian Method
- Lazy Learner (kNN) Method
- Support Vector Machine
- Model Evaluation and Selection

Decision Tree
Definition

- Decision tree is a tree shaped diagram used to determine a course of action. Each branch of the tree represents a possible decision, occurrence or reaction

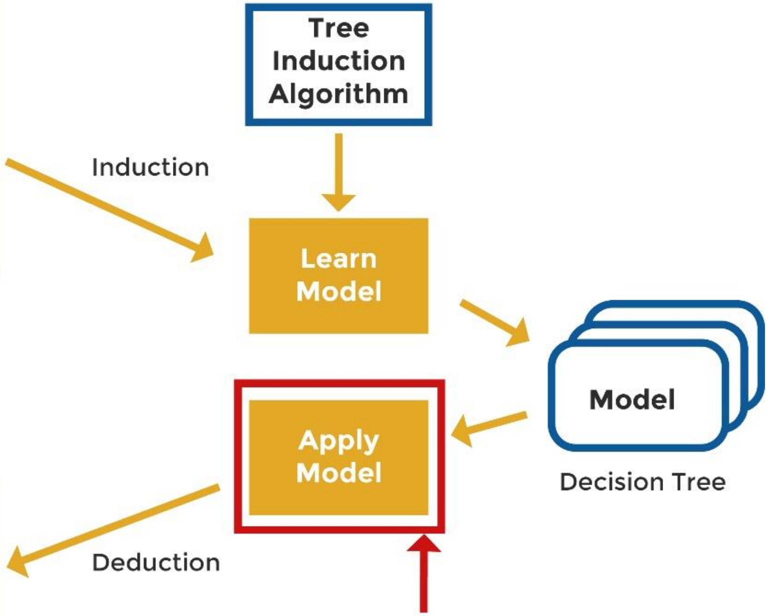
- Important Terms
  - Entropy : measure of randomness in the dataset
  - Information gain : measure of decrease in entropy after dataset is split
  - Leaf node : carries the classification or decision
  - Root node: top most decision node

Decision Tree

Classification Task

TID	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set



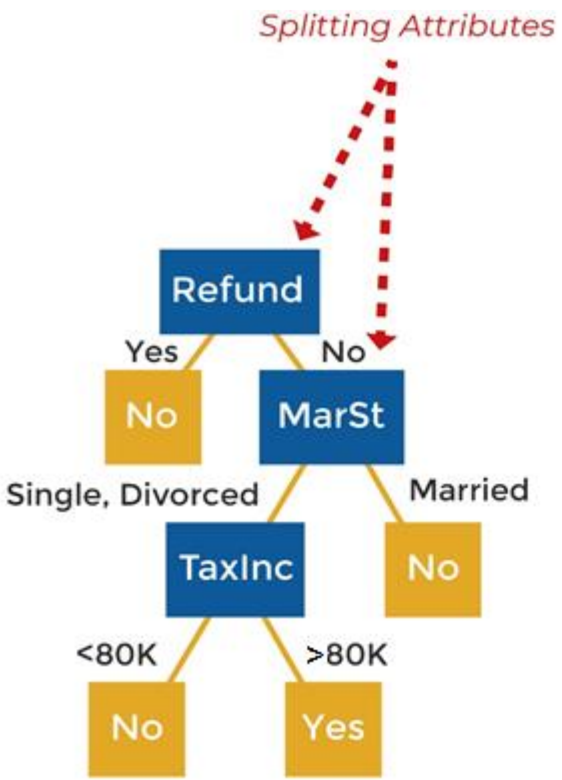
TID	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

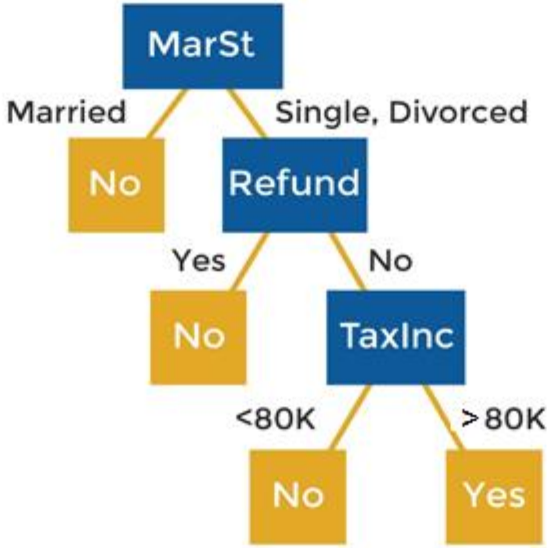
Test Set

Example of
a Decision Tree

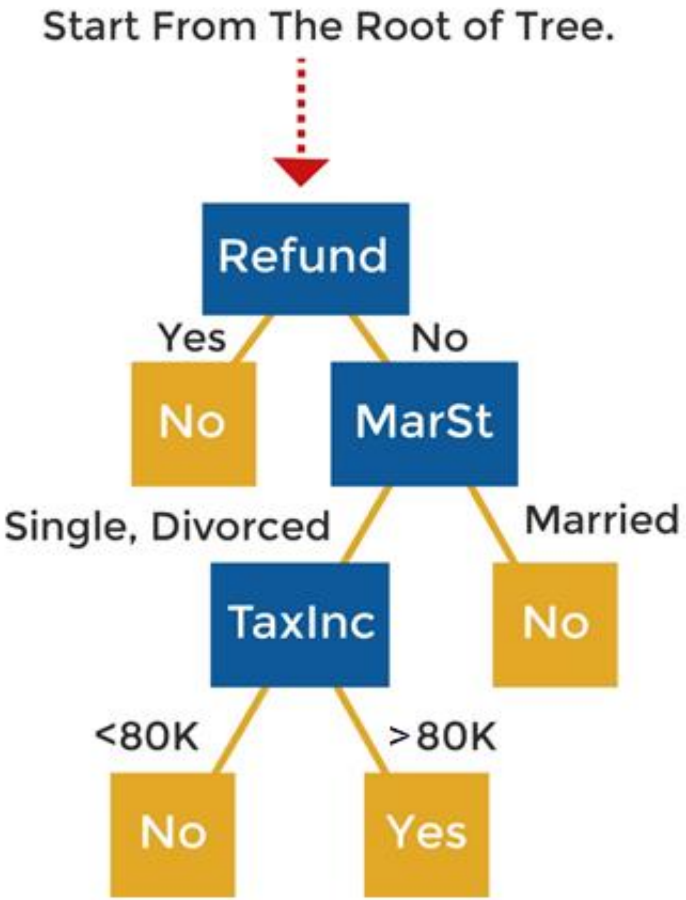
	Categorical	*Categorical*	*Continuous*	*Class*
TID	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Data





Apply Model to
Test Data



Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Decision Tree Induction

- Many Algorithms:
  - Hunt’s Algorithm (one of the earliest)
  - CART
  - ID3, C4.5
  - SLIQ,SPRINT
- Greedy strategy.
  - Split the records based on an attribute test that optimizes certain criterion.
- Issues
  - Determine how to split the records
  - How to specify the attribute test condition?
  - How to determine the best split?
  - Determine when to stop splitting

How to Specify
Attribute Test Condition?

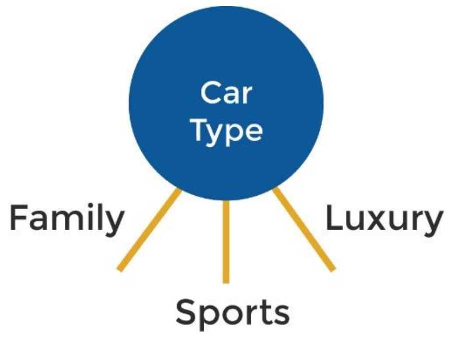
- Depends on attribute types
  - Discrete
    - Nominal
    - Ordinal
  - Continuous

- Depends on number of ways to split
  - Multi-way split
  - Binary split

Splitting Based on Nominal Attributes

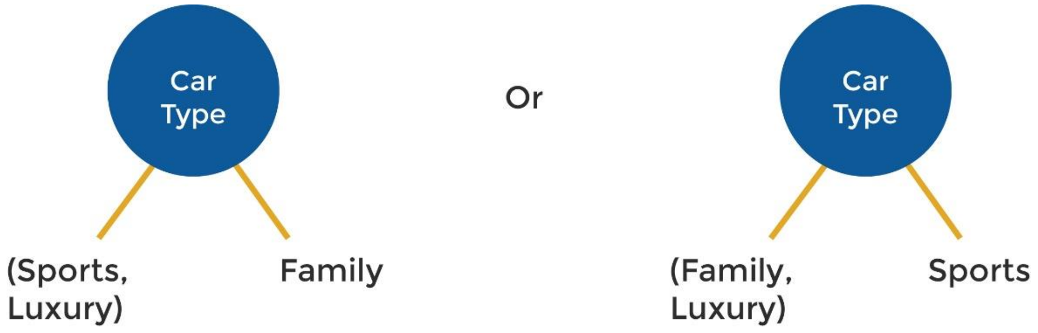
Multi-way split:

Use as many partitions as distinct values.



