Overview of Data Mining

Mehedy Masud

Lecture slides modified from:

Jiawei Han (http://www-sal.cs.uiuc.edu/~hanj/DM_Book.html)

Vipin Kumar (http://www-users.cs.umn.edu/~kumar/csci5980/index.html)

Ad Feelders (http://www.cs.uu.nl/docs/vakken/adm/)

Zdravko Markov (http://www.cs.ccsu.edu/~markov/ccsu_courses/DataMining-1.html)

Outline

- Definition, motivation & application
- Branches of data mining
- Classification, clustering, Association rule mining
- Some classification techniques

What Is Data Mining?

- Data mining (knowledge discovery in databases):
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases

- Alternative names and their "inside stories":
  - Data mining: a misnomer?
  - Knowledge discovery(mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, business intelligence, etc.

Data Mining Definition

- Finding hidden information in a database
- Fit data to a model
- Similar terms
  - Exploratory data analysis
  - Data driven discovery
  - Deductive learning

Motivation:

- Data explosion problem
  - Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories
- We are drowning in data, but starving for knowledge!
- Solution: Data warehousing and data mining
  - Data warehousing and on-line analytical processing
  - Extraction of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases

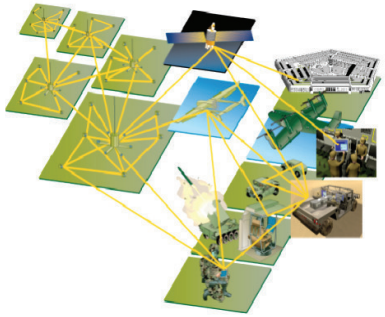
Why Mine Data? Commercial Viewpoint

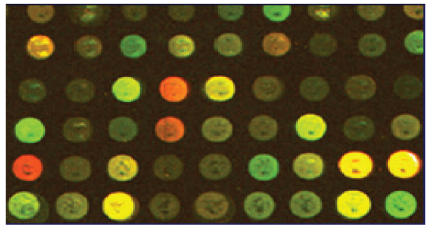
- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions

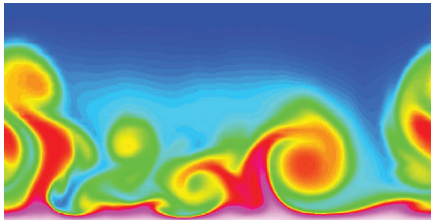
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

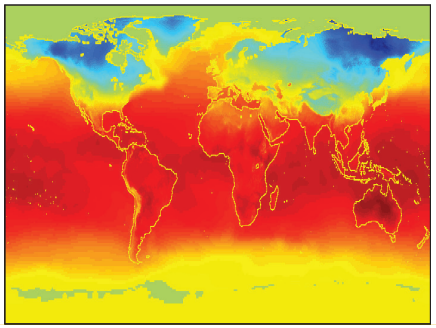
Why Mine Data? Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation









Examples: What is (not) Data Mining?

What is not Data Mining?

- Look up phone number in phone directory
- Query a Web search engine for information about "Amazon"

What is Data Mining?

- Certain names are more prevalent in certain US locations (O'Brien, O'Rurke, O'Reilly… in Boston area)
- Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)

Database Processing vs. Data Mining Processing

- Query
  - Well defined
  - SQL
- Data
  - Operational data
- Output
  - Precise
  - Subset of database

- Query
  - Poorly defined
  - No precise query language
- Data
  - Not operational data
- Output
  - Fuzzy
  - Not a subset of database

Query Examples

- Database
  - Find all credit applicants with last name of Smith.
  - Identify customers who have purchased more than $10,000 in the last month.
  - Find all customers who have purchased milk
- Data Mining
  - Find all credit applicants who are poor credit risks. (classification)
  - Identify customers with similar buying habits. (Clustering)
  - Find all items which are frequently purchased with milk. (association rules)

Data Mining: Classification Schemes

- Decisions in data mining
  - Kinds of databases to be mined
  - Kinds of knowledge to be discovered
  - Kinds of techniques utilized
  - Kinds of applications adapted
- Data mining tasks
  - Descriptive data mining
  - Predictive data mining

Decisions in Data Mining

- **Databases to be mined**
  - Relational, transactional, object-oriented, object-relational, active, spatial, time-series, text, multi-media, heterogeneous, legacy, WWW, etc.
- **Knowledge to be mined**
  - Characterization, discrimination, association, classification, clustering, trend, deviation and outlier analysis, etc.
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
  - Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, neural network, etc.
- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, DNA mining, stock market analysis, Web mining, Weblog analysis, etc.

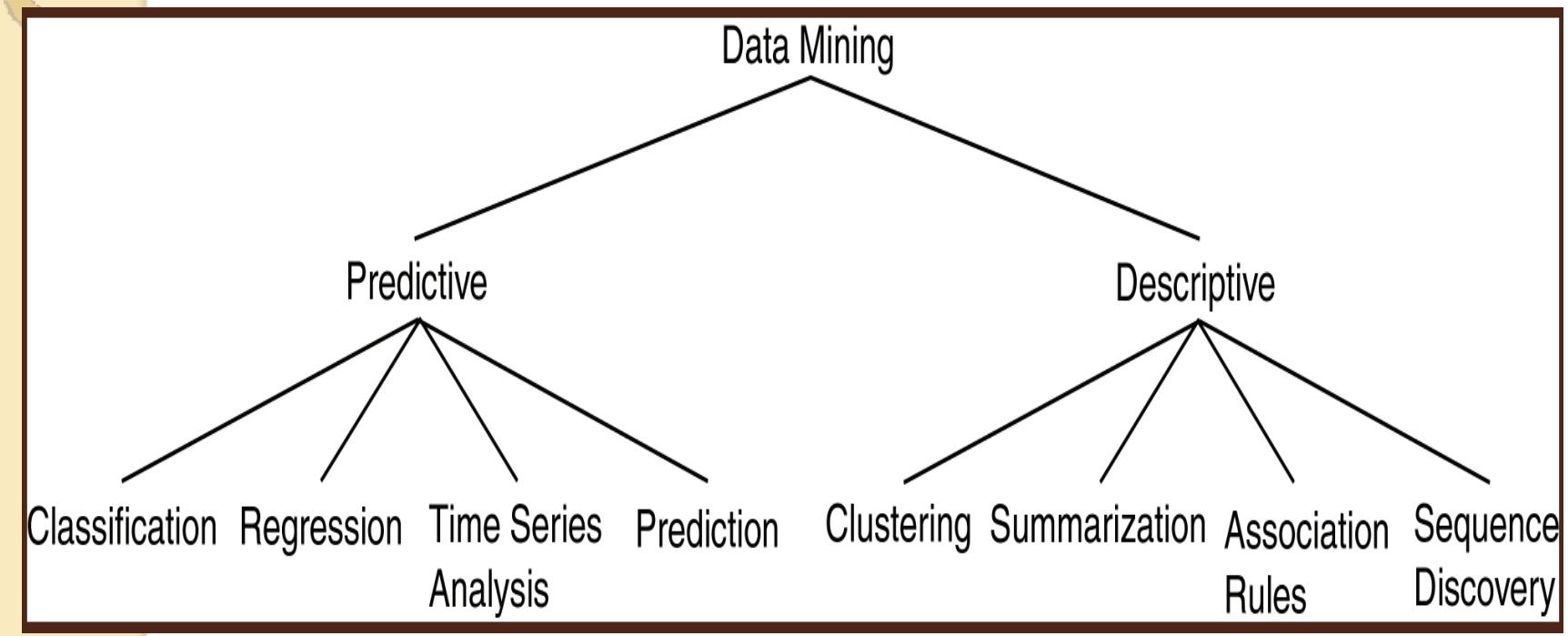
Data Mining Tasks

- Prediction Tasks
  - Use some variables to predict unknown or future values of other variables
- Description Tasks
  - Find human-interpretable patterns that describe the data.

