

EECS 510: SOCIAL MEDIA MINING

SPRING 2016

DATA MINING ESSENTIALS 1:

DATA MINING BASICS, DATA, DATA PREPROCESSING

ANKIT AGRAWAL

NORTHWESTERN UNIVERSITY

Source material from: Reza Zafarani, Mohammad-Ali Abbasi, Huan Liu, Alok Choudhary

INTRODUCTION

- Data production rate has been increased dramatically (Big Data) and we are able store much more data than before
 - E.g., purchase data, social media data, mobile phone data
 - 5Vs: Volume, Velocity, Value, Veracity, Variety
- Businesses and customers need useful or actionable knowledge and gain insight from raw data for various purposes
 - It's not just searching data or databases

Data mining helps us to extract new information and uncover hidden patterns out of the stored and streaming data

The process of discovering hidden patterns in large data sets

It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems

- *Extracting or “mining” knowledge from large amounts of data, or big data*
- Data-driven discovery and modeling of hidden patterns in big data
- Extracting implicit, previously unknown, unexpected, and potentially useful information/knowledge from data

DATA MINING STORIES

- “My bank called and said that they saw that I bought two surfboards at Laguna Beach, California.” - credit card fraud detection
- The NSA is using data mining to analyze telephone call data to track al’Qaeda activities
- Walmart uses data mining to control product distribution based on typical customer buying patterns at individual stores

The Unknown

**As we know,
There are known knowns.
There are things we know we know.**

Conventional Wisdom

- High Humidity results in outbreak of Meningitis
- Customers switch carriers when contract is over

Validate Hypothesis

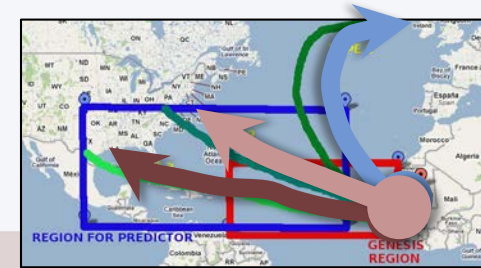
- Nuclear Reaction happens under these conditions
- Did combustion occur at the expected parameter values

e.g., Statistics, Query, Transformation, Viz

The Unknown

As we know,
There are known knowns.
There are things we know we know.

We also know
There are known unknowns.
That is to say
We know there are some things
We do not know.



Top-Down Discovery -
We know the question
to ask

- Will this hurricane strike the Atlantic coast?
- What is the likelihood of this patient to develop cancer
- Will this customer buy a new smart phone?

Predictive Modeling...; e.g., SVM, Decision Trees

The Unknown

As we know,
There are known knowns.
There are things we know we know.
We also know
There are known unknowns.
That is to say
We know there are some things
We do not know.

**But there are also unknown unknowns,
The ones we don't know
We don't know.**

Bottom up Discovery -
We don't know the
question to ask

- Wow! I found a new galaxy?
- Switch C fails when switch A fails followed by switch B failing
- On Thursday people buy beer and diaper together.
- The ratio $K/P > X$ is an indicator of onset of diabetes.



Relationship Mining, Clustering etc...

DATA MINING VS. DATABASES

- **Data mining** is the *process* of extracting hidden and actionable patterns from data
- **Database systems** store and manage data
 - Queries return part of stored data
 - Queries do not extract hidden patterns
- Examples of querying databases
 - Find all employees with income more than \$250K
 - Find top spending customers in last month
 - Find all students from *engineering college* with GPA more than average

EXAMPLES OF DATA MINING APPLICATIONS

- Identifying fraudulent transactions of a credit card or spam emails
 - You are given a user's purchase history and a new transaction, identify whether the transaction is fraud or not;
 - Determine whether a given email is spam or not
- Extracting purchase patterns from existing records
 - beer \Rightarrow diapers (80%)
- Forecasting future sales and needs according to some given samples
- Extracting groups of like-minded people in a given network

DATA

DATA INSTANCES

- A collection of properties and features related to an object or person
 - A patient's medical record
 - A user's profile
 - A gene's information
- Instances are also called examples, records, data points, or observations

Data Instance:

Patient Name	Blood Pressure	Chest Pain	Fatigued	Heart Disease
John	High	Yes	Yes	Yes

Features or Attributes

Class Label

DATA TYPES

- **Nominal** (categorical)
 - No comparison is defined
 - E.g., {male, female}
- **Ordinal**
 - Comparable but the difference is not defined
 - E.g., {Low, medium, high}
- **Interval**
 - Deduction and addition is defined but not division
 - E.g., 3:08 PM, calendar dates
- **Ratio**
 - E.g., Height, weight, money quantities

SAMPLE DATASET

outlook	temperature	humidity	windy	play
sunny	85	85	FALSE	no
sunny	80	90	TRUE	no
overcast	83	86	FALSE	yes
rainy	70	96	FALSE	yes
rainy	68	80	FALSE	yes
rainy	65	70	TRUE	no
overcast	64	65	TRUE	yes
sunny	72	95	FALSE	no
sunny	69	70	FALSE	yes
rainy	75	80	FALSE	yes
sunny	75	70	TRUE	yes
overcast	72	90	TRUE	yes
overcast	81	75	FALSE	yes
rainy	71	91	TRUE	no

No.	Outlook (O)	Temperature (T)	Humidity (H)	Play Golf (PG)
1	sunny	hot	high	N
2	sunny	mild	high	N
3	overcast	hot	high	Y
4	rain	mild	high	Y
5	sunny	cool	normal	Y
6	rain	cool	normal	N
7	overcast	cool	normal	Y
8	sunny	mild	high	?

When making data ready for data mining algorithms, data quality need to be assured, o/w GIGO!

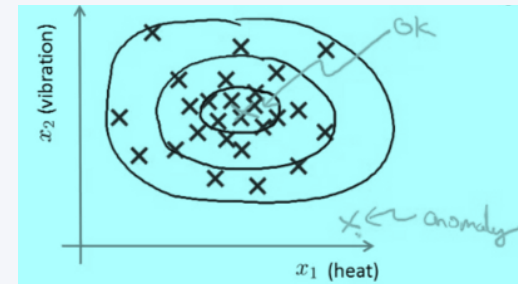
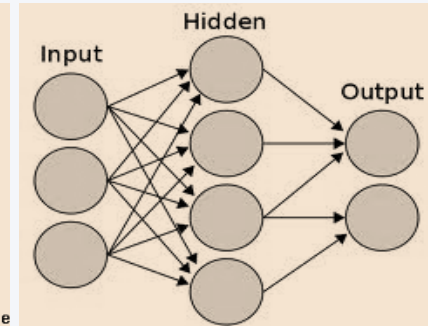
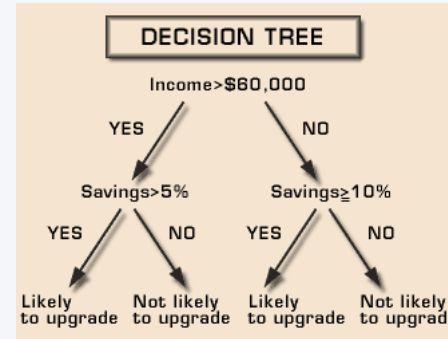
- **Noise**
 - Noise is the distortion of the data
- **Outliers**
 - Outliers are data points that are considerably different from other data points in the dataset
- **Missing Values**
 - Missing feature values in data instances
- **Duplicate data**

- **Aggregation**
 - when multiple attributes need to be combined into a single attribute or when the scale of the attributes change
- **Discretization**
 - From continuous values to discrete values
- **Feature Selection**
 - Choose relevant features
- **Feature Extraction**
 - Creating a mapping of new features from original features
- **Sampling**
 - Random Sampling
 - Sampling with or without replacement
 - Stratified Sampling

Data Mining Models

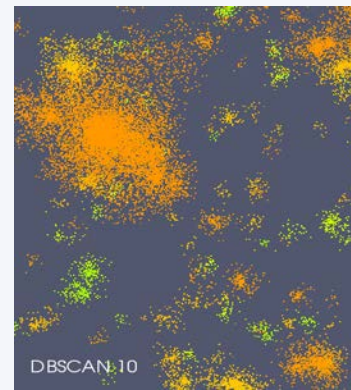
• Predictive

- Classification: learning a model to classify new records based on training data (e.g., decision trees, NN, SVM, etc.)
- Regression: learning a function to model the data while minimizing the error
- Anomaly Detection: Identification of outlier records that might lead to interesting discoveries



• Descriptive

- Clustering: Discovering groups of records that have similarities
- Association Rule Mining: Discovering relations between different attributes of the dataset