Binary split:

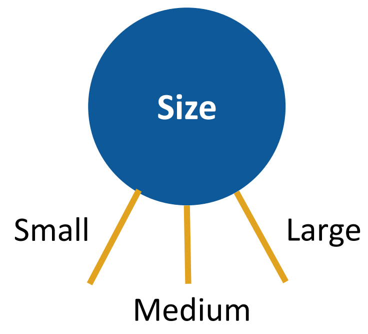
- Divides values into two subsets
- Need to find optimal partitioning



Splitting Based on Ordinal Attributes

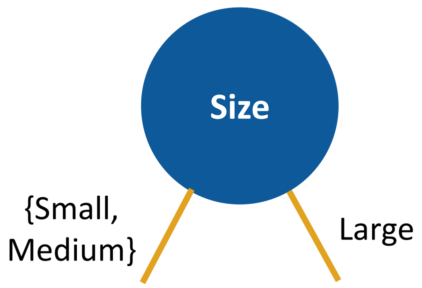
Multi-way split:

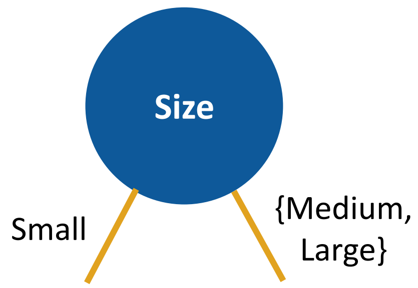
Use as many partitions as distinct values.



Binary split:

- Divides values into two subsets
- Need to find optimal partitioning

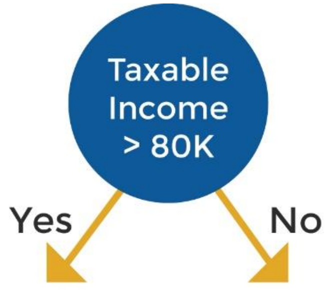




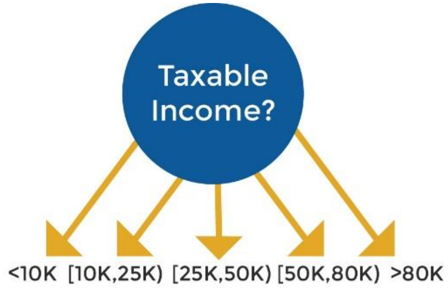
Splitting Based on Continuous Attributes

- Different ways of handling
  - Discretization to form an ordinal categorical attribute
    - Static – discretize once at the beginning
    - Dynamic – ranges can be found by equal interval bucketing, equal frequency bucketing (percentiles), or clustering.

- Binary Decision: $(A < v)$ or $(A \geq v)$
  - consider all possible splits and finds the best cut
  - can be more compute intensive



(i) Binary Split



(ii) Multi-way Split

Decision Tree
Summary

- Advantages:
  - Inexpensive to construct
  - Extremely fast at classifying unknown records
  - Easy to interpret for small-sized trees
  - Accuracy is comparable to other classification techniques for many simple data sets

- Disadvantages:
  - Over-fitting when algorithm capture noise in the data
  - The model can get unstable due to small variation of data
  - Low biased tree: difficult for the model to work with new data

Agenda

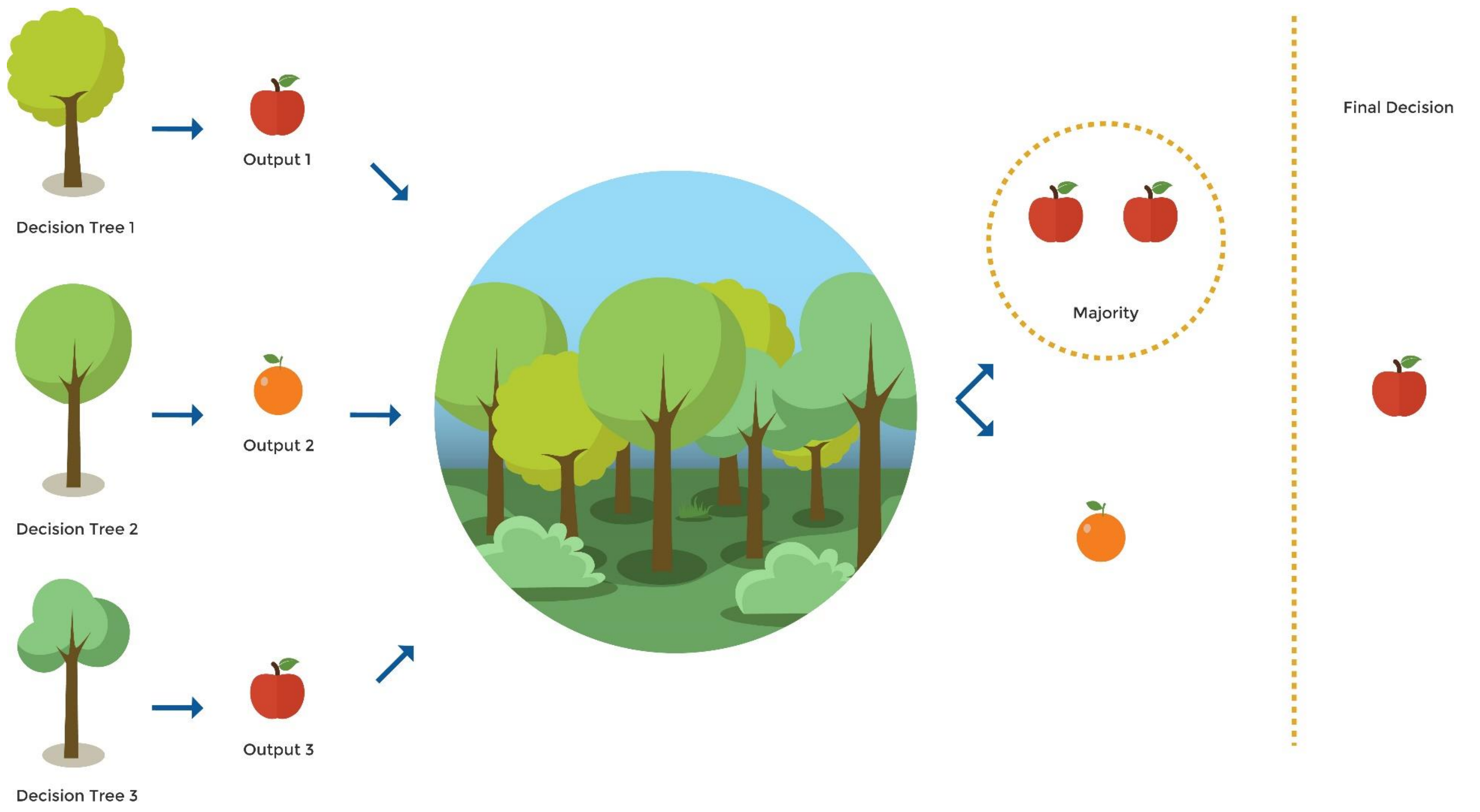
- Introduction to Classification
- Decision Tree
- **Random Forest**
- Bayesian
- Lazy Learner (kNN)
- Support Vector Machine
- Model Evaluation and Selection

Random Forest
Definition

- Random forest or Random Decision Forest is a method that operates by constructing multiple decision trees during training phases

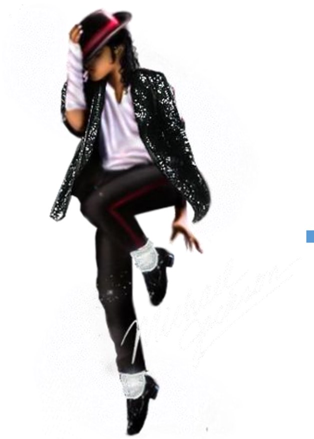
- The Decision of the majority of the trees is chosen by the random forest as final decision