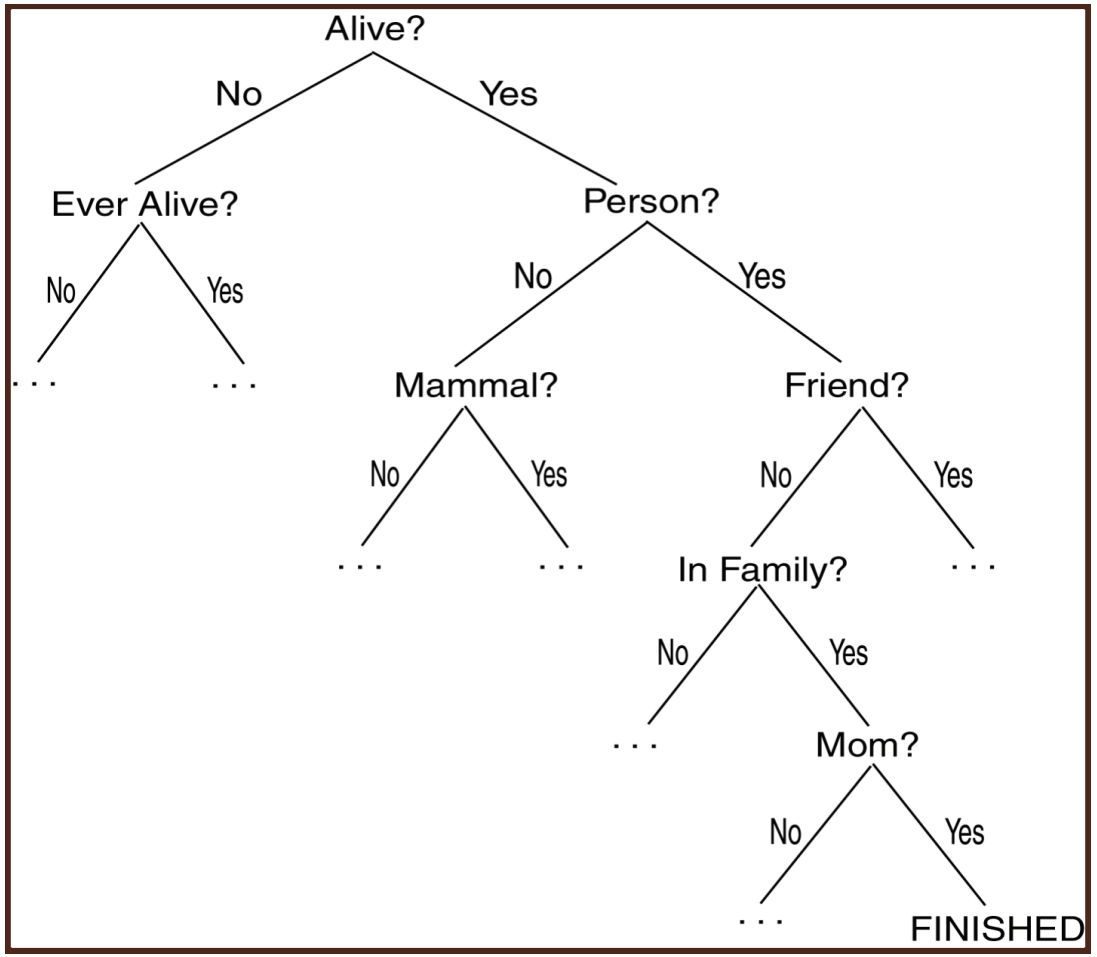
Common data mining tasks

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

Data Mining Models and Tasks



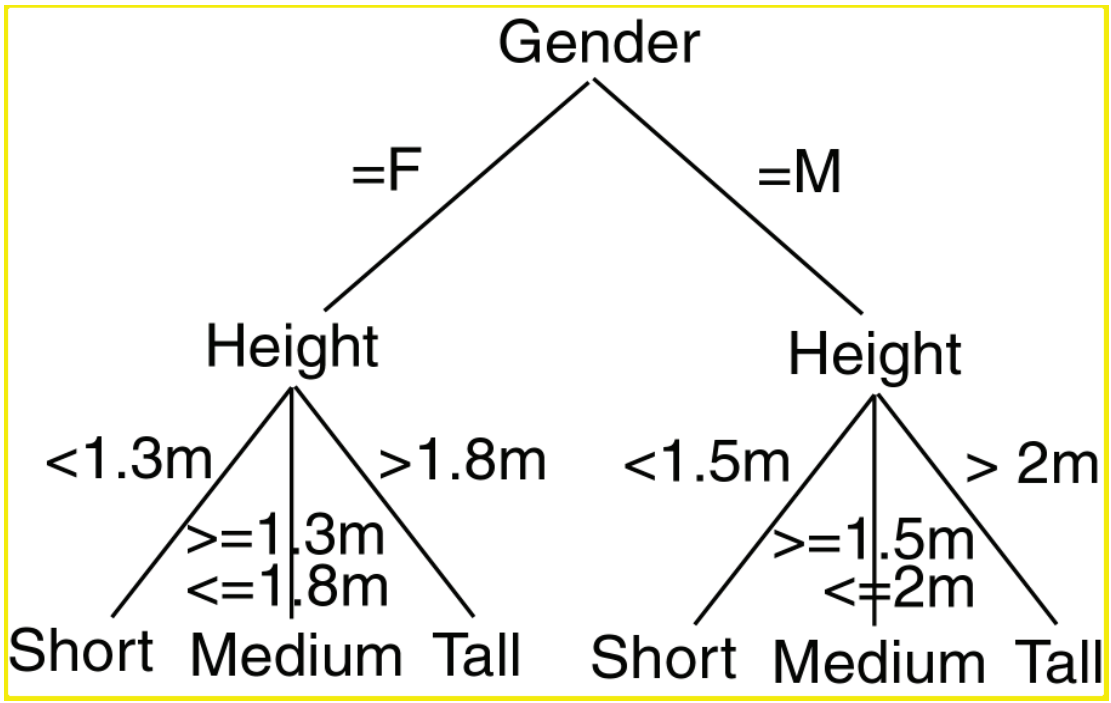
Twenty Questions Game



Decision Trees

- ***Decision Tree (DT):***
  - Tree where the root and each internal node is labeled with a question.
  - The arcs represent each possible answer to the associated question.
  - Each leaf node represents a prediction of a solution to the problem.
- Popular technique for classification; Leaf node indicates class to which the corresponding tuple belongs.

Decision Tree Example



Decision Trees

- A ***Decision Tree Model*** is a computational model consisting of three parts:
  - Decision Tree
  - Algorithm to create the tree
  - Algorithm that applies the tree to data
- Creation of the tree is the most difficult part.
- Processing is basically a search similar to that in a binary search tree (although DT may not be binary).

Decision Tree Algorithm

```
Input:
   T      //Decision Tree
   D      //Input Database
Output:
   M      //Model Prediction
DTProc Algorithm:
                //Illustrates Prediction Technique using DT
   for each t ∈ D do
      n = root node of T;
      while n not leaf node do
         Obtain answer to question on n applied t;
         Identify arc from t which contains correct answer;
         n = node at end of this arc;
      Make prediction for t based on labeling of n;
```

DT Advantages/Disadvantages

- Advantages:
  - Easy to understand.
  - Easy to generate rules
- Disadvantages:
  - May suffer from overfitting.
  - Classifies by rectangular partitioning.
  - Does not easily handle nonnumeric data.
  - Can be quite large – pruning is necessary.