$\{\text{milk, bread}\} \Rightarrow \{\text{butter}\}$

$$\begin{aligned} \text{Rule: } X \Rightarrow Y & \begin{cases} \text{Support} = \frac{\text{freq}(X, Y)}{N} \\ \text{Confidence} = \frac{\text{freq}(X, Y)}{\text{freq}(X)} \\ \text{Lift} = \frac{\text{Support}}{\text{Supp}(X) \times \text{Supp}(Y)} \end{cases} \end{aligned}$$

CLASSIFICATION

CLASSIFICATION

Learning patterns from labeled data and classify new data with labels (categories)

- For example, we want to classify an e-mail as "legitimate" or "spam"



CLASSIFICATION: THE PROCESS

- In classification, we are given a set of labeled examples
- These examples are records/instances in the format (\mathbf{x}, y) where \mathbf{x} is a vector and y is the class attribute, commonly a scalar
- The classification task is to build model that maps \mathbf{x} to y
- Our task is to find a mapping f such that $f(\mathbf{x}) = y$

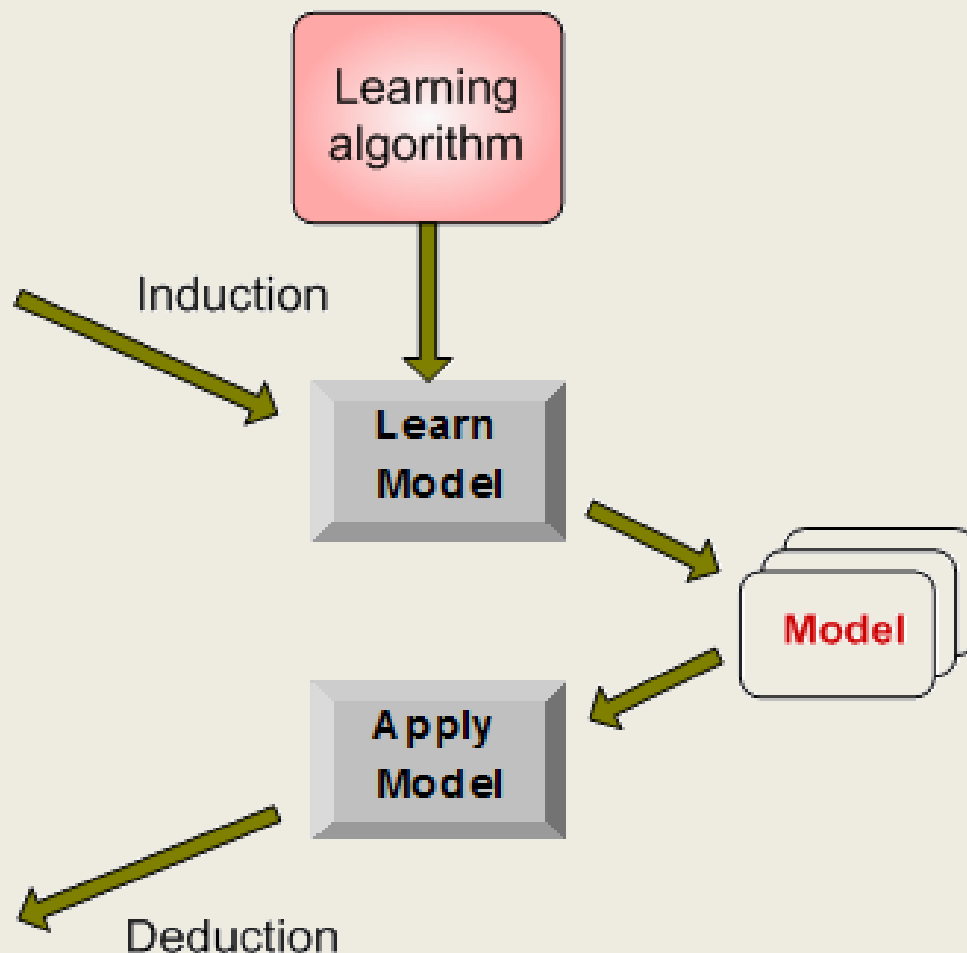
CLASSIFICATION: THE PROCESS

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	80K	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



CLASSIFICATION ALGORITHMS

- Decision tree learning
- Naive Bayes learning
- K-nearest neighbor classifier
- Artificial Neural Networks

DECISION TREE

- A decision tree is learned from the dataset (training data with known classes) and later applied to predict the class attribute value of new data (test data with unknown classes) where only the feature values are known, using a tree-like model.

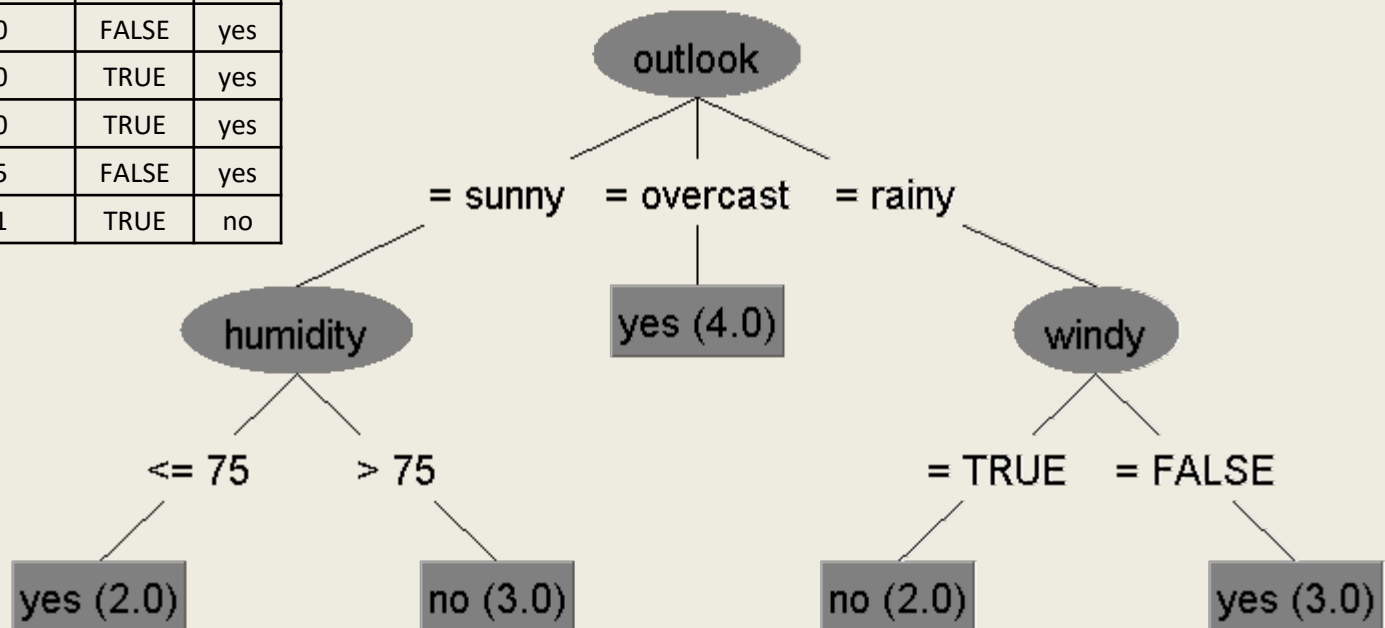
DECISION TREE: EXAMPLE 1

outlook	temperature	humidity	windy	play
sunny	85	85	FALSE	no
sunny	80	90	TRUE	no
overcast	83	86	FALSE	yes
rainy	70	96	FALSE	yes
rainy	68	80	FALSE	yes
rainy	65	70	TRUE	no
overcast	64	65	TRUE	yes
sunny	72	95	FALSE	no
sunny	69	70	FALSE	yes
rainy	75	80	FALSE	yes
sunny	75	70	TRUE	yes
overcast	72	90	TRUE	yes
overcast	81	75	FALSE	yes
rainy	71	91	TRUE	no

{sunny, 75,75,FALSE} → {?}

{overcast,85,90,TRUE} → {?}

{rainy,65,70,TRUE} → {?}



ID3, A DECISION TREE ALGORITHM

Use *information gain (entropy)* to determine how well an attribute separates the training data according to the class attribute value

$$\text{entropy}(D) = -p_+ \log p_+ - p_- \log p_-$$

- p_+ is the proportion of positive examples in D
- p_- is the proportion of negative examples in D

