UNSW SYDNEY Australia's Global University

COMP9321

Data Services Engineering

Term 1, 2019

Week 5 Lecture 1

Which type of graph is more suitable when you want to illustrate the correlation between two variables?

Line Graph

Scatter Plot

Bar Graph

Tree Diagram

Which type of graph is more suitable when the focus of the graph is to compare size/strength of the segments to the total of each group?

Scatter Plot

Stacked Bar Graph

Bar Graph

Tree Diagram

Which statement is NOT correct about Graphs and Charts?

Both rely on a repeated pattern to show data

Charts are always restricted to numerical axes

Graphs must have at least one numerical axe

They cannot be used interchangeably

Which of the followings is correct about PCA?

PCA cannot be used in the field of Feature Selection. This application of PCA is well known for ineffective.

Data visualization: To take 3D data, and find a different way of plotting it in 2D (using k=2)

As a replacement for (or alternative to) linear regression: For most learning applications, PCA and linear regression give substantially similar results

Data compression: Reduce the dimension of your data, so that it takes up less memory/disk space. You must do this when processing your data.

Given *d*-dimensional data *X* , you run principle component analysis and pick P principle components. Can you always reconstruct any data point x_i for $i \in \{1 \ldots N\}$ from the *P* principle components with zero reconstruction error?

Yes, if $P < d$

Yes, if $P = d$

Yes, if $P > d$

No, you cannot

Data Analytics

COMP9321 2019T1

Wrap-up of lecture 1.1

COMP9321 2019T1

Wrap up of Lecture 1.1
What can we do with the data?

Life Sciences

Clinical research is a slow and expensive process, with trials failing for a variety of reasons. Advanced analytics, artificial intelligence (AI) and the Internet of Medical Things (IoMT) unlocks the potential of improving speed and efficiency at every stage of clinical research by delivering more intelligent, automated solutions.

Banking

Financial institutions gather and access analytical insight from large volumes of unstructured data in order to make sound financial decisions. Big data analytics allows them to access the information they need when they need it, by eliminating overlapping, redundant tools and systems.

Source: https://www.sas.com/en_au/insights/analytics/big-data-analytics.html

Wrap up of Lecture 1.1

What can we do with the data?

Manufacturing

For manufacturers, solving problems is nothing new. They wrestle with difficult problems on a daily basis - from complex supply chains, to motion applications, to labor constraints and equipment breakdowns. That's why big data analytics is essential in the manufacturing industry, as it has allowed competitive organizations to discover new cost saving opportunities and revenue opportunities.

Health Care

Big data is a given in the health care industry. Patient records, health plans, insurance information and other types of information can be difficult to manage – but are full of key insights once analytics are applied. That's why big data analytics technology is so important to heath care. By analyzing large amounts of information – both structured and unstructured – quickly, health care providers can provide lifesaving diagnoses or treatment options almost immediately.

Source: https://www.sas.com/en_au/insights/analytics/big-data-analytics.html

Wrap up of Lecture 1.1

What can we do with the data?

Government

Certain government agencies face a big challenge: tighten the budget without