CLUSTERING

Clustering Definition

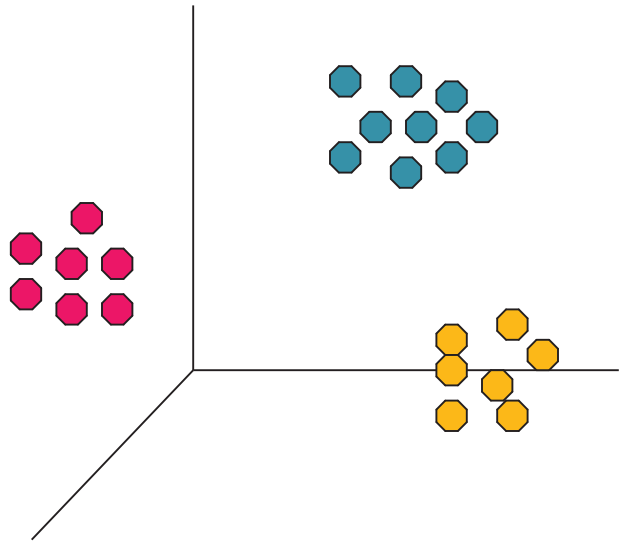
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

Illustrating Clustering

- Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized



Clustering: Application 1

- Market Segmentation:
  - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - Approach:
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering: Application 2

- Document Clustering:
  - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
  - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
  - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

ASSOCIATION RULE MINING

Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection;
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Association Rule Discovery: Application I

- Marketing and Sales Promotion:
  - Let the rule discovered be

    {Bagels, … } --> {Potato Chips}
  - Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
  - Bagels in the antecedent => Can be used to see which products would be affected if the store discontinues selling bagels.
  - Bagels in antecedent *and* Potato chips in consequent => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

Association Rule Discovery: Application 2

- Supermarket shelf management.
  - Goal: To identify items that are bought together by sufficiently many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - A classic rule --
    - If a customer buys diaper and milk, then he is very likely to buy beer:

$$Diapers \rightarrow Beer,\ support = 20\%,\ confidence = 85\%$$

CLASSIFICATION

Classification: Definition

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

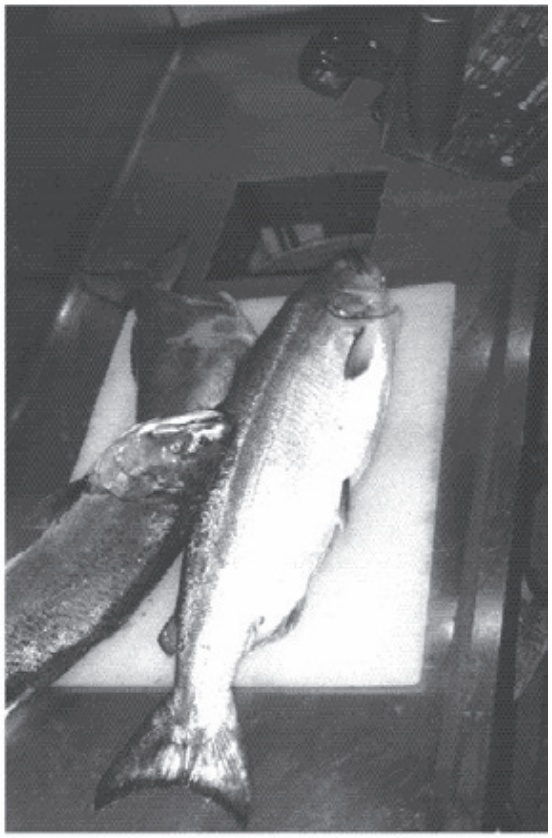
An Example

(from *Pattern Classification by* Duda & Hart & Stork – Second Edition, 2001)

- A fish-packing plant wants to automate the process of sorting incoming fish according to species

- As a pilot project, it is decided to try to separate sea bass from salmon using optical sensing

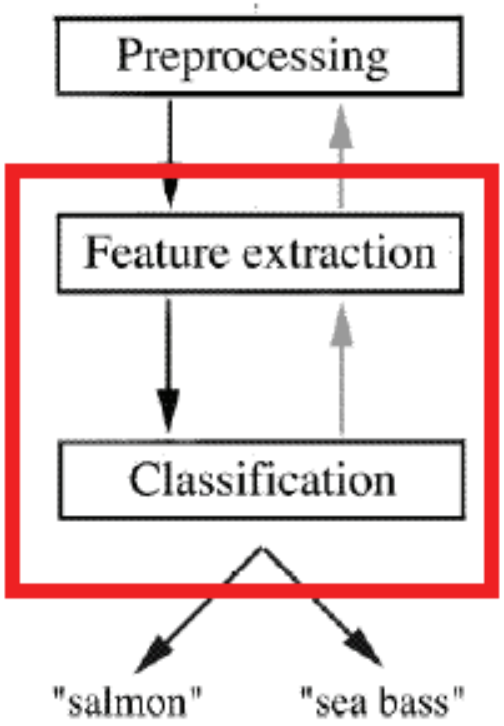
An Example (continued)



Features (to distinguish):

- Length
- Lightness
- Width
- Position of mouth

An Example (continued)



- **Preprocessing:** Images of different fishes are isolated from one another and from background;
- **Feature extraction:** The information of a single fish is then sent to a feature extractor, that measure certain "features" or "properties";
- **Classification:** The values of these features are passed to a classifier that evaluates the evidence presented, and build a model to discriminate between the two species

An Example (continued)

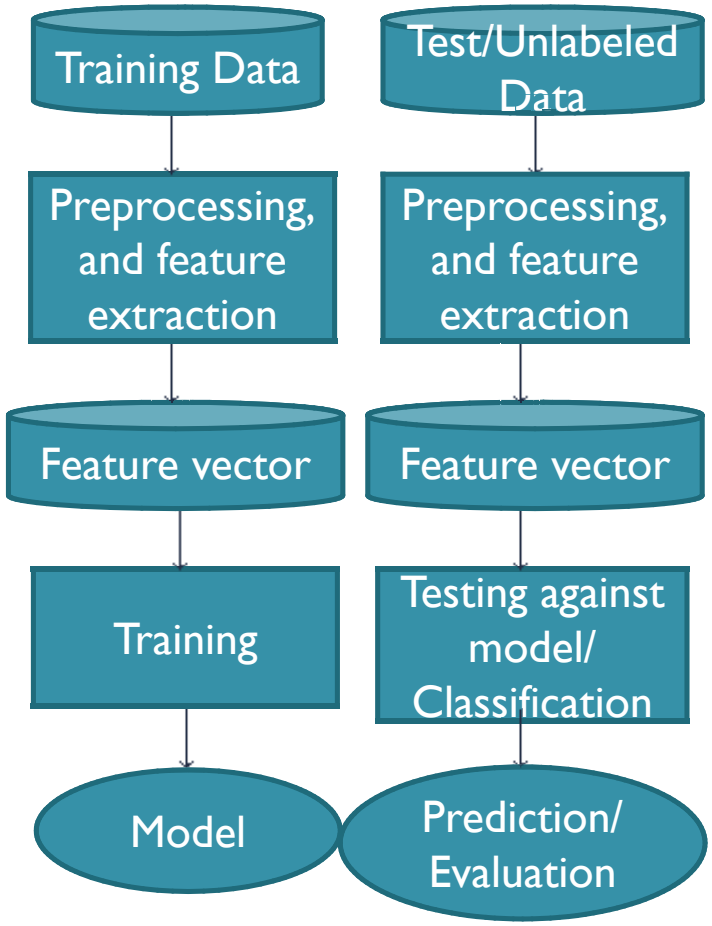
- Domain knowledge:
  - A sea bass is generally longer than a salmon
- Related feature: (or attribute)
  - Length
- Training the classifier:
  - Some examples are provided to the classifier in this form: <fish_length, fish_name>
  - These examples are called training examples
  - The classifier *learns* itself from the training examples, how to distinguish Salmon from Bass based on the *fish_length*

An Example (continued)