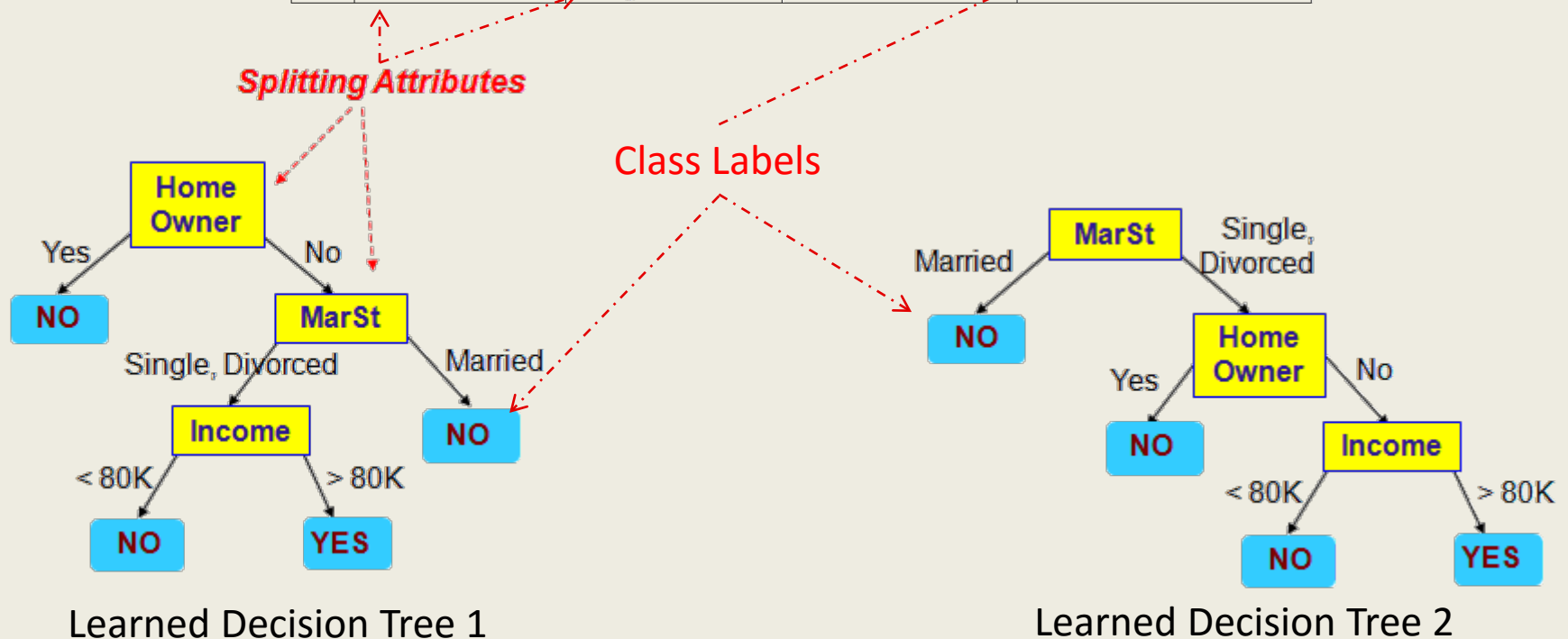
In a dataset containing ten examples, 7 have a positive class attribute value and 3 have a negative class attribute value [7+, 3-]:

$$\text{entropy}(D) = -\frac{7}{10} \log \frac{7}{10} - \frac{3}{10} \log \frac{3}{10} = 0.881$$

If the numbers of positive and negative examples in the set are equal, then the entropy is 1

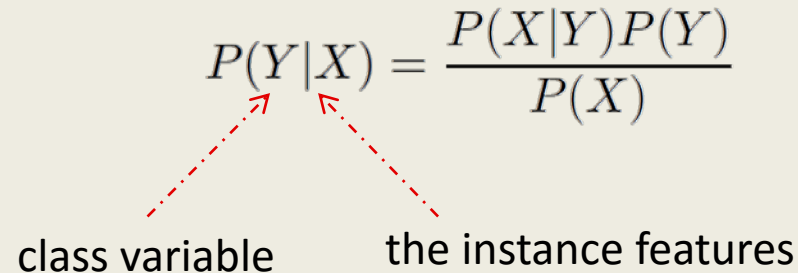
DECISION TREE: EXAMPLE 2

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
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



NAIVE BAYES CLASSIFIER

For two random variables X and Y , Bayes theorem states that,


$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$


class variable the instance features

Then class attribute value for instance x $\arg \max_{y_i} P(y_i|X)$

Assuming that variables
are independent

$$P(y_i|X) = \frac{P(X|y_i)P(y_i)}{P(X)}$$


$$P(X|y_i) = \prod_{j=1}^n P(x_j|y_i) \quad \Rightarrow \quad P(y_i|X) = \frac{(\prod_{j=1}^n P(x_j|y_i))P(y_i)}{P(X)}$$

NBC: AN EXAMPLE

No.	Outlook (O)	Temperature (T)	Humidity (H)	Play Golf (PG)
1	sunny	hot	high	N
2	sunny	mild	high	N
3	overcast	hot	high	Y
4	rain	mild	high	Y
5	sunny	cool	normal	Y
6	rain	cool	normal	N
7	overcast	cool	normal	Y
8	sunny	mild	high	?

$$\begin{aligned}
 (PG = Y|i_8) &= \frac{P(i_8|PG = Y)P(PG = Y)}{P(i_8)} \\
 &= P(O = \text{Sunny}, T = \text{mild}, H = \text{high}|PG = Y) \\
 &\quad \times \frac{P(PG = Y)}{P(i_8)} \\
 &= P(O = \text{Sunny}|PG = Y) \times P(T = \text{mild}|PG = Y) \\
 &\quad \times P(H = \text{high}|PG = Y) \times \frac{P(PG = Y)}{P(i_8)} \\
 &= \frac{1}{4} \times \frac{1}{4} \times \frac{2}{4} \times \frac{\frac{4}{7}}{P(i_8)} = \frac{1}{56P(i_8)}.
 \end{aligned}$$

$$\begin{aligned}
 P(PG = N|i_8) &= \frac{P(i_8|PG = N)P(PG = N)}{P(i_8)} \\
 &= P(O = \text{Sunny}, T = \text{mild}, H = \text{high}|PG = N) \\
 &\quad \times \frac{P(PG = N)}{P(i_8)} \\
 &= P(O = \text{Sunny}|PG = N) \times P(T = \text{mild}|PG = N) \\
 &\quad \times P(H = \text{high}|PG = N) \times \frac{P(PG = N)}{P(i_8)} \\
 &= \frac{2}{3} \times \frac{1}{3} \times \frac{2}{3} \times \frac{\frac{3}{7}}{P(i_8)} = \frac{4}{63P(i_8)}.
 \end{aligned}$$