Illustration of Random Forest



Classification Model

Application of
Random Forest



User Perfoms a Step

Kinect Registers the
Movement

Marks the User Based on
Accuracy

Training Set to Identify Body Parts

Random Forest Classifier Learns

Identifies the Body Parts
While Dancing

Score Game Avatar Based on Accuracy

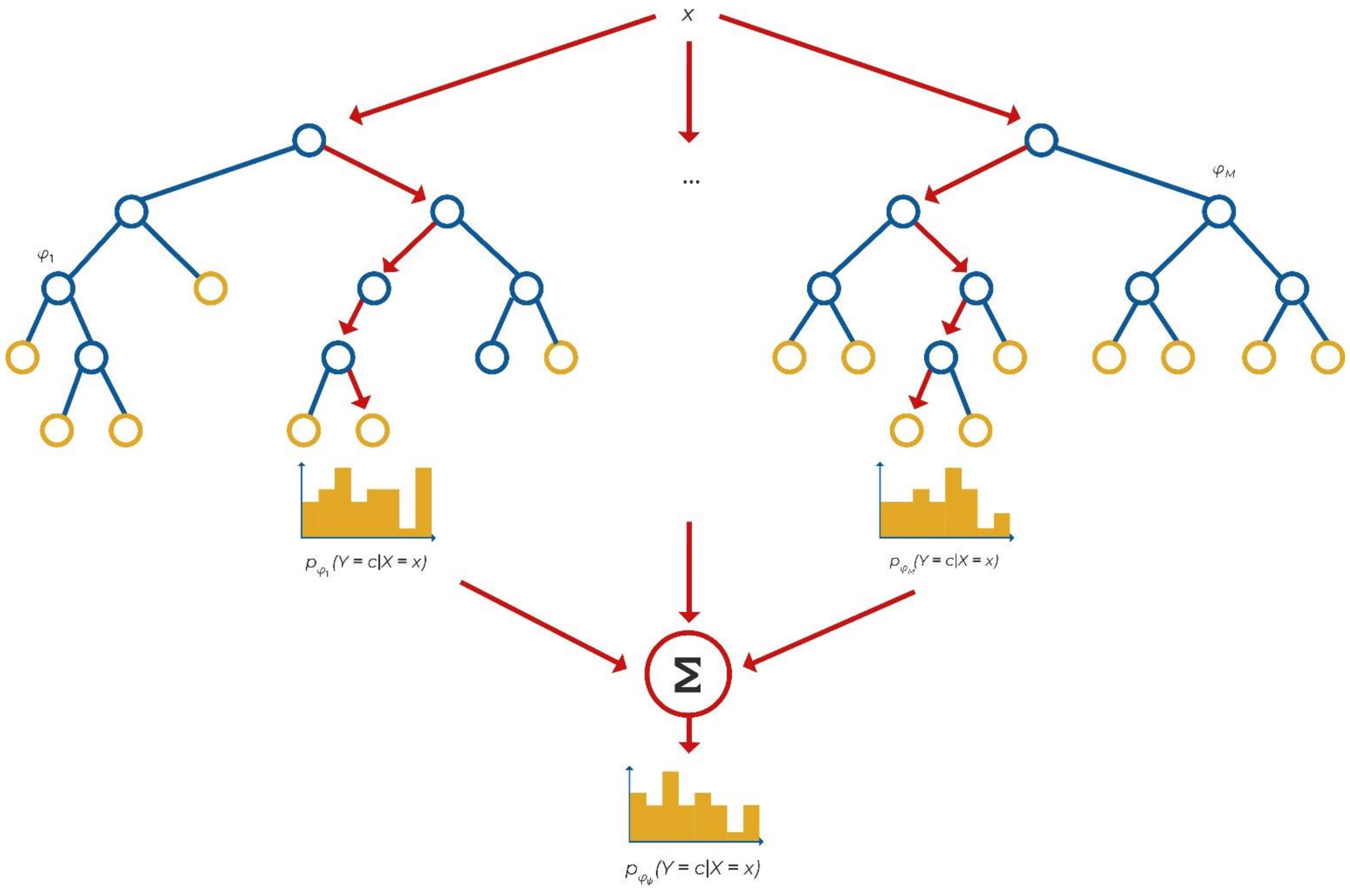
Classification Model

Random Forests

- Randomization
  - Bootstrap Samples
  - Random Selection of $K \leq p$ split variables
  - Random Selection of the Treshold

Random Forests

Extra Trees



Random Forest
(Breiman 2001)

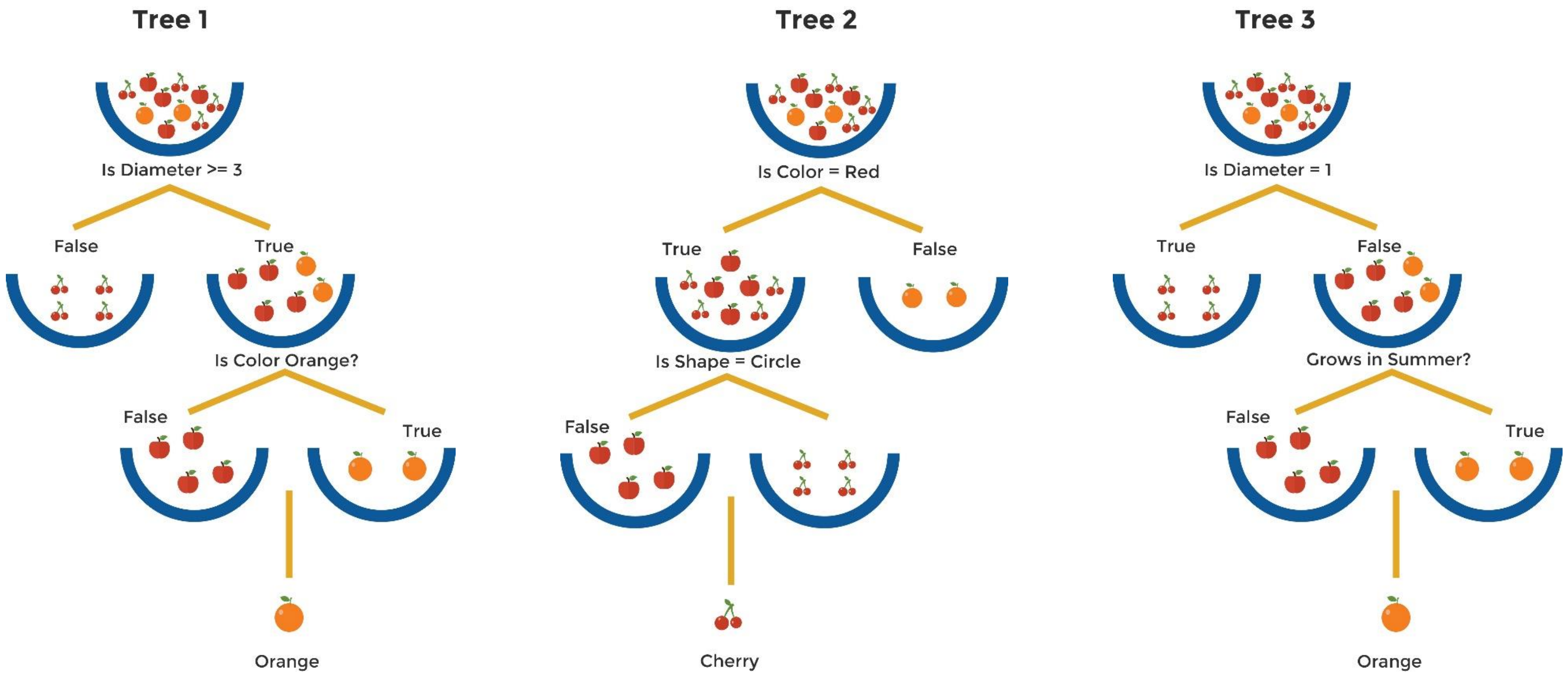
- Random Forest
  - Each classifier in the ensemble is a *decision tree* classifier and is generated using a random selection of attributes at each node to determine the split
  - During classification, each tree votes and the most popular class is returned

- Two Methods to construct Random Forest
  - Forest-RI (*random input selection*): Randomly select, at each node, F attributes as candidates for the split at the node. The CART methodology is used to grow the trees to maximum size
  - Forest-RC (*random linear combinations*): Creates new attributes (or features) that are a linear combination of the existing attributes (reduces the correlation between individual classifiers)

- Insensitive to the number of attributes selected for consideration at each split, and faster than bagging (grouping based on frequency) or boosting

How Does
Random Forest Work?

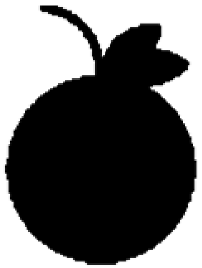
- We have 3 trees in the forest…



Major Vote is ORANGE

How Does
Random Forest Work?