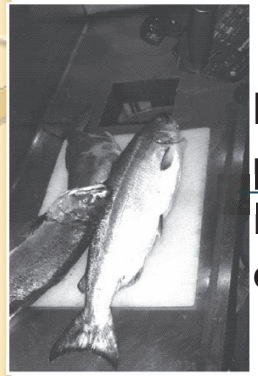
- Classification model (hypothesis):
  - The classifier generates a model from the training data to classify future examples (test examples)
  - An example of the model is a rule like this:
  - If *Length* >= *l* then sea bass* otherwise *salmon*
  - Here the value of *l** determined by the classifier
- Testing the model
  - Once we get a model out of the classifier, we may use the classifier to test future examples
  - The test data is provided in the form <fish_length>
  - The classifier outputs <fish_type> by checking *fish_length* against the model

An Example (continued)

- So the overall classification process goes like this →



An Example (continued)



Training data

Pre-processing, Feature extraction

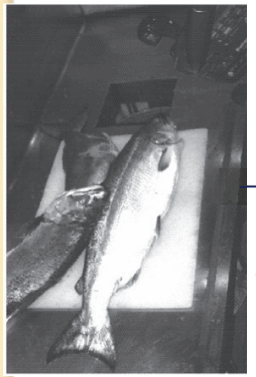
12, salmon
15, sea bass
8, salmon
5, sea bass

Feature vector

Training

If len > 12,
then sea bass
else salmon

Model



Test data

Pre-processing, Feature extraction

Labeled data

15, salmon
10, salmon

18, ?
8, ?

Unlabeled data

Feature vector

Test/ Classify

sea bass (error!)
salmon (correct)

sea bass
salmon

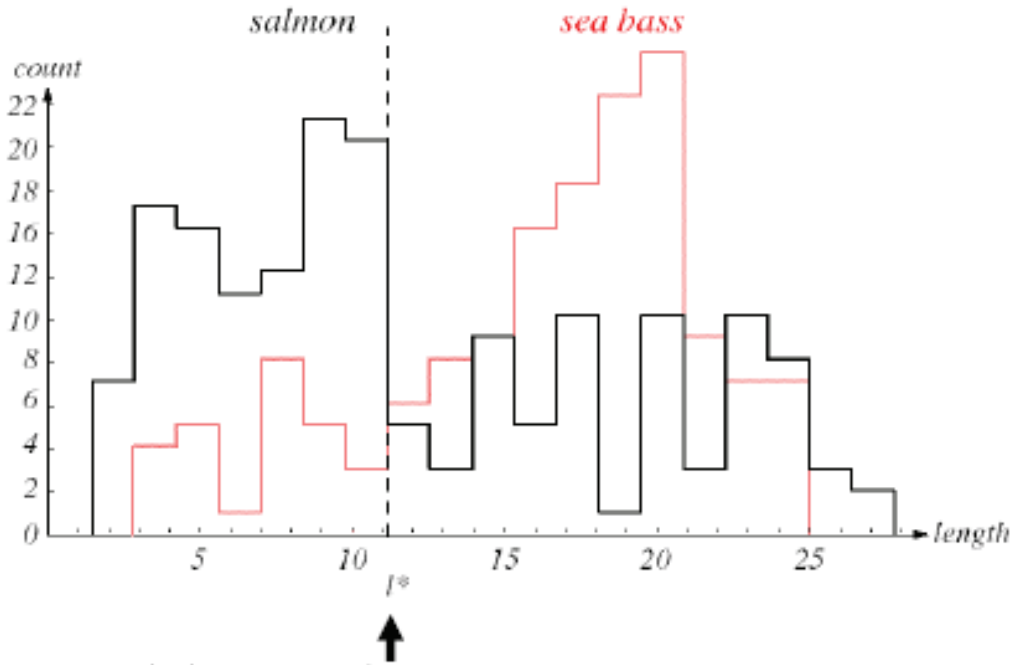
Evaluation/Prediction

An Example (continued)

- Why error?
  - Insufficient training data
  - Too few features
  - Too many/irrelevant features
  - Overfitting / specialization

An Example (continued)

Histograms of the length feature for the two categories



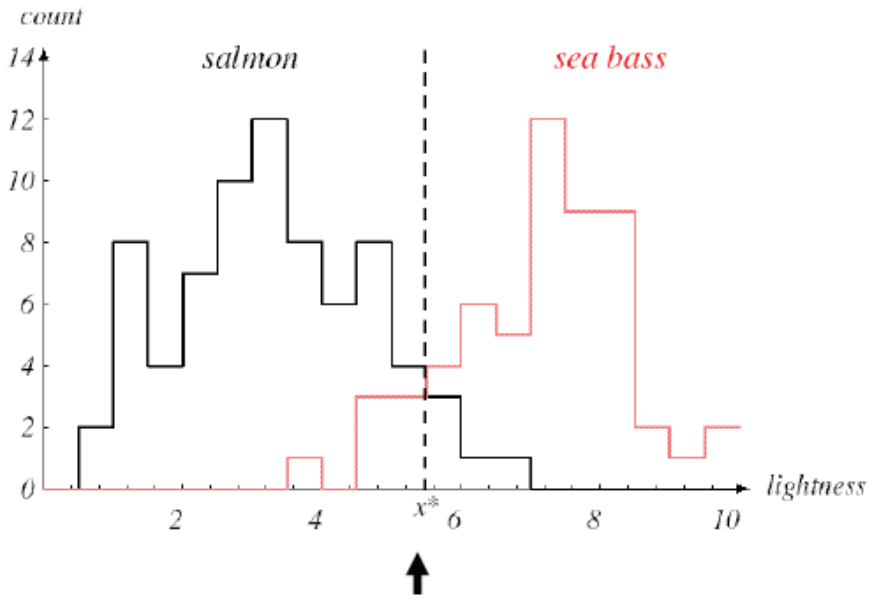
Leads to the smallest number of errors on average

We cannot reliably separate sea bass from salmon by length alone!

An Example (continued)

- New Feature:
  - *Average lightness of the fish scales*

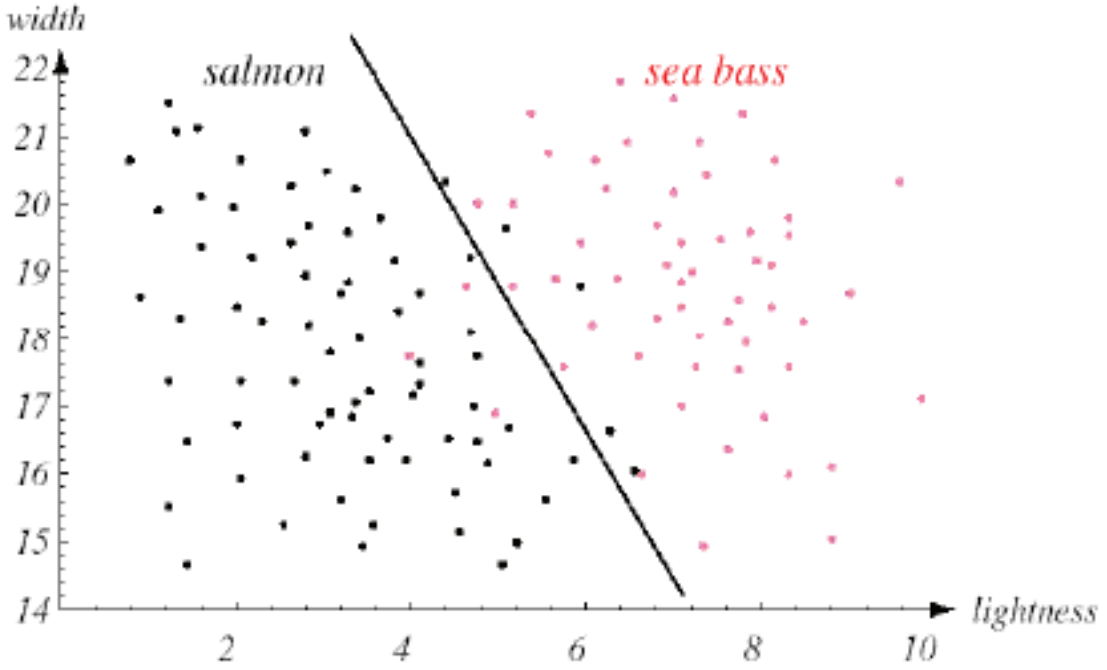
Histograms of the lightness feature for the two categories



Leads to the smallest number of errors on average

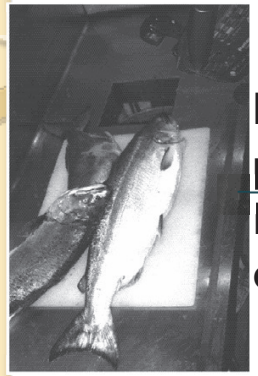
The two classes are much better separated!