$$\frac{1}{56P(i_8)} < \frac{4}{63P(i_8)}$$

Play Golf = N

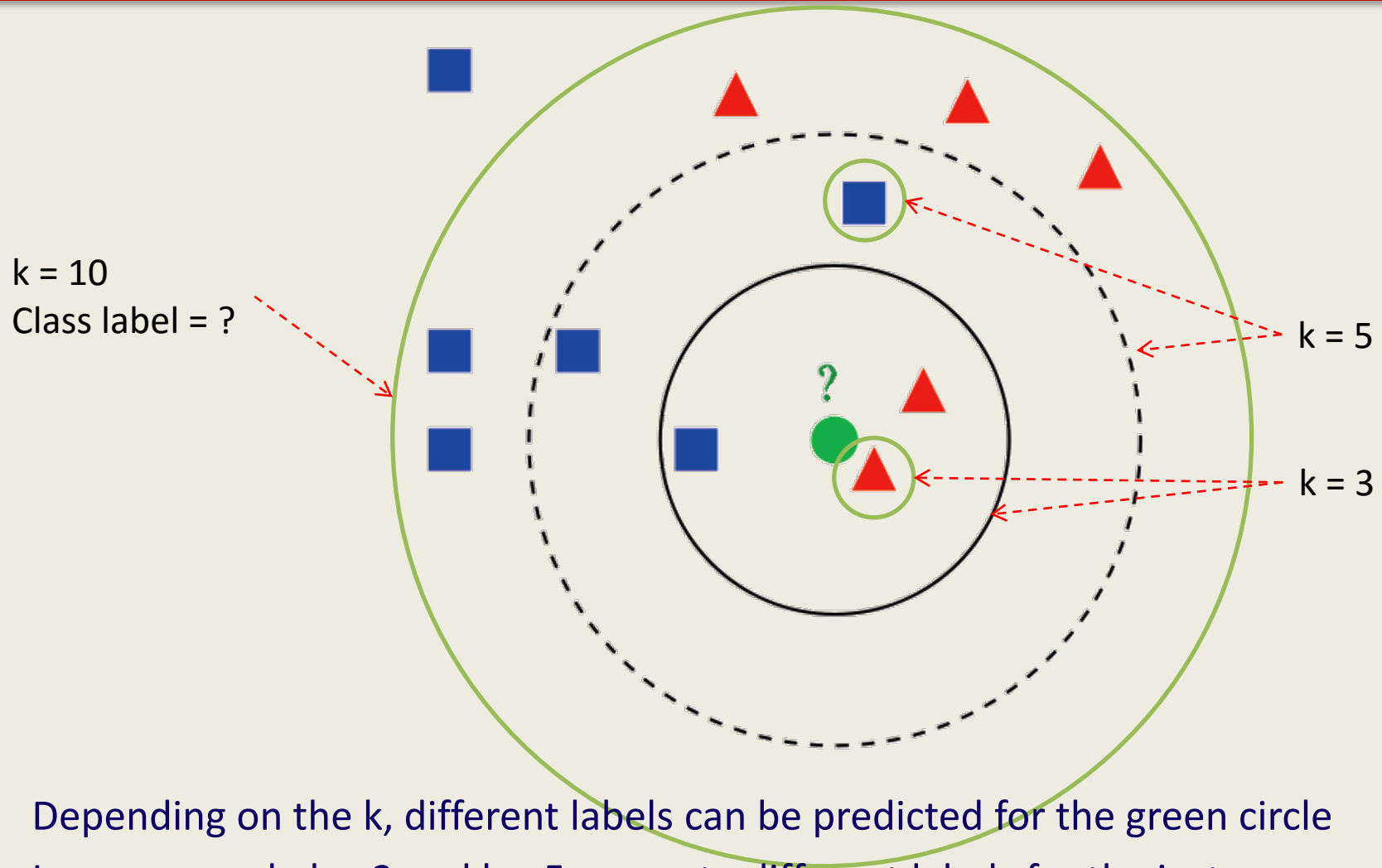
NEAREST NEIGHBOR CLASSIFIER

- k-nearest neighbor employs the neighbors of a data point to perform classification
- The instance being classified is assigned the label that the majority of k neighbors' labels
- When $k = 1$, the closest neighbor's label is used as the predicted label for the instance being classified
- For determining the neighbors, distance is computed based on some distance metric, e.g., Euclidean distance

K-NN: ALGORITHM

1. The dataset, number of neighbors (k), and the instance i is given
2. Compute the distance between i and all other data points in the dataset
3. Pick k closest neighbors
4. The class label for the data point i is the one that the majority holds (if there are more than one class, select one of them randomly)

K-NEAREST NEIGHBOR: EXAMPLE



- Depending on the k , different labels can be predicted for the green circle
- In our example $k = 3$ and $k = 5$ generate different labels for the instance
- $K = 10$ we can choose either triangle or rectangle

K-NEAREST NEIGHBOR: EXAMPLE

No.	Outlook (O)	Temperature (T)	Humidity (H)	Play Golf (PG)
1	sunny	hot	high	N
2	sunny	mild	high	N
3	overcast	hot	high	Y
4	rain	mild	high	Y
5	sunny	cool	normal	Y
6	rain	cool	normal	N
7	overcast	cool	normal	Y
8	sunny	mild	high	?

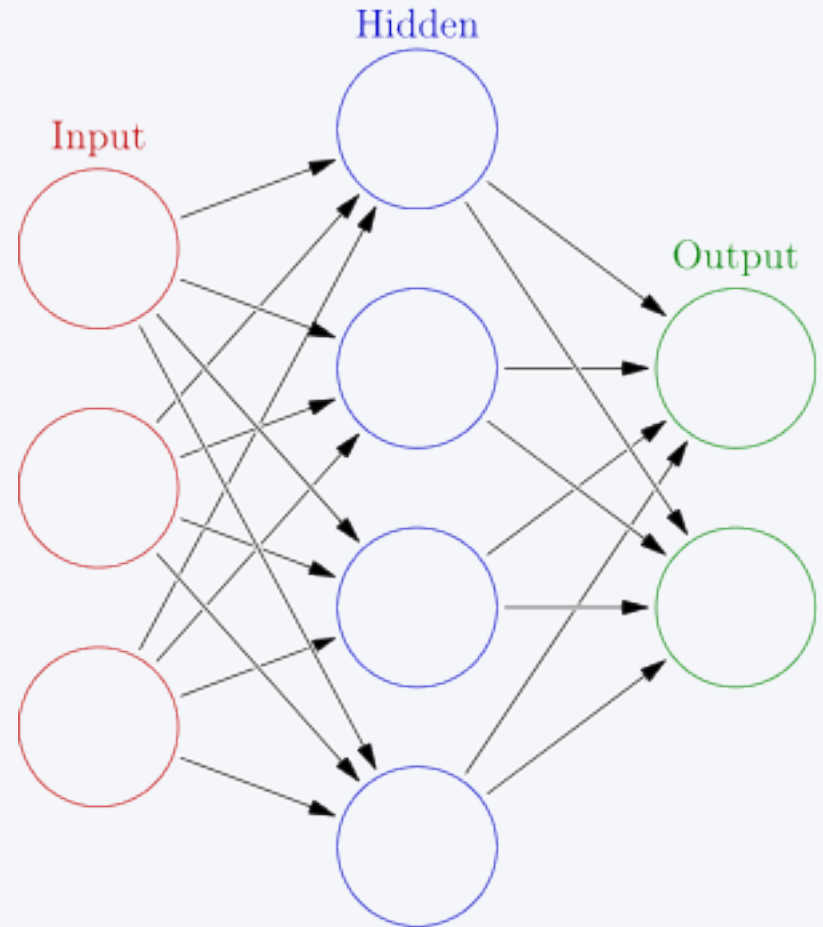
Similarity between row 8 and other data instances;

(Similarity = 1 if attributes have the same value, otherwise similarity = 0)

Data instance	Outlook	Temperature	Humidity	Similarity	Label	K	Prediction
2	1	1	1	3	N	1	
1	1	0	1	2	N	2	
4	0	1	1	2	Y	3	
3	0	0	1	1	Y	4	
5	1	0	0	1	Y	5	
6	0	0	0	0	N	6	
7	0	0	0	0	Y	7	

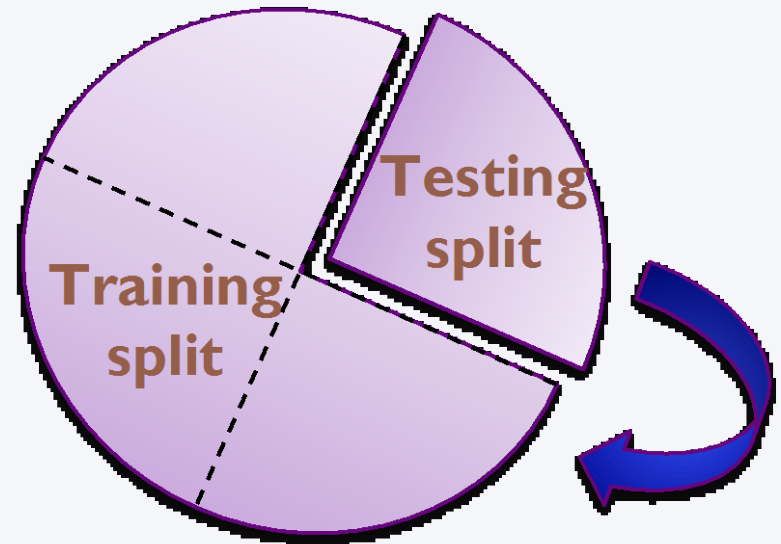
Artificial Neural Networks

- Computational models inspired by neuron connections in the brain
- Can have multiple layers of hidden artificial neurons
- Edges between artificial neurons have weights, which are refined during the model learning process
- Capable of learning non-linear functions.