- What fruit is this?



Diameter : 3
Colour : Orange
Grows in summer : Yes
Shape : Circle

- From Tree 1, we classify it as ORANGE
- From Tree 2, we classify it as CHERRY
- From Tree 3, we classify it as ORANGE

- Majority voted as ORANGE, so we classify it as ORANGE

Random Forest Summary

- Advantages:
  - It can be used for both regression and classification tasks and that it’s easy to view the relative importance it assigns to the input features
  - It is also considered as a very handy and easy to use algorithm, because it’s default hyper-parameters often produce a good prediction result

- Disadvantages:
  - a large number of trees can make the algorithm to slow and ineffective for real-time predictions. A more accurate prediction requires more trees, which results in a slower model
  - It is a predictive modeling tool and not a descriptive tool

Agenda

- Introduction to Classification
- Decision Tree
- Random Forest
- **Bayesian**
- Lazy Learner (kNN)
- Support Vector Machine
- Model Evaluation and Selection

Bayesian Classifier
Definition

- A probabilistic framework for solving classification problems

- Conditional Probability:

$$P(C \mid A) = \frac{P(A,C)}{P(A)}$$

$$P(A \mid C) = \frac{P(A,C)}{P(C)}$$

- Bayes theorem:

$$P(C \mid A) = \frac{P(A \mid C)P(C)}{P(A)}$$

Examples of Bayes Theorem

- Given:
  - A doctor knows that meningitis causes stiff neck 50% of the time
  - Prior probability of any patient having meningitis is 1/50,000
  - Prior probability of any patient having stiff neck is 1/20

- If a patient has stiff neck, what's the probability he/she has meningitis?

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

Bayesian Classifier

- Consider each attribute and class label as random variables

- Given a record with attributes (A_1, A_2,…,A_n)
  - Goal is to predict class C
  - Specifically, we want to find the value of C that maximizes P(C| A_1, A_2,…,A_n)

- Can we estimate P(C| A_1, A_2,…,A_n) directly from data?

Bayesian Classifier

- Approach:
  - compute the posterior probability $P(C \mid A_1, A_2, \ldots, A_n)$ for all values of C using the Bayes theorem

$$P(C \mid A_1 A_2 \ldots A_n) = \frac{P(A_1 A_2 \ldots A_n \mid C) P(C)}{P(A_1 A_2 \ldots A_n)}$$

  - Choose value of C that maximizes $P(C \mid A_1, A_2, \ldots, A_n)$
  - Equivalent to choosing value of C that maximizes $P(A_1, A_2, \ldots, A_n \mid C)\, P(C)$
- How to estimate $P(A_1, A_2, \ldots, A_n \mid C)$?

Naïve Bayes
Classifier

Assume independence among attributes A_i when class is given:

- $P(A_1, A_2, \ldots, A_n | C) = P(A_1 | C_j)\, P(A_2 | C_j) \ldots P(A_n | C_j)$
- Can estimate $P(A_i | C_j)$ for all A_i and C_j.
- New point is classified to C_j if $P(C_j) \prod P(A_i | C_j)$ is maximal.

How to Estimate Probabilities from Data?

- Class: $P(C) = Nc/N$

  e.g., $P(No) = 7/10$,

  $P(Yes) = 3/10$

- For discrete attributes:

  $P(Ai \mid Ck) = |Aik| / Nc$

  - where $|Aik|$ is number of instances having attribute Ai and belongs to class Ck

  Examples:
  $P(Status=Married|No) = 4/7$
  $P(Refund=Yes|Yes)=0$

TID	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

How to Estimate
Probabilities from Data?

TID	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

$$P(Income = 120 \mid No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

- Normal distribution:

$$P(A_i \mid c_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(A_i-\mu_{ij})^2}{2\sigma_{ij}^2}}$$

  - One for each (Ai,ci) pair

- For (Income, Class=No):
  - If Class=No
    - sample mean = 110
    - sample variance = 2975

Example of
Naïve Bayes Classifier

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
Human	Yes	No	No	Yes	Mammals
Phyton	No	No	No	No	Non-Mammals
Salmon	No	No	Yes	No	Non-Mammals
Whale	Yes	No	Yes	No	Mammals
Frog	No	No	Sometimes	Yes	Non-Mammals
Komodo	No	No	No	Yes	Non-Mammals
Bat	Yes	Yes	No	Yes	Mammals
Pigeon	No	Yes	No	Yes	Non-Mammals
Cat	Yes	No	No	Yes	Mammals
Leopard Shark	Yes	No	Yes	No	Non-Mammals
Turtle	No	No	Sometimes	Yes	Non-Mammals
Penguin	No	No	Sometimes	Yes	Non-Mammals
Porcupine	Yes	No	No	Yes	Mammals
Eel	No	No	Yes	No	Non-Mammals
Salamander	No	No	Sometimes	Yes	Non-Mammals
Gila Monster	No	No	No	Yes	Non-Mammals
Platypus	No	No	No	Yes	Mammals
Owl	No	Yes	No	Yes	Non-Mammals
Dolphin	Yes	No	Yes	No	Mammals
Eagle	No	Yes	No	Yes	Non-Mammals

Give Birth	Can Fly	Live in Water	Have Legs	Class
Yes	No	Yes	No	?

- A: Attributes
- M: Mammals
- N: Non-Mammals

$$P(A|M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A|N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A|M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A|N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

P(A|M)P(M) > P(A|N)P(N)

=> Mammals

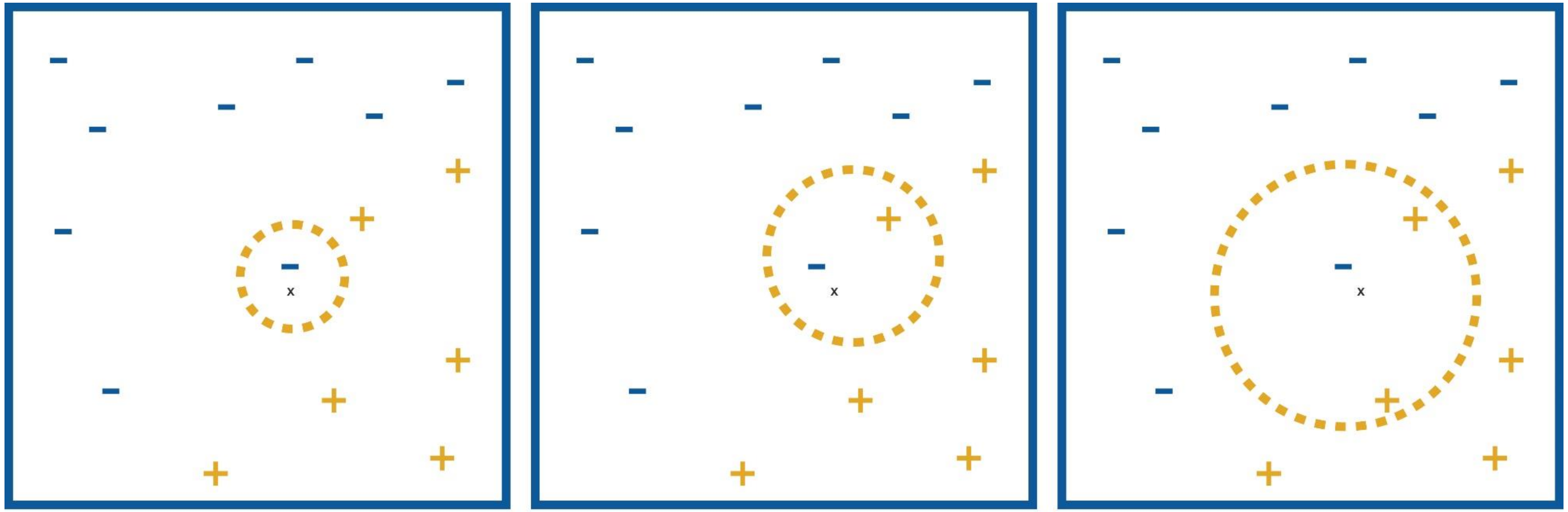
Naïve Bayes
Summary

- Robust to isolated noise points
- Handle missing values by ignoring the instance during probability estimate calculations
- Robust to irrelevant attributes
- Independence assumption may not hold for some attributes
  - Use other techniques such as Bayesian Belief Networks (BBN)

Agenda

- Introduction to Classification
- Decision Tree
- Random Forest
- Bayesian
- **Lazy Learner (kNN)**
- Support Vector
- Model Evaluation and Selection