An Example (continued)



Decision rule: Classify the fish as a sea bass if its feature vector falls above the decision boundary shown, and as salmon otherwise

An Example (continued)



Training data

Pre-processing, Feature extraction

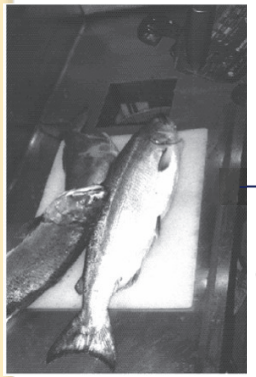
12, 4, salmon
15, 8, sea bass
8, 2, salmon
5, 10, sea bass

Feature vector

Training

If ltns > 6 **or** len*5+ltns*2>100 ***then*** sea bass ***else*** salmon

Model



Test data

Pre-processing, Feature extraction

15, 2, salmon
10, 7, salmon
18, 7, ?
8, 5, ?

Feature vector

Test/ Classify

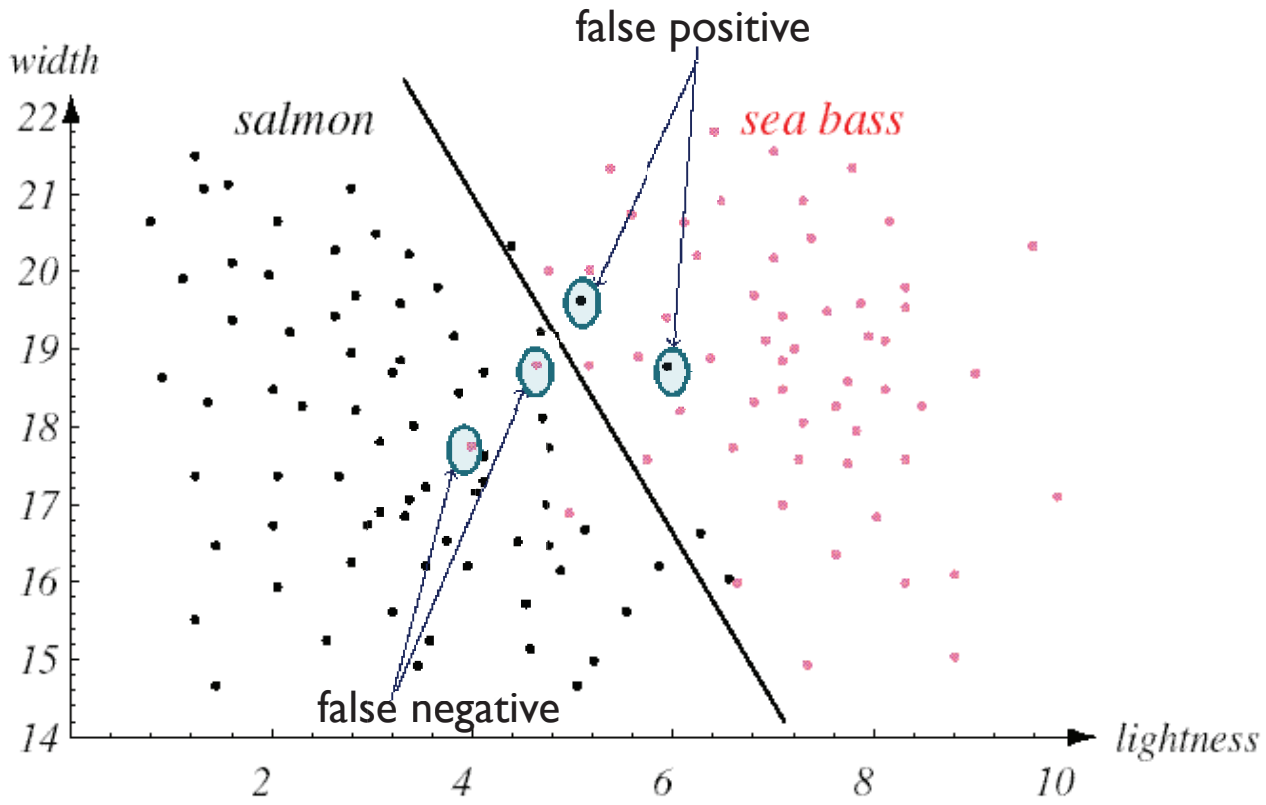
salmon (correct)
salmon (correct)
sea bass
salmon

Evaluation/Prediction

Terms

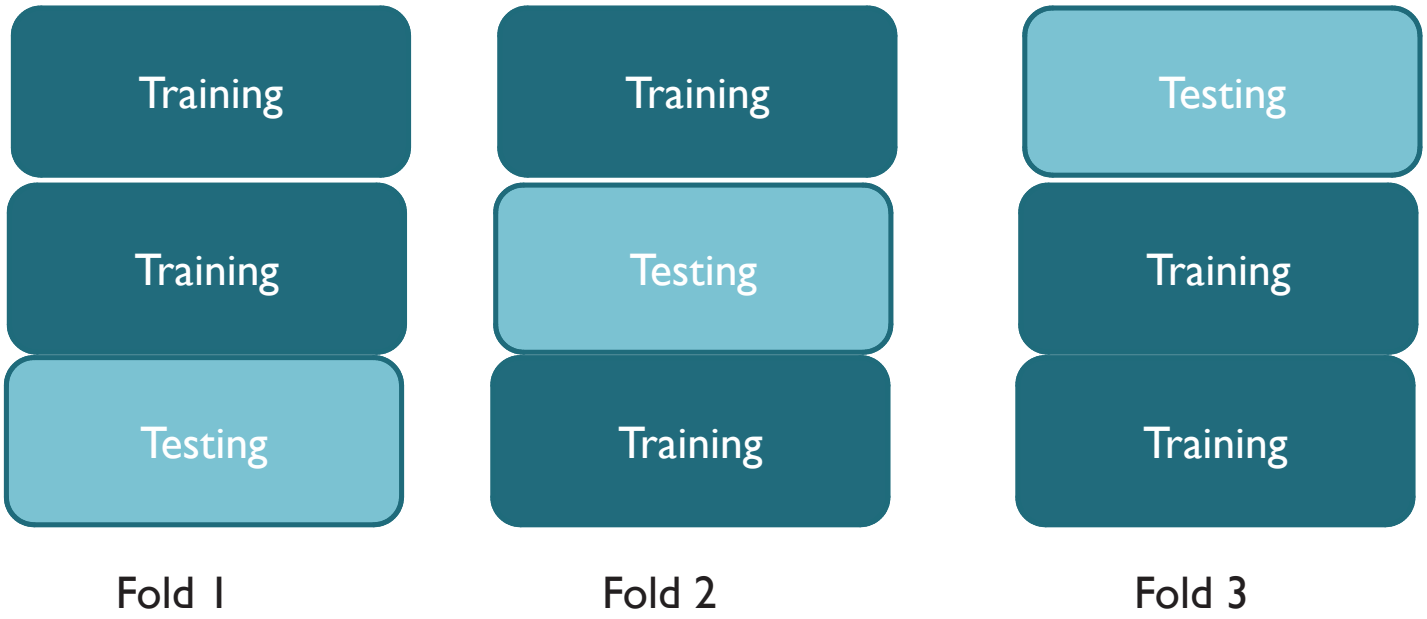
- Accuracy:
  - % of test data correctly classified
  - In our first example, accuracy was 3 out 4 = 75%
  - In our second example, accuracy was 4 out 4 = 100%
- False positive:
  - Negative class incorrectly classified as positive
  - Usually, the larger class is the negative class
  - Suppose
    - **salmon is negative** class
    - **sea bass is positive class**