Model Evaluation

- Test-train split
 - Split the labeled data into training and testing sets
- Cross-validation
 - Test every instance in the dataset using a model that has not seen that instance
 - Types
 - k -fold cross validation
 - Leave-one-out cross-validation (LOOCV) with $k=n$



k -fold cross validation

		Prediction	
		0	1
Actual	0	TN	FP
	1	FN	TP

Example

Confusion Matrix

		Prediction	
		0	1
Actual	0	TN	FP
	1	FN	TP

Spam
(1)

Test Instance ID	Model 1	Model 2	Model 3
1	0.7	0.6	0.4
2	0.8	0.7	0.3
3	0.95	0.9	0.45
4	0.4	0.3	0.35
5	0.6	0.2	0.45
6	0.4	0.6	0.25
7	0.6	0.4	0.2
8	0.3	0.3	0.15
9	0.2	0.7	0.1
10	0.1	0.2	0.05

Non-spam
(0)

M1

	0	1
0		
1		

Accuracy = %

M2

	0	1
0		
1		

Accuracy = %

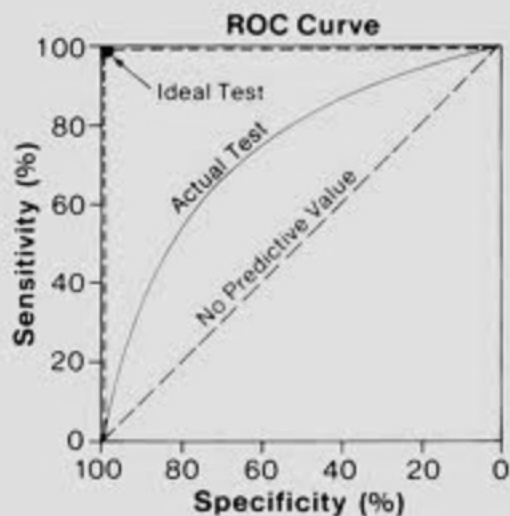
M3

	0	1
0		
1		

Accuracy = %

Evaluation Metrics - Classification

- Confusion matrix based
 - Accuracy, precision, recall, F-score,...
- Receiver operating characteristic curve (ROC) based
 - Area under the curve (AUC)



Confusion Matrix

		Prediction	
		0	1
Actual	0	TN	FP
	1	FN	TP

true positive (TP)

eqv. with hit

true negative (TN)

eqv. with correct rejection

false positive (FP)

eqv. with false alarm, Type I error

false negative (FN)

eqv. with miss, Type II error

sensitivity or true positive rate (TPR)

eqv. with hit rate, recall

$$TPR = TP/P = TP/(TP + FN)$$

specificity (SPC) or True Negative Rate

$$SPC = TN/N = TN/(FP + TN)$$

precision or positive predictive value (PPV)

$$PPV = TP/(TP + FP)$$

negative predictive value (NPV)

$$NPV = TN/(TN + FN)$$

fall-out or false positive rate (FPR)

$$FPR = FP/N = FP/(FP + TN) = 1 - SPC$$

false discovery rate (FDR)

$$FDR = FP/(TP + FP) = 1 - PPV$$

accuracy (ACC)

$$ACC = (TP + TN)/(P + N)$$

F1 score

is the harmonic mean of precision and sensitivity

$$F1 = 2TP/(2TP + FP + FN)$$

Matthews correlation coefficient (MCC)

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

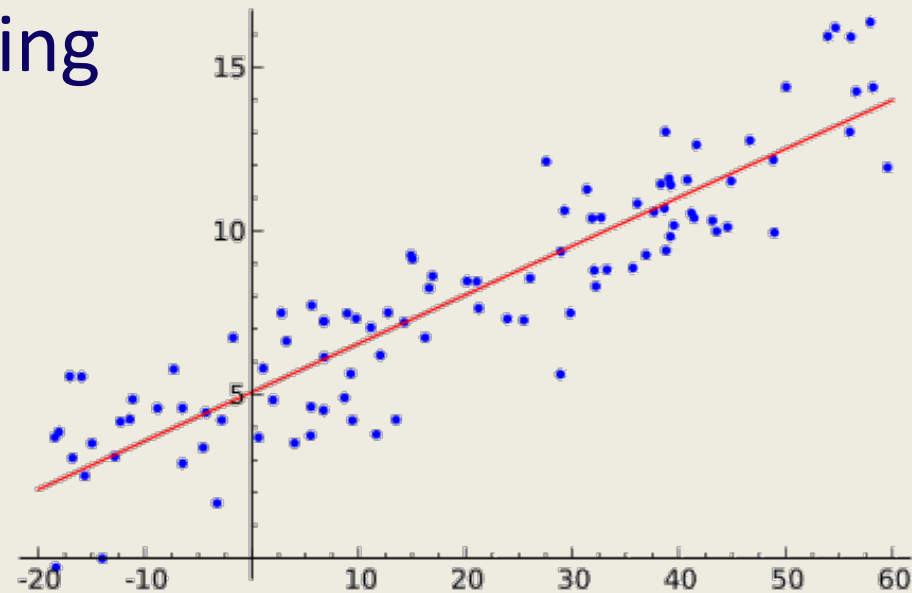
Source: Fawcett (2006).

REGRESSION

REGRESSION

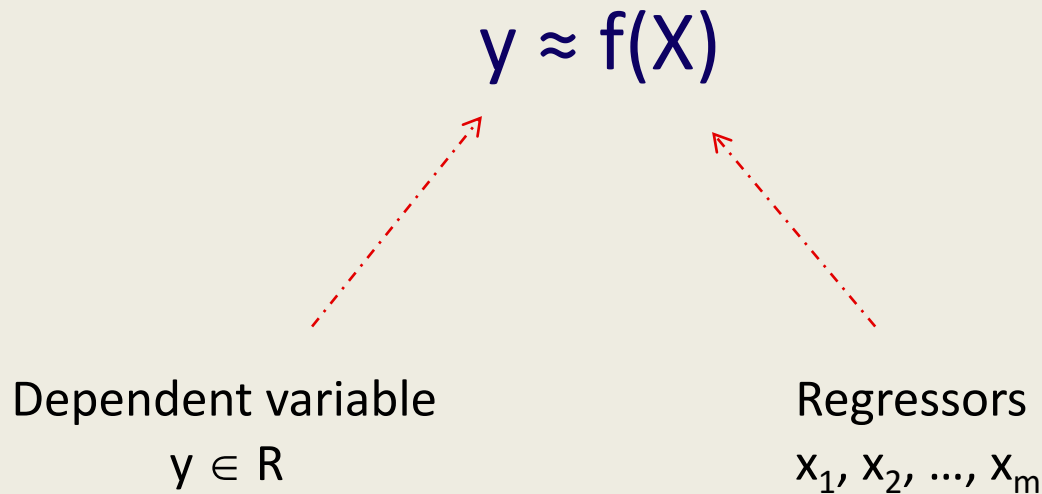
Regression analysis includes techniques of modeling and analyzing the relationship between a dependent variable and one or more independent variables

- Regression analysis is widely used for prediction and forecasting
- It can be used to infer relationships between the independent and dependent variables



REGRESSION

In regression, we deal with real numbers as class values (Recall that in classification, class values or labels are categories)



Our task is to find the relation between y and the vector (x_1, x_2, \dots, x_m)

LINEAR REGRESSION

In linear regression, we assume the relation between the class attribute y and feature set \mathbf{x} to be linear

$$y = \mathbf{x}\mathbf{w} + \epsilon$$

where \mathbf{w} represents the vector of regression coefficients

- The problem of regression can be solved by estimating \mathbf{w} and ϵ using the provided dataset and the labels y
 - The least squares is often used to solve the problem

SOLVING LINEAR REGRESSION PROBLEMS

- The problem of regression can be solved by estimating w and ε using the dataset provided and the labels y
 - “Least squares” is a popular method to solve regression problems

$$\|y - xw - w_0\|^2$$

REGRESSION COEFFICIENTS

- When there is only one independent variable:

$$y = w_0 + w_1 x$$

$$w_0 = \bar{y} - w_1 \bar{x}$$

$$\bar{x} = \frac{1}{n} \sum_i^n x_i$$

$$w_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- Two independent variables

$$y = w_0 + w_1 x_1 + w_2 x_2$$

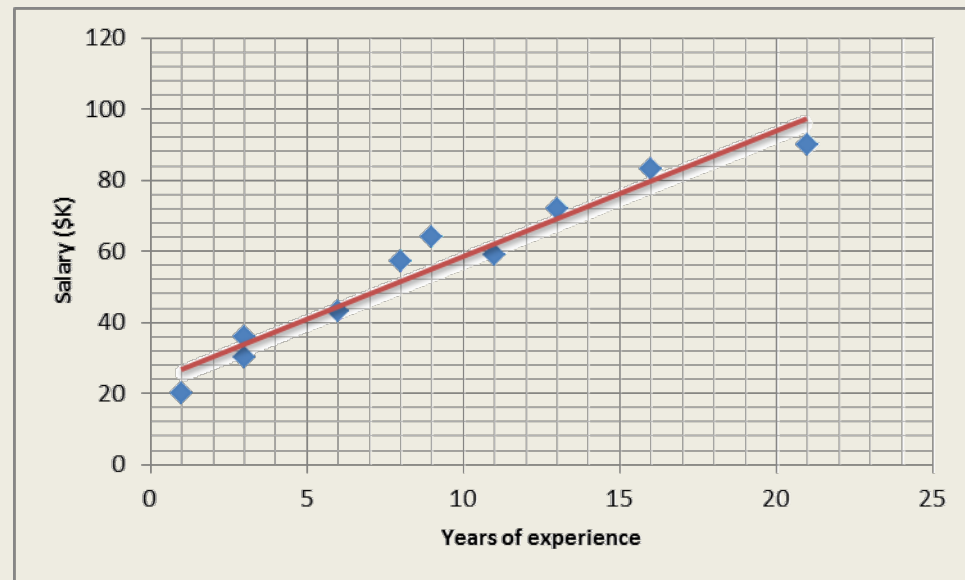
LINEAR REGRESSION: EXAMPLE

Years of experience	Salary (\$K)
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83

$$w_1 = \frac{(3 - 9.1)(30 - 55.4) + \dots + (16 - 9.1)(83 - 55.4)}{(3 - 9.1)^2 + \dots + (16 - 9.1)^2} = 3.5$$

$$w_0 = 55.4 - 3.5 \times 9.1 = 23.6$$

$$y = 23.6 + 3.5x$$



EVALUATING REGRESSION PERFORMANCE

- The labels cannot be predicted precisely
- It is needed to set a margin to accept or reject the predictions
 - For example, when the observed temperature is 71 any prediction in the range of 71 ± 0.5 can be considered as correct prediction