Nearest Neighbor
Definition



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

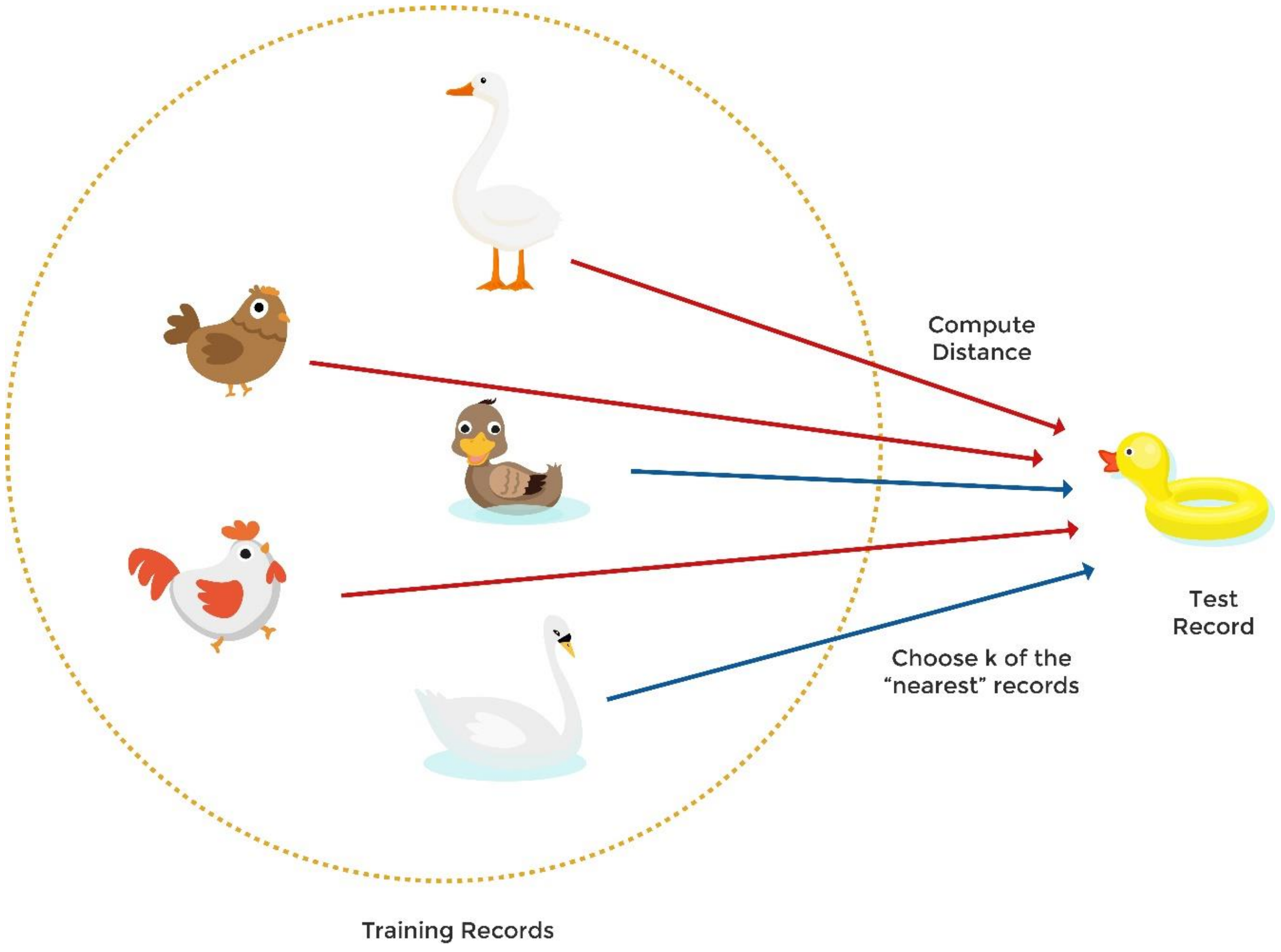
K-nearest neighbors of a record x are data points that have the k smallest distance to x

Lazy Learner (kNN) Technique

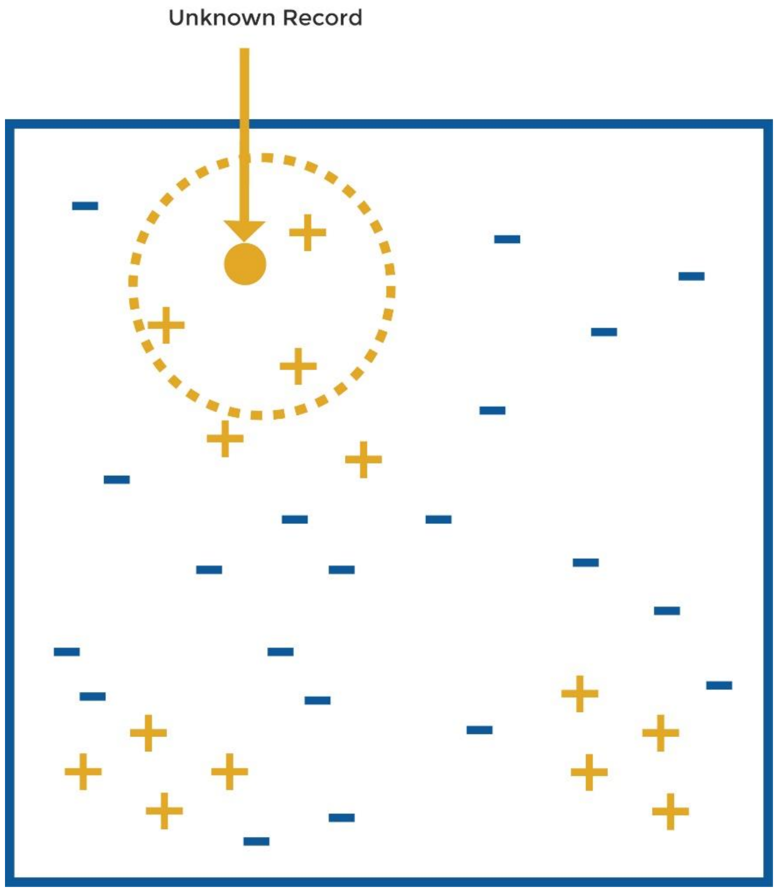
Nearest Neighbor Classifiers

- Basic idea:

  If it walks like a duck, quacks like a duck, then it's probably a duck



Nearest-Neighbor Classifiers

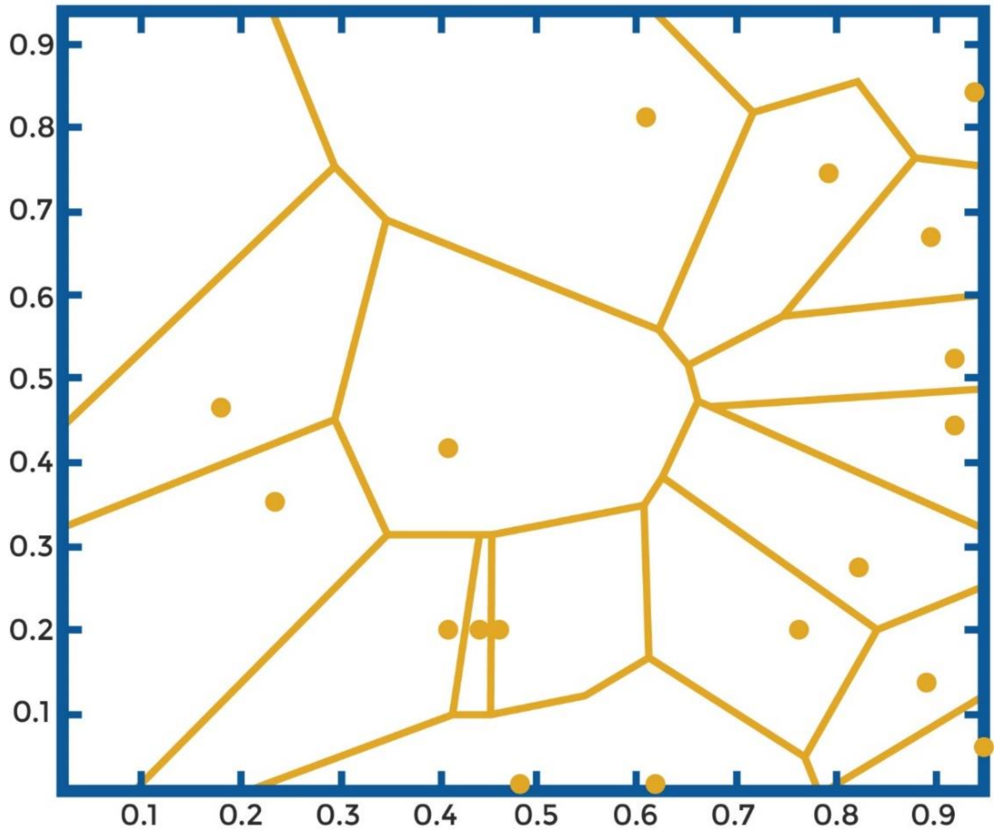


- Requirement
  - The set of stored records
  - Distance Metric to compute distance between records
  - The value of k, the number of nearest neighbors to retrieve