Terms


false positive
width
salmon
sea bass
22
21
20
19
18
17
16
15
14
false negative
2
4
6
8
10
lightness

Terms

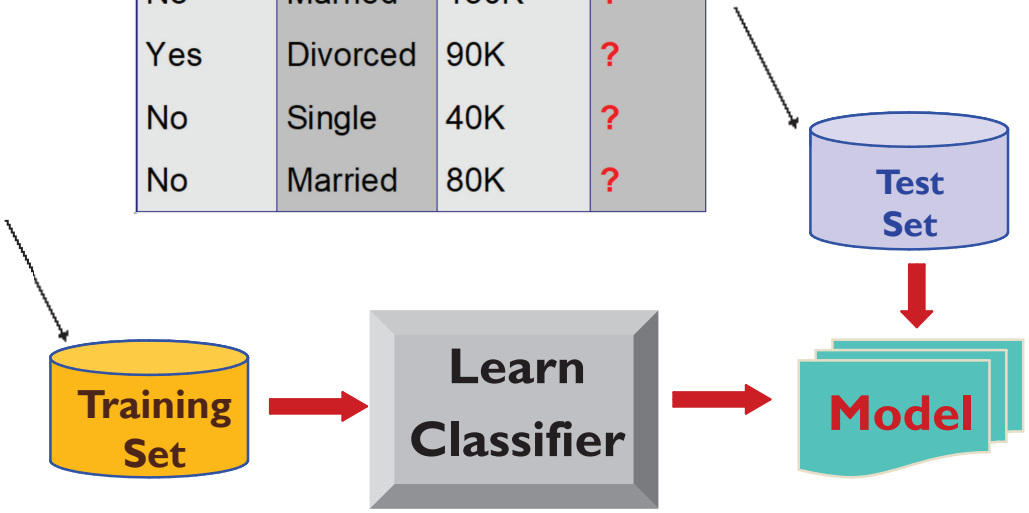
- Cross validation (3 fold)



Classification Example 2

Tid	Refund	Marital Status	Taxable Income	Cheat
	categorical	categorical	continuous	class
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

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Classification: Application 1

- Direct Marketing
  - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
  - Approach:
    - Use the data for a similar product introduced before.
    - We know which customers decided to buy and which decided otherwise. This {*buy, don't buy*} decision forms the *class attribute*.
    - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
      - Type of business, where they stay, how much they earn, etc.
    - Use this information as input attributes to learn a classifier model.

Classification: Application 2

- Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - Use credit card transactions and the information on its account-holder as attributes.
      - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions as fraud or fair transactions. This forms the class attribute.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

Classification: Application 3

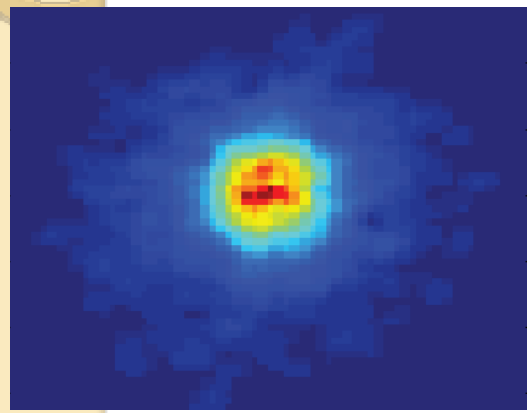
- Customer Attrition/Churn:
  - Goal: To predict whether a customer is likely to be lost to a competitor.
  - Approach:
    - Use detailed record of transactions with each of the past and present customers, to find attributes.
      - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
    - Label the customers as loyal or disloyal.
    - Find a model for loyalty.

Classification: Application 4

- Sky Survey Cataloging
  - Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
    - 3000 images with 23,040 x 23,040 pixels per image.
  - Approach:
    - Segment the image.
    - Measure image attributes (features) - 40 of them per object.
    - Model the class based on these features.
    - Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

Classifying Galaxies

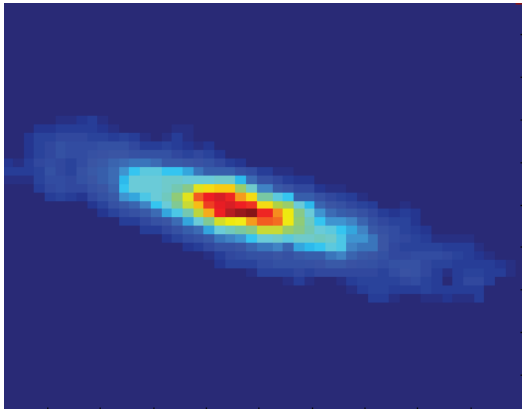
Early



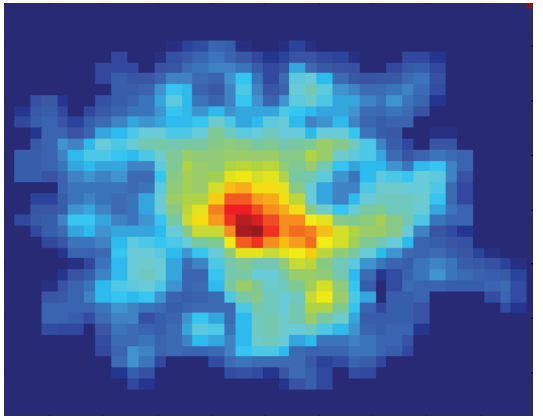
Class:
- **Stages of Formation**

Attributes:
- **Image features,**
- **Characteristics of light waves received, etc.**

Intermediate



Late



Data Size:
- **72 million stars, 20 million galaxies**
- **Object Catalog: 9 GB**
- **Image Database: 150 GB**