Evaluation Metrics - Regression

- ◆ Compare vectors of actual and predicted values
 - Coefficient of correlation (R)
 - Coefficient of determination (R^2)
 - Mean Absolute Error (MAE)
 - Root Mean Squared Error (RMSE)
 - Standard Deviation of Error (SDE)
 - Mean Absolute Error Fraction (MAE_f)
 - Root Mean Squared Error Fraction (RMSE_f)
 - Standard Deviation of Error Fraction (SDE_f)

$$R = \frac{\sum_{i=1}^N (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^N (y_i - \bar{y})^2 \sum_{i=1}^N (\hat{y}_i - \bar{\hat{y}})^2}}$$

$$MAE = \bar{e} = \frac{1}{N} \sum_N |y - \hat{y}|$$

$$RMSE = \sqrt{\frac{1}{N} \sum_N (y - \hat{y})^2}$$

$$SDE = \sqrt{\frac{1}{N} \sum_N (|y - \hat{y}| - \bar{e})^2}$$

$$MAE_f = \bar{e}_f = \frac{1}{N} \sum_N \left| \frac{y - \hat{y}}{y} \right|$$

$$RMSE_f = \sqrt{\frac{1}{N} \sum_N \left(\frac{y - \hat{y}}{y} \right)^2}$$

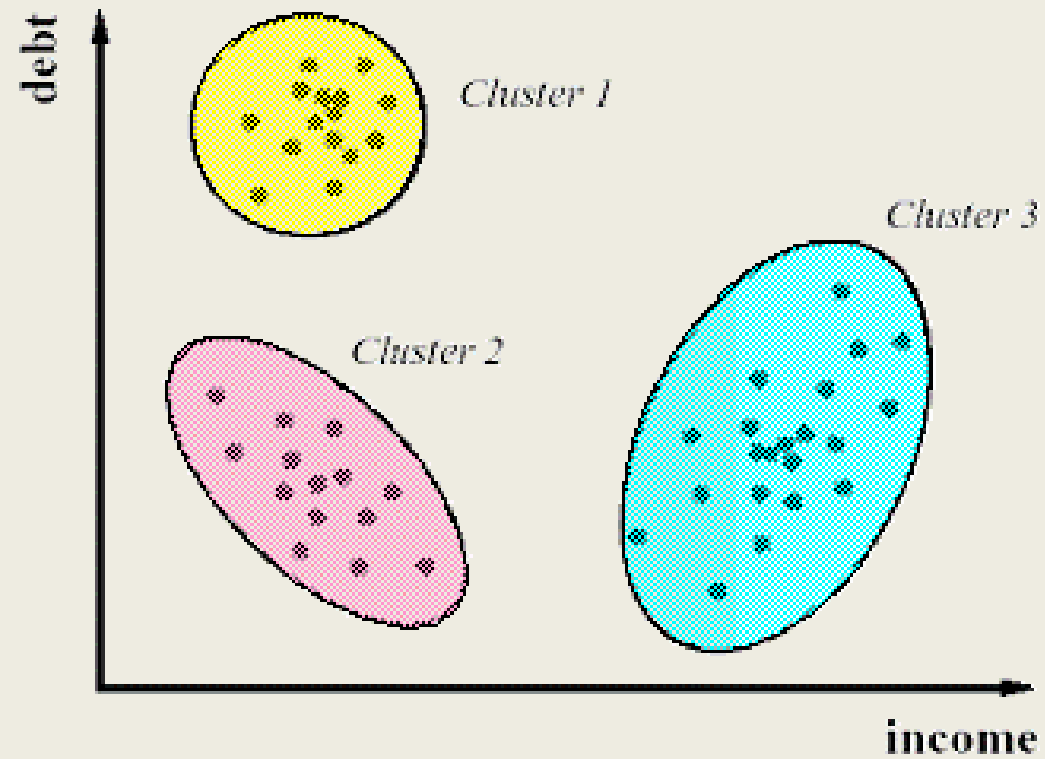
$$SDE_f = \sqrt{\frac{1}{N} \sum_N \left(\left| \frac{y - \hat{y}}{y} \right| - \bar{e}_f \right)^2}$$

CLUSTERING

Grouping together items that are similar in some way – according to some criteria

- Clustering is a form of **unsupervised learning**
 - The clustering algorithms do not have examples showing how the samples should be group together
- The clustering algorithms look for patterns or structures in the data that are of interest
- Clustering algorithms group together **similar items**

CLUSTERING: EXAMPLE

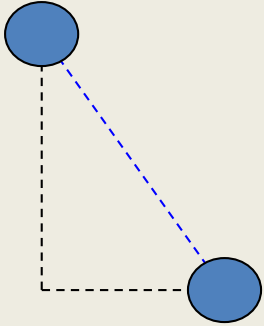


MEASURING SIMILARITY IN CLUSTERING ALGORITHMS

- The goal is to group together similar items
- Different similarity measures can be used to find similar items
- Usually similarity measures are critical to clustering algorithms

The most popular (dis)similarity measure for continuous features is ***Euclidean Distance***

EUCLIDEAN DISTANCE – A DISSIMILAR MEASURE



$$d(X, Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

- Here n is the number of dimensions in the data vector

SIMILARITY MEASURES: MORE DEFINITIONS

Measure Name	Formula	Type	Description
Mahalanobis	$d(X, Y) = \sqrt{(X - Y)^T Co^{-1}(X - Y)}$	Dissimilarity	X, Y are features vectors and Co is the Covariance matrix of the dataset
Manhattan	$d(X, Y) = \sum_i x_i - y_i $	Dissimilarity	X, Y are features vectors
L_p -norm	$d(X, Y) = (\sum_i x_i - y_i ^n)^{\frac{1}{n}}$	Dissimilarity	X, Y are features vectors
Cosine	$c(X, Y) = \frac{X \cdot Y}{ X Y }$	Similarity	X, Y are features vectors and ' \cdot ' represents the inner product

- Distance-based algorithms
 - K-Means
- Hierarchical algorithms

k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean

- Finding the global optimal of k partitions is computationally expensive (NP-hard). However, there are efficient heuristic algorithms that are commonly employed and converge quickly to an optimum that might not be global.

The algorithm is the most commonly used clustering algorithm.