- To classify an unknown record:
  - Compute distance to other training records
  - Identify k nearest neighbors
  - Use class labels of nearest neighbors to determine the class label of unknown record (e.g., by taking majority vote)

1 Nearest-Neighbor

Voronoi Diagram



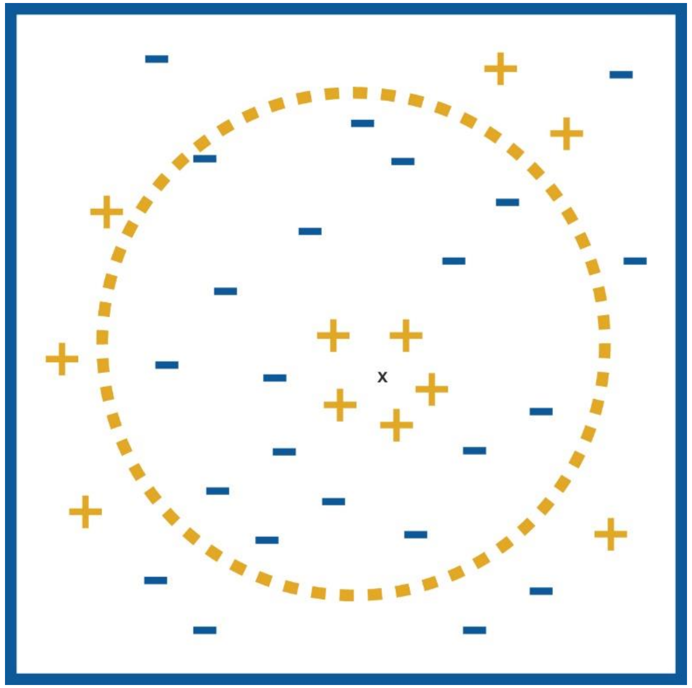
Nearest-Neighbor Classification

- Compute distance between two points:
  - Euclidean distance

$$d(p,q) = \sqrt{\sum_i (p_i - q_i)^2}$$

- Determine the class from nearest neighbor list
  - take the majority vote of class labels among the k-nearest neighbors
  - Weigh the vote according to distance
    - weight factor, w = 1/d2

Nearest-Neighbor Classification



- Choosing the value of k:
  - If k is too small, sensitive to noise points
  - If k is too large, neighborhood may include points from other classess
- k-NN classifiers are lazy learners
  - It does not build models explicitly
  - Unlike eager learners such as decision tree induction and rule-based systems
  - Classifying unknown records are relatively expensive

Example:
PEBLS

- PEBLS: Parallel Exemplar-Based Learning System (Cost & Salzberg)
  - Works with both continuous and nominal features
    - For nominal features, distance between two nominal values is computed using modified value difference metric (MVDM)
  - Each record is assigned a weight factor
  - Number of nearest neighbor, k = 1

Example:
PEBLS

TID	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Class	Marital Status		
	Single	Married	Divorced
Yes	2	0	1
No	2	4	1

Class	Refund	
	Single	Married
Yes	0	3
No	3	4

Distance between nominal attribute values:

d(Single,Married)

= | 2/4 – 0/4 | + | 2/4 – 4/4 | = 1

d(Single,Divorced)

= | 2/4 – 1/2 | + | 2/4 – 1/2 | = 0

d(Married,Divorced)

= | 0/4 – 1/2 | + | 4/4 – 1/2 | = 1

d(Refund=Yes,Refund=No)

= | 0/3 – 3/7 | + | 3/3 – 4/7 | = 6/7

$$d(V_1,V_2) = \sum_i \left| \frac{n_{1i}}{n_1} - \frac{n_{2i}}{n_2} \right|$$

Example:
PEBLS

TID	Refund	Marital Status	Taxable Income	Cheat
X	Yes	Single	125K	No
Y	No	Married	100K	No

Distance between record X and record Y:

$$\Delta(X,Y) = w_X w_Y \sum_{i=1}^{d} d(X_i,Y_i)^2$$

where: $$w_X = \frac{\text{Number of times X is used for prediction}}{\text{Number of times X predicts correctly}}$$

$w_X \cong 1$ if X makes accurate prediction most of the time

$w_X > 1$ if X is not reliable for making predictions

K-NN
Summary

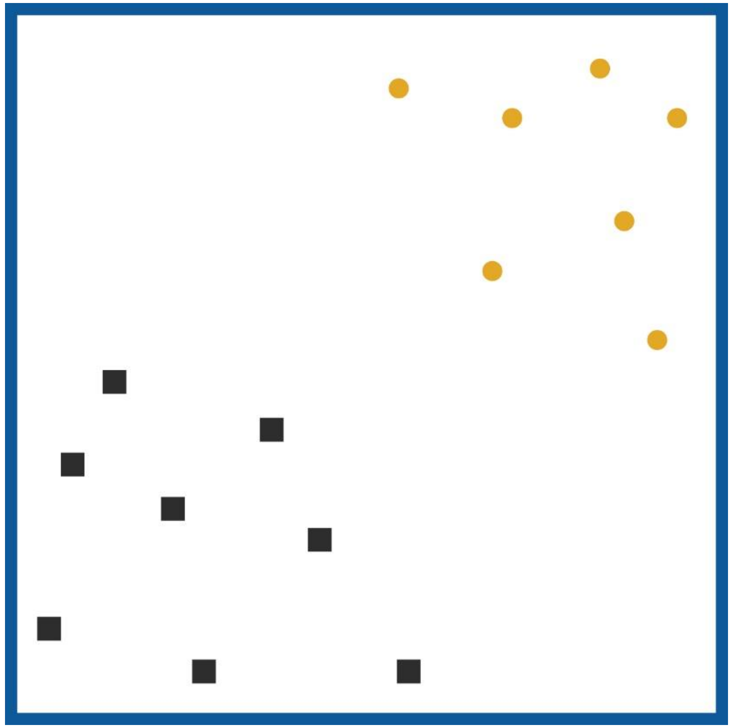
- Advantages:
  - Simple technique that is easily implemented
  - Building model is cheap
  - Extremely flexible classification scheme
  - Well suited for
    - Multi-modal classes
    - Records with multiple class labels
  - Error rate at most twice that of Bayes error rate

- Disadvantages:
  - Classifying unknown records are relatively expensive
    - Requires distance computation of k-nearest neighbors
    - Computationally intensive, especially when the size of the training set grows
  - Accuracy can be severely degraded by the presence of noisy or irrelevant features

Agenda

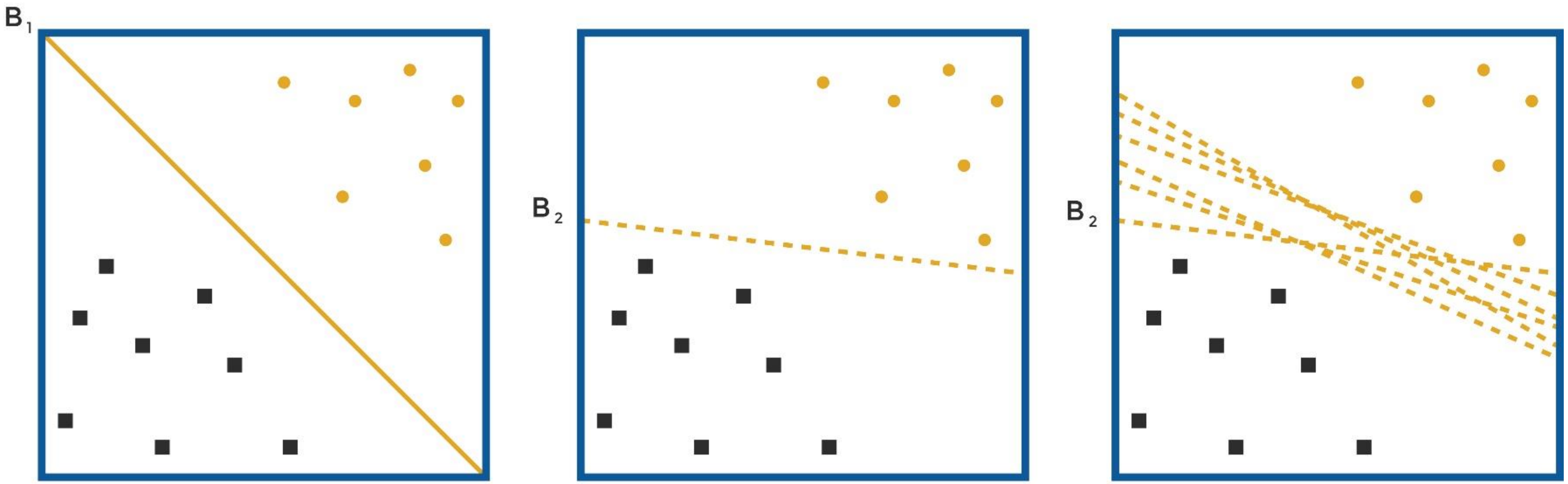
- Introduction to Classification
- Decision Tree
- Random Forest
- Bayesian
- Lazy Learner (kNN)
- Support Vector Machine
- Model Evaluation and Selection