SOME CLASSIFICATION TECHNIQUES

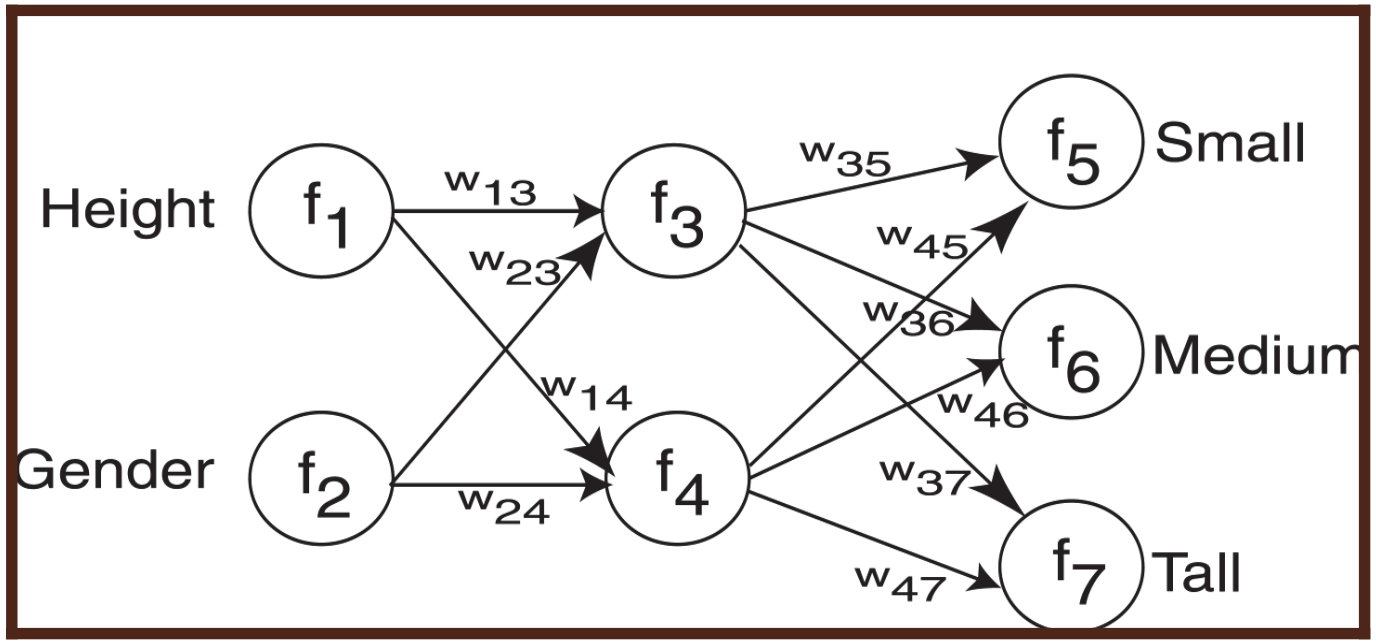
Neural Networks

- Based on observed functioning of human brain.
- ***(Artificial Neural Networks (ANN)***
- Our view of neural networks is very simplistic.
- We view a neural network (NN) from a graphical viewpoint.
- Alternatively, a NN may be viewed from the perspective of matrices.
- Used in pattern recognition, speech recognition, computer vision, and classification.

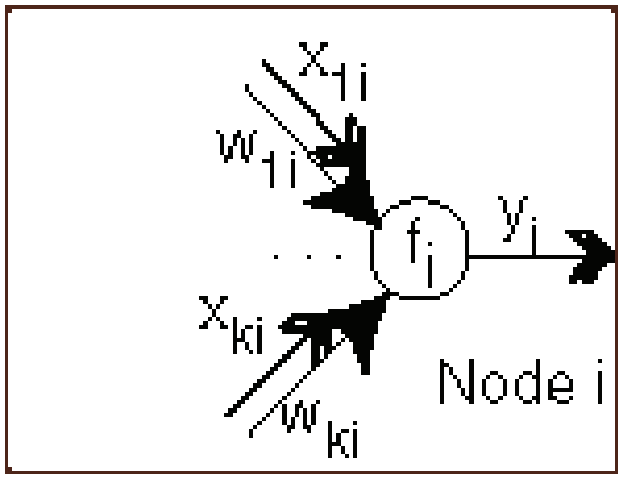
Neural Networks

- ***Neural Network (NN)*** is a directed graph $F=<V,A>$ with vertices $V=\{1,2,\ldots,n\}$ and arcs $A=\{<i,j>|1<=i,j<=n\}$, with the following restrictions:
  - V is partitioned into a set of input nodes, V_I, hidden nodes, V_H, and output nodes, V_O.
  - The vertices are also partitioned into layers
  - Any arc <i,j> must have node i in layer h-1 and node j in layer h.
  - Arc <i,j> is labeled with a numeric value w_{ij}.
  - Node i is labeled with a function f_i.

Neural Network Example



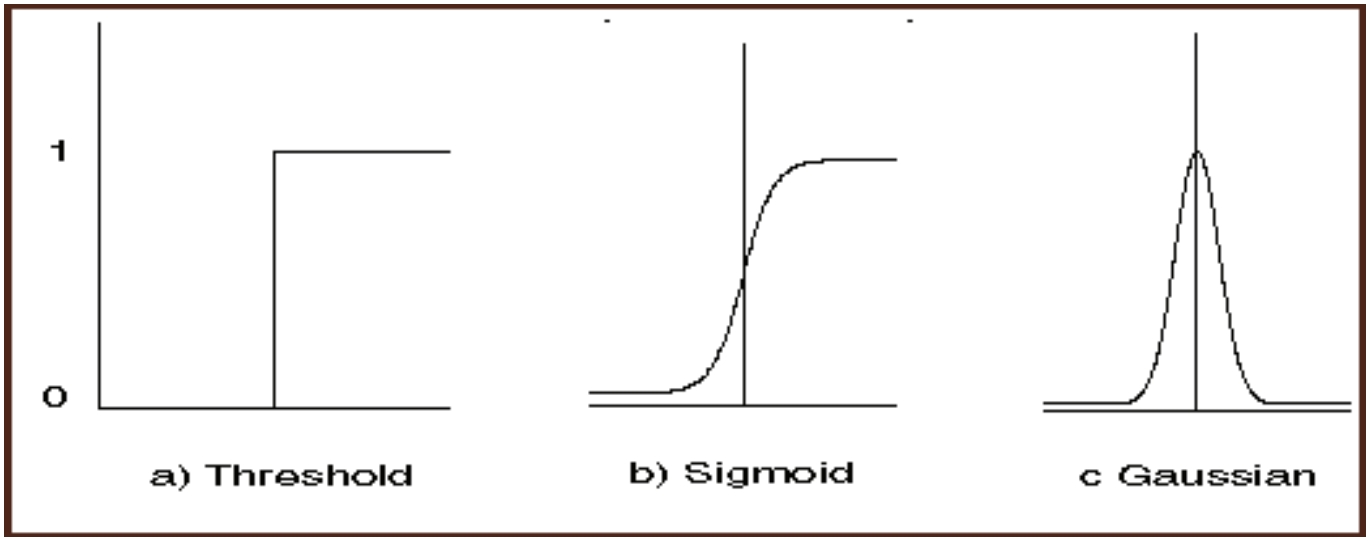
NN Node



$$y_i = f_i(\sum_{j=1}^{k} w_{ji}\ x_{ji}) = f_i([w_{1i}...w_{ki}] \begin{bmatrix} x_{1i} \\ ... \\ x_{ki} \end{bmatrix})$$

NN Activation Functions

- Functions associated with nodes in graph.
- Output may be in range [-1,1] or [0,1]



NN Activation Functions

Linear:

$$f_i(S) = c\,S$$

Threshold or Step:

$$f_i(S) = \left\{ \begin{array}{ll} 1 & if S > T \\ 0 & otherwise \end{array} \right\}$$

Ramp:

$$f_i(S) = \left\{ \begin{array}{ll} 1 & if S > T_2 \\ \frac{S-T_1}{T_2-T_1} & if T_1 \leq S \leq T_2 \\ 0 & if S < T_1 \end{array} \right\}$$

Sigmoid:

$$f_i(S) = \frac{1}{(1+e^{-c\,S})}$$

Hyperbolic Tangent:

$$f_i(S) = \frac{(1-e^{-S})}{(1+e^{-c\,S})}$$

Gaussian:

$$f_i(S) = e^{\frac{-S^2}{v}}$$

NN Learning

- Propagate input values through graph.
- Compare output to desired output.
- Adjust weights in graph accordingly.

Neural Networks

- A ***Neural Network Model*** is a computational model consisting of three parts:
  - Neural Network graph
  - Learning algorithm that indicates how learning takes place.
  - Recall techniques that determine hew information is obtained from the network.
- We will look at propagation as the recall technique.

NN Advantages

- Learning
- Can continue learning even after training set has been applied.
- Easy parallelization
- Solves many problems

NN Disadvantages

- Difficult to understand
- May suffer from overfitting
- Structure of graph must be determined a priori.
- Input values must be numeric.
- Verification difficult.

Bayes Theorem

- ***Posterior Probability:*** $P(h_1|x_i)$
- ***Prior Probability:*** $P(h_1)$
- ***Bayes Theorem:***

$$P(x_i) = \sum_{j=1}^{m} P(x_i \mid h_j)\ P(h_j).$$

$$P(h_1 \mid x_i) = \frac{P(x_i \mid h_1)\ P(h_1)}{P(x_i)}.$$

- Assign probabilities of hypotheses given a data value.