Algorithm 2 K-Means Algorithm

Require: A Dataset of Real-Value Attributes, K (number of Clusters)

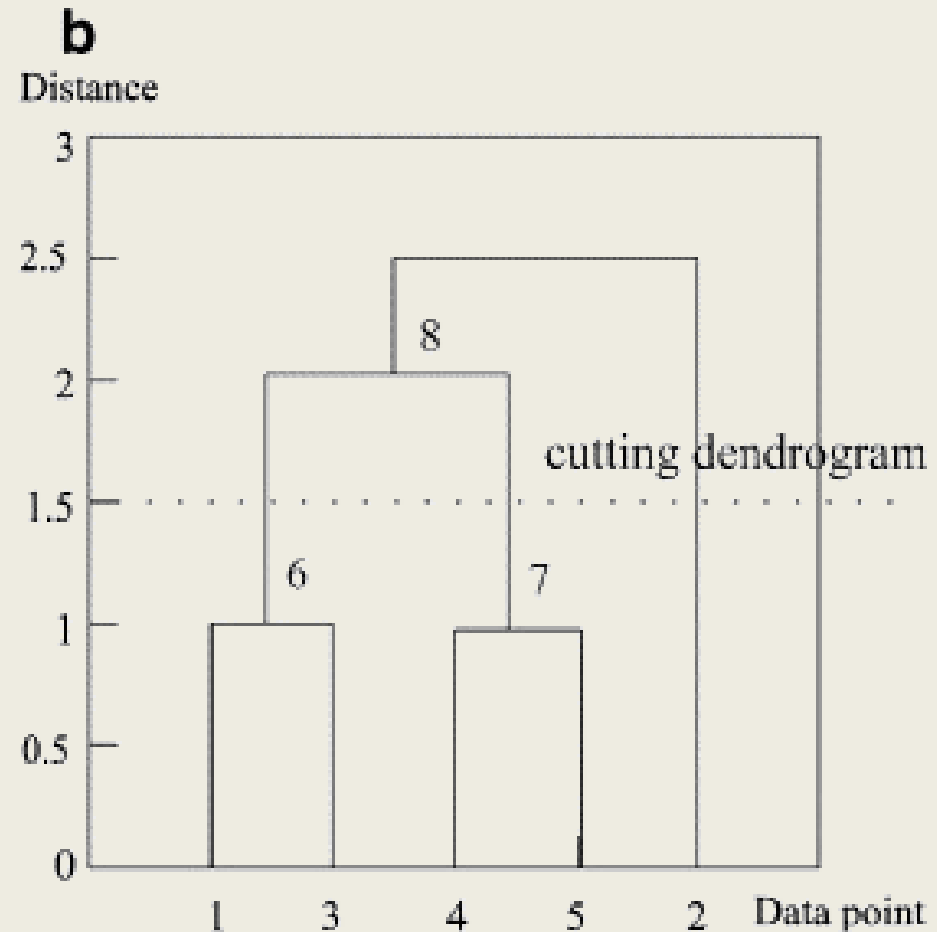
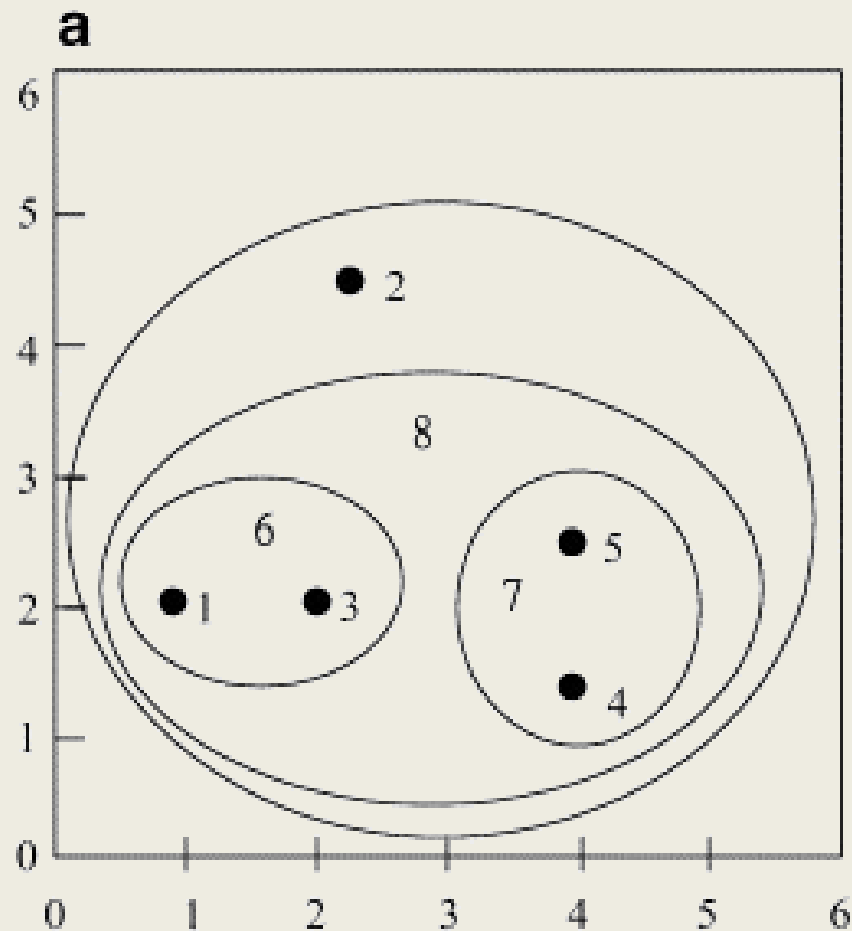
- 1: **return** A Clustering of Data into K Clusters
- 2: Consider K random points in the data space as the initial cluster centroids.
- 3: **while** centroids have not converged **do**
- 4: Assign each data point to the cluster which has the closest cluster centroid.
- 5: If all data points have been assigned then recalculate the cluster centroids by averaging datapoints inside each cluster
- 6: **end while**

HIERARCHICAL CLUSTERING

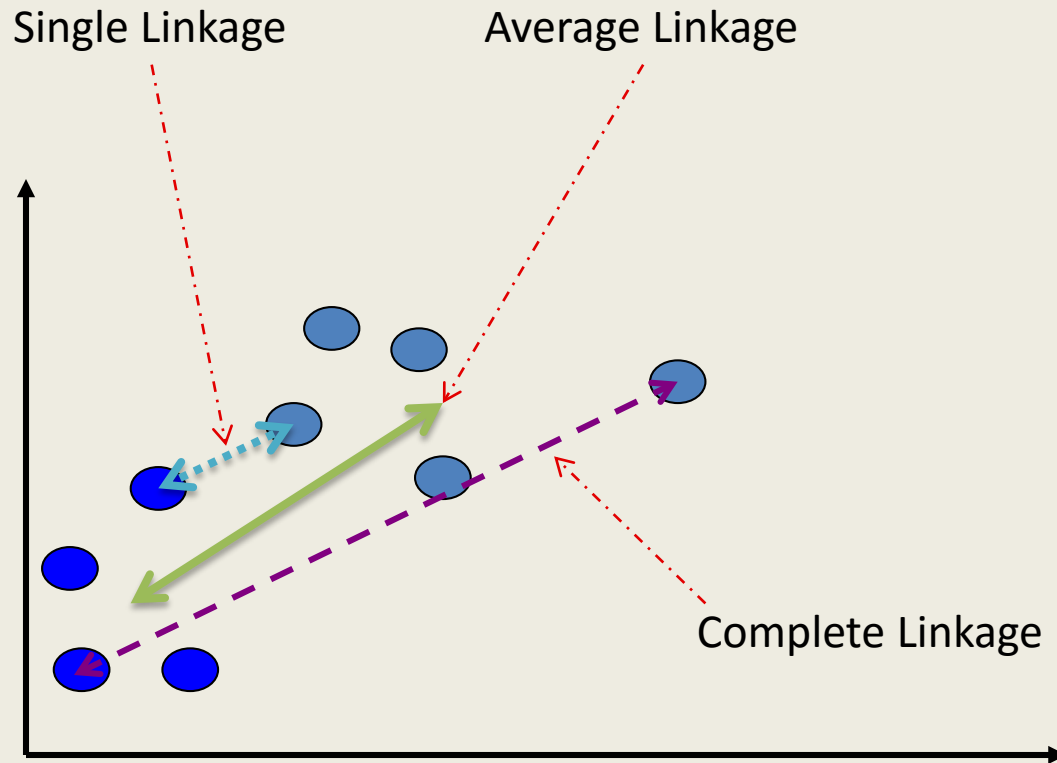
Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters.

- Strategies for hierarchical clustering generally fall into two types:
 - **Agglomerative:** This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
 - **Divisive:** This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

HIERARCHICAL CLUSTERING: AN EXAMPLE

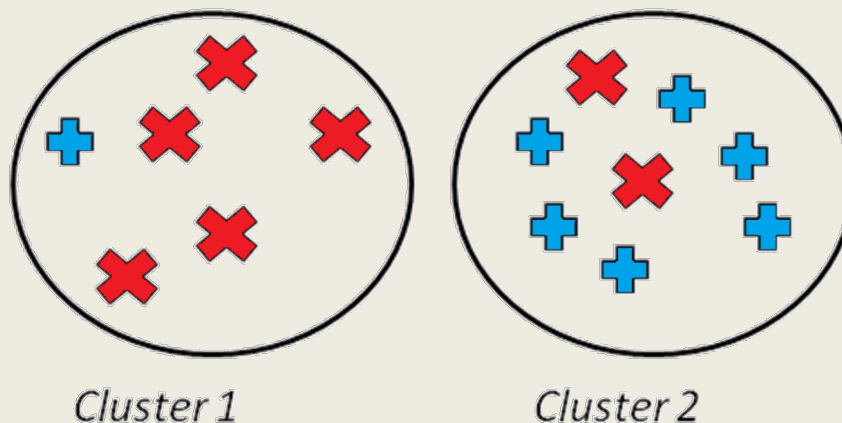


LINKAGE IN HIERARCHICAL CLUSTERING: EXAMPLE



EVALUATING THE CLUSTERINGS

When we are given objects of two different kinds, the perfect clustering would be that objects of the same type are clustered together.