Support Vector
Machine Objectives



Find a linear hyperplane (decision boundary) that will separate the data

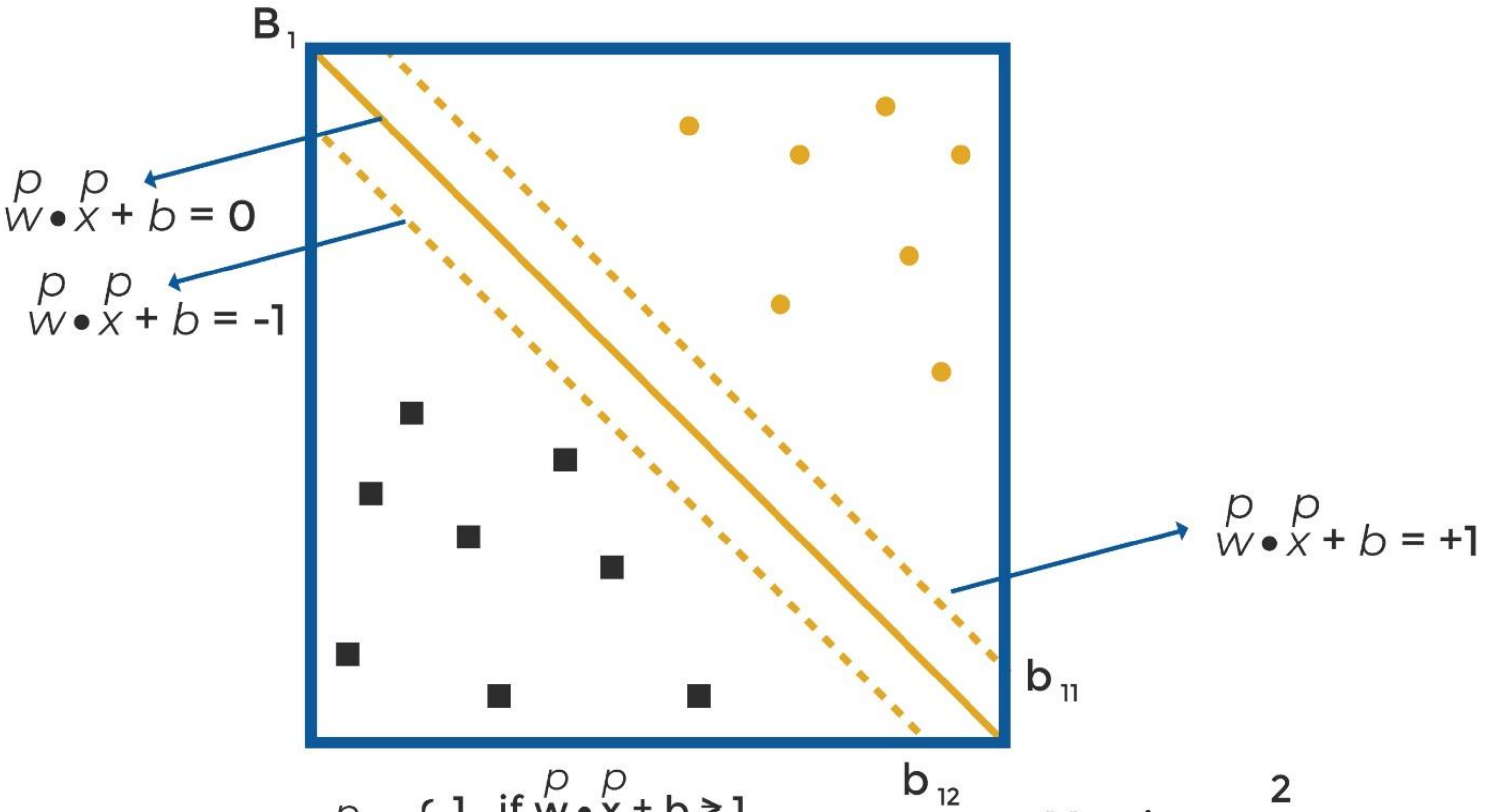
Support Vector
Machine



One Possible Solution | Another Possible Solution | Other Possible Solution

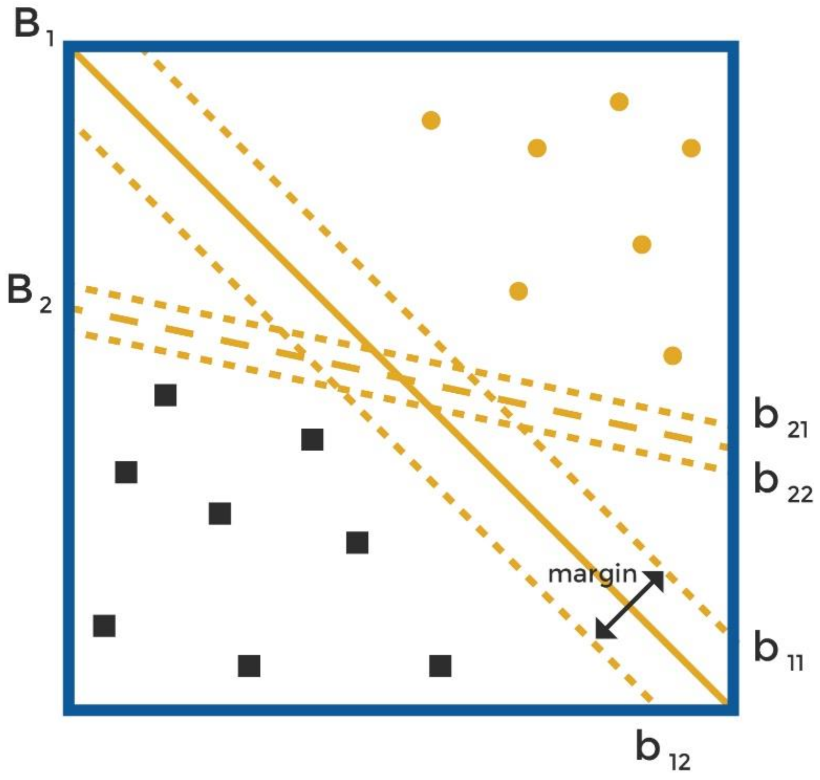
- Which one is better? B1 or B2?
- How do you define better?