Bayes Theorem Example

- Credit authorizations (hypotheses): h_1=authorize purchase, h_2 = authorize after further identification, h_3=do not authorize, h_4= do not authorize but contact police
- Assign twelve data values for all combinations of credit and income:

	1	2	3	4
Excellent	x_1	x_2	x_3	x_4
Good	x_5	x_6	x_7	x_8
Bad	x_9	x_{10}	x_{11}	x_{12}

- From training data: $P(h_1) = 60\%$; $P(h_2)=20\%$; $P(h_3)=10\%$; $P(h_4)=10\%$.

Bayes Example(cont'd)

- Training Data:

ID	Income	Credit	Class	x_i
1	4	Excellent	h_1	x_4
2	3	Good	h_1	x_7
3	2	Excellent	h_1	x_2
4	3	Good	h_1	x_7
5	4	Good	h_1	x_8
6	2	Excellent	h_1	x_2
7	3	Bad	h_2	x_{11}
8	2	Bad	h_2	x_{10}
9	3	Bad	h_3	x_{11}
10	1	Bad	h_4	x_9

Bayes Example(cont'd)

- Calculate $P(x_i|h_j)$ and $P(x_i)$
- Ex: $P(x_7|h_1)=2/6$; $P(x_4|h_1)=1/6$; $P(x_2|h_1)=2/6$; $P(x_8|h_1)=1/6$; $P(x_i|h_1)=0$ for all other x_i.
- Predict the class for x_4:
  - Calculate $P(h_j|x_4)$ for all h_j.
  - Place x_4 in class with largest value.
  - Ex:
    - $P(h_1|x_4)=(P(x_4|h_1)(P(h_1))/P(x_4)$
      $=(1/6)(0.6)/0.1=1$.
    - x_4 in class h_1.

Hypothesis Testing

- Find model to explain behavior by creating and then testing a hypothesis about the data.
- Exact opposite of usual DM approach.
- H_0 – Null hypothesis; Hypothesis to be tested.
- H_1 – Alternative hypothesis

Chi Squared Statistic

- O – observed value
- E – Expected value based on hypothesis.

$$\chi^2 = \sum \frac{(O-E)^2}{E}.$$

- Ex:
  - O={50,93,67,78,87}
  - E=75
  - χ^2=15.55 and therefore significant

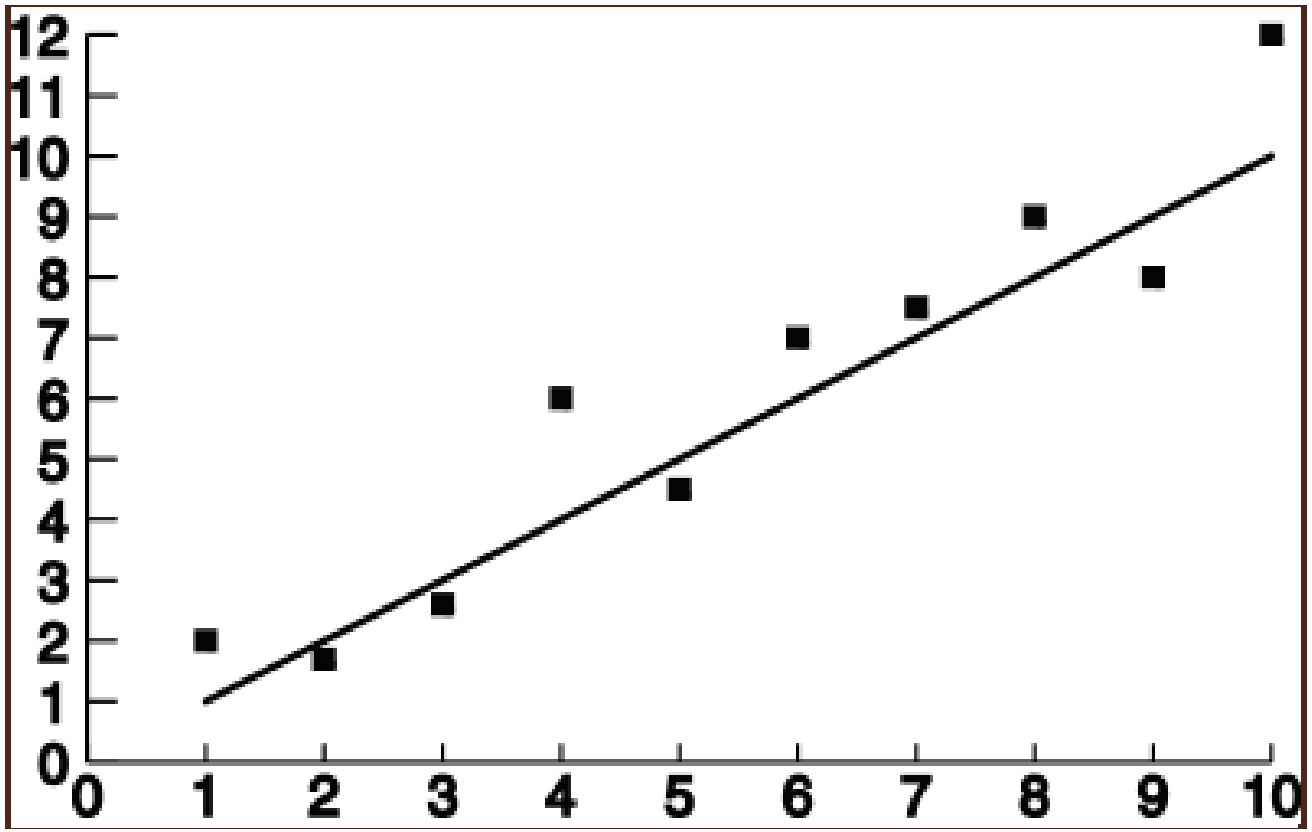
Regression

- Predict future values based on past values
- ***Linear Regression*** assumes linear relationship exists.

$$y = c_0 + c_1 x_1 + \ldots + c_n x_n$$

- Find values to best fit the data

Linear Regression



Correlation

- Examine the degree to which the values for two variables behave similarly.
- Correlation coefficient r:
  - 1 = perfect correlation
  - -1 = perfect but opposite correlation
  - 0 = no correlation

$$r = \frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum (x_i - \bar{X})^2 \sum (y_i - \bar{Y})^2}}$$

Similarity Measures

- Determine similarity between two objects.
- Similarity characteristics:

- $\forall t_i \in D, sim(t_i, t_i) = 1$
- $\forall t_i, t_j \in D, sim(t_i, t_j) = 0$ if t_i and t_j are not alike at all.
- $\forall t_i, t_j, t_k \in D, sim(t_i, t_j) < sim(t_i, t_k)$ if t_i is more like t_k than it is like t_j

- Alternatively, distance measure measure how unlike or dissimilar objects are.

Similarity Measures

Dice: $$sim(t_i, t_j) = \frac{2\sum_{h=1}^{k} t_{ih} t_{jh}}{\sum_{h=1}^{k} t_{ih}^2 + \sum_{h=1}^{k} t_{jh}^2}$$

Jaccard: $$sim(t_i, t_j) = \frac{\sum_{h=1}^{k} t_{ih} t_{jh}}{\sum_{h=1}^{k} t_{ih}^2 + \sum_{h=1}^{k} t_{jh}^2 - \sum_{h=1}^{k} t_{ih} t_{jh}}$$

Cosine: $$sim(t_i, t_j) = \frac{\sum_{h=1}^{k} t_{ih} t_{jh}}{\sqrt{\sum_{h=1}^{k} t_{ih}^2 \sum_{h=1}^{k} t_{jh}^2}}$$

Overlap: $$sim(t_i, t_j) = \frac{\sum_{h=1}^{k} t_{ih} t_{jh}}{min(\sum_{h=1}^{k} t_{ih}^2, \sum_{h=1}^{k} t_{jh}^2)}$$

Distance Measures

- Measure dissimilarity between objects

Euclidean: $dis(t_i, t_j) = \sqrt{\sum_{h=1}^{k} (t_{ih} - t_{jh})^2}$

Manhattan: $dis(t_i, t_j) = \sum_{h=1}^{k} \mid (t_{ih} - t_{jh}) \mid$