- Evaluation with ground truth
- Evaluation without ground truth

EVALUATION WITH GROUND TRUTH

When ground truth is available, the evaluator has prior knowledge of what a clustering should be

- That is, we know the correct clustering assignments.
- Measures
 - Precision and Recall, or F-Measure

PRECISION AND RECALL

$$\text{Precision} = \frac{\text{Relevant and retrieved}}{\text{Retrieved}}$$

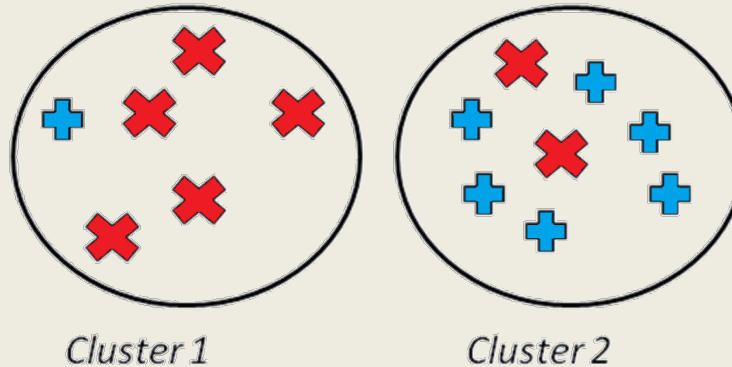
$$P = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{\text{Relevant and retrieved}}{\text{Relevant}}$$

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

- **True Positive (TP) :**
 - when similar points are assigned to the same clusters
 - This is considered a correct decision.
- **True Negative (TN) :**
 - when dissimilar points are assigned to different clusters
 - This is considered a correct decision
- **False Negative (FN) :**
 - when similar points are assigned to different clusters
 - This is considered an incorrect decision
- **False Positive (FP) :**
 - when dissimilar points are assigned to the same clusters
 - This is considered an incorrect decision

PRECISION AND RECALL: EXAMPLE 1



$$TP = \binom{5}{2} + \binom{6}{2} + \binom{2}{2} = 26,$$

$$FP = (5 \times 1) + (6 \times 2) = 17,$$

$$FN = (5 \times 2) + (6 \times 1) = 16,$$

$$TN = (6 \times 5) + (2 \times 1) = 32.$$

$$P = \frac{26}{26+17} = 0.60$$

$$R = \frac{26}{26+16} = 0.61$$

- To consolidate precision and recall into one measure, we can use the harmonic mean of precision of recall

$$F = 2 \cdot \frac{P \cdot R}{P + R}$$

Computed for the same example, we get $F = 0.6049$

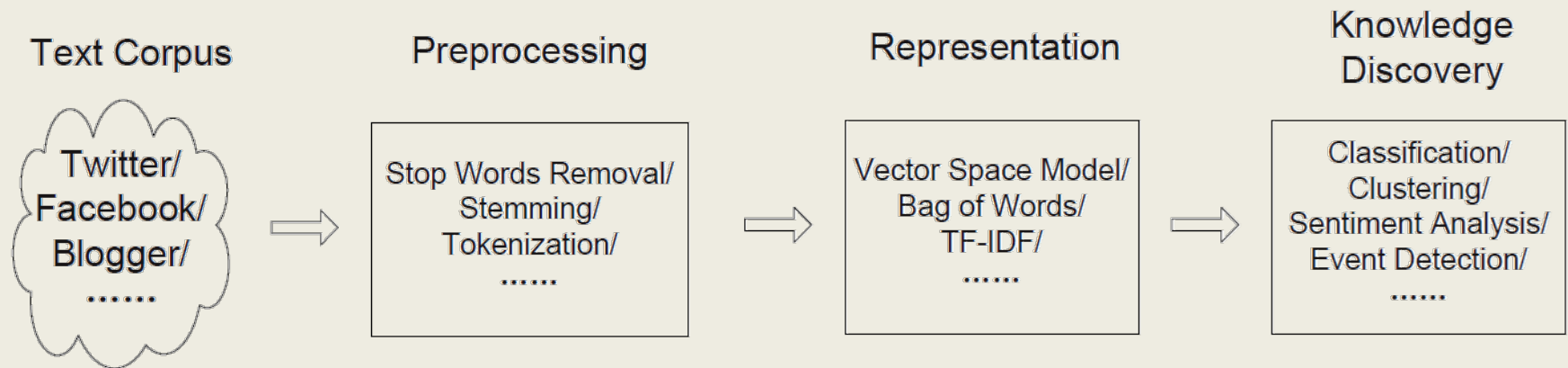
EVALUATION WITHOUT GROUND TRUTH

- Use domain experts
- Use quality measures such as SSE
 - SSE: the **sum of the squared error** for all clusters
 - Intra-cluster variation
- Use more than two clustering algorithms and compare the results and pick the algorithm with better quality measure

TEXT MINING

- In social media, most of the data that is available online is in text format
- In general, the way to perform data mining is to convert text data into tabular format and then perform data mining on this data
- The process of converting text data into tabular data is called *vectorization*

TEXT MINING PROCESS



A set of linguistic, statistical, and machine learning techniques that model and structure the information content of textual sources for business intelligence, exploratory data analysis, research, or investigation

Text preprocessing aims to make the input documents more consistent to facilitate text representation, which is necessary for most text analytics tasks

- **Methods:**
 - **Stop word removal**
 - Stop word removal eliminates words using a stop word list, in which the words are considered more general and meaningless
 - e.g. the, a, is, at, which
 - **Stemming**
 - Stemming reduces inflected (or sometimes derived) words to their stem, base or root form
 - For example, “watch”, “watching”, “watched” are represented as “watch”

TEXT REPRESENTATION

- The most common way to model documents is to transform them into sparse numeric vectors and then deal with them with linear algebraic operations
- This representation is called “Bag of Words”
- Methods:
 - Vector space model
 - tf-idf

VECTOR SPACE MODEL

- In the vector space model, we start with a set of documents, D
- Each document is a set of words
- The goal is to convert these textual documents to vectors

$$d_i = (w_{1,i}, w_{2,i}, \dots, w_{N,i})$$

- d_i : document i , $w_{j,i}$: the weight for word j in document i

The weight can be set to 1 when the word exist in the document and 0 when it does not. Or we can set this weight to the number of times the word is observed in the document

VECTOR SPACE MODEL: AN EXAMPLE

- Documents:
 - d1: data mining and social media mining
 - d2: social network analysis
 - d3: data mining
- Reference vector:
 - (social, media, mining, network, analysis, data)
- Vector representation:

	analysis	data	media	mining	network	social
d1	0	1	1	1	0	1
d2	1	0	0	0	1	1
d3	0	1	0	1	0	0

TF-IDF (TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY)

tf-idf of term t , document d , and document corpus D is calculated as follows:

$$\text{tf-idf}(t, d, D) = \text{tf}(t, d) * \text{idf}(t, D)$$

$$\text{idf}(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

The total number of documents in the corpus

The number of documents where the term t appears

TF-IDF: AN EXAMPLE

Consider words “apple” and “orange” that appear 10 and 20 times in document 1 (d1).

Consider $|D| = 20$ and word “apple” only appearing in d1 and word “orange” appearing in all 20 documents

$$tf - idf(\text{“apple”}, d_1) = 10 \times \log_2 \frac{20}{1} = 10,$$

$$tf - idf(\text{“orange”}, d_1) = 20 \times \log_2 \frac{20}{20} = 0.$$

TF-IDF: AN EXAMPLE

- Documents:
 - d1: data mining and social media mining
 - d2: social network analysis
 - d3: data mining
- tf-idf representation:

	analysis	data	media	mining	network	social
df(w)	1	2	1	2	1	2
log(N/df(w))	0.48	0.18	0.48	0.18	0.48	0.18
d1, tf	0	1	1	2	0	1
d2, tf	1	0	0	0	1	1
d3, tf	0	1	0	1	0	0
d1, tf-idf	0.00	0.18	0.48	0.35	0.00	0.18
d2, tf-idf	0.48	0.00	0.00	0.00	0.48	0.18
d3, tf-idf	0.00	0.18	0.00	0.18	0.00	0.00



Thank You !

Ankit Agrawal

Research Associate Professor
Dept. of Electrical Engineering and
Computer Science
Northwestern University
ankitag@eecs.northwestern.edu