Support Vector Machine



$$f(\vec{x}) = \begin{cases} 1 & \text{if } \vec{w} \bullet \vec{x} + b \geq 1 \\ -1 & \text{if } \vec{w} \bullet \vec{x} + b \leq 1 \end{cases}$$

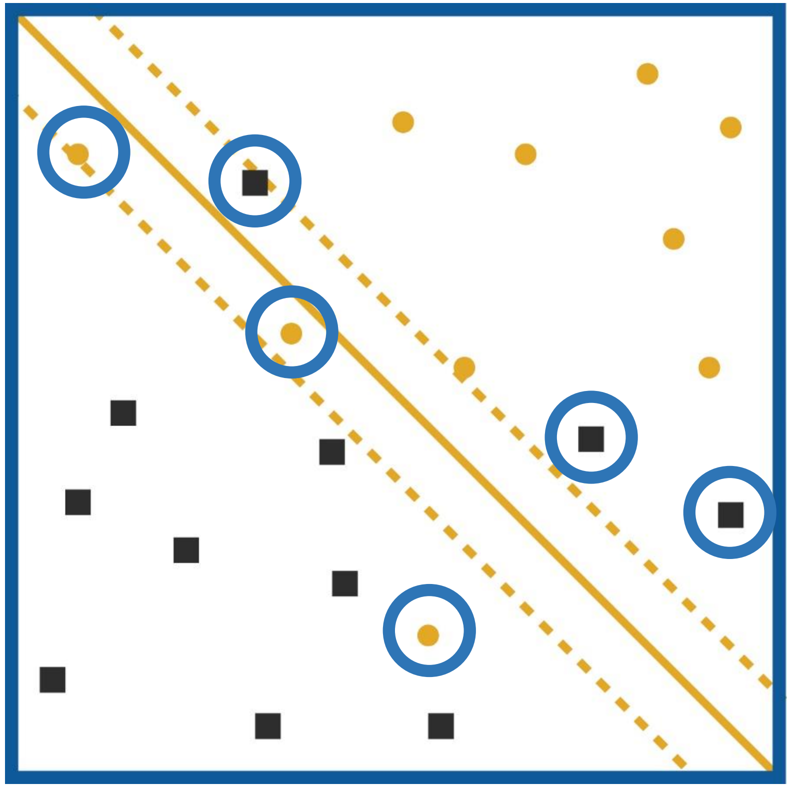
$$\text{Margin} = \frac{2}{\|\vec{w}\|^2}$$



Find hyperplane maximizes the margin => B1 is better than B2

Support Vector Machines Issue

- What if the problem is not linearly separable?



Support Vector Machines Issue

What if the problem is not linearly separable?

- Introduce slack variables
  - Need to minimize:

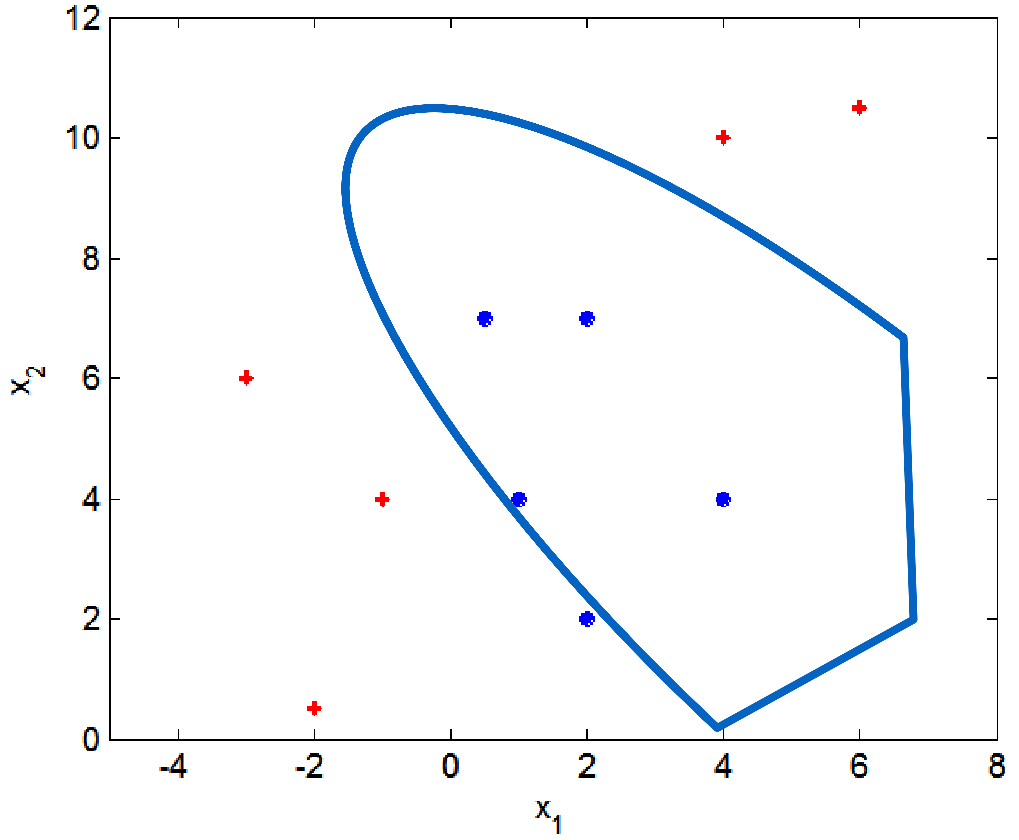
$$L(w) = \frac{\|\vec{w}\|^2}{2} + C\left(\sum_{i=1}^{N} \xi_i^k\right)$$

  - Subject to:

$$f(\vec{x}_i) = \begin{cases} 1 & \text{if } \vec{w} \bullet \vec{x}_i + b \geq 1 - \xi_i \\ -1 & \text{if } \vec{w} \bullet \vec{x}_i + b \leq -1 + \xi_i \end{cases}$$

Nonlinear Support Vector Machines

- What if decision boundary is not linear?



SVM
Summary

- Advantages:
  - SVM’s are very good when we have no idea on the data.
  - Works well with even unstructured and semi structured data like text, Images and trees.
  - The kernel trick is real strength of SVM. With an appropriate kernel function, we can solve any complex problem.
  - Unlike in neural networks, SVM is not solved for local optima.
  - It scales relatively well to high dimensional data.
  - SVM models have generalization in practice, the risk of overfitting is less in SVM.

- Disadvantages:
  - Choosing a “good” kernel function is not easy.
  - Long training time for large datasets.
  - Difficult to understand and interpret the final model, variable weights and individual impact.
  - Since the final model is not so easy to see, we can not do small calibrations to the model hence its tough to incorporate our business logic.

Agenda

- Introduction to Classification
- Decision Tree
- Random Forest
- Bayesian
- Lazy Learner (kNN)
- Support Vector Machine
- **Model Evaluation and Selection**

Metrics for
Performance Evaluation

- Focus on the predictive capability of a model
  - Rather than how fast it takes to classify or build models, scalability, etc.
- Confusion Matrix:

	Predicted Class		
Actual Class		Class = Yes	Class = No
	Class = Yes	a (TP)	b (TN)
	Class = No	c (FP)	d (FN)

- Most widely-used metric:

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

Limitation of Accuracy

- Consider a 2-class problem
  - Number of Class 0 examples = 9990
  - Number of Class 1 examples = 10
- If model predicts everything to be class 0, accuracy is 9990/10000 = 99.9 %
  - Accuracy is misleading because model does not detect any class 1 example

Cost Matrix

	Predicted Class		
Actual Class	C(i|j)	Class = Yes	Class = No
	Class = Yes	C(Yes|Yes)	C(No|Yes)
	Class = No	C(Yes|No)	C(No|No)

C(i|j): Cost of misclassifying class j example as class i

Computing Cost of Classification

Cost Matrix	Predicted Class		
Actual Class	C(i|j)	+	-
	+	-1	100
	-	1	0

Model M_1	Predicted Class		
Actual Class		+	-
	+	150	40
	-	60	250

Accuracy = 80%

Cost = 3910

Model M_2	Predicted Class		
Actual Class		+	-
	+	250	45
	-	5	200

Accuracy = 90%

Cost = 4255

Cost vs
Accuracy

	Predicted Class	
Actual Class	Class = Yes	Class = No
Class = Yes	a	b
Class = No	c	d

	Predicted Class	
Actual Class	Class = Yes	Class = No
Class = Yes	p	q
Class = No	q	p

- Accuracy is proportional to cost if

  1. C(Yes|No) =C(No|Yes) = q
  2. C(Yes|Yes) =C(No|No) = p

- N = a + b + c + d

- Accuracy = (a + d)/N

- Cost = p (a + d) + q (b + c)

  = p (a + d) + q (N – a – d)

  = q N – (q – p)(a + d)

  = N [q – (q-p) × Accuracy]

Cost-Sensitive Measures

$$\text{Precision (p)} = \frac{a}{a+c}$$

$$\text{Recall (r)} = \frac{a}{a+b}$$

$$\text{F - measure (F)} = \frac{2rp}{r+p} = \frac{2a}{2a+b+c}$$

- Precision is biased towards C(Yes|Yes) & C(Yes|No)
- Recall is biased towards C(Yes|Yes) & C(No|Yes)
- F-measure is biased towards all except C(No|No)

$$\text{Weighted Accuracy} = \frac{w_1 a + w_4 d}{w_1 a + w_2 b + w_3 c + w_4 d}$$

		Predicted Class	
Actual Class		Class = Yes	Class = No
	Class = Yes	a (TP)	b (TN)
	Class = No	c (FP)	d (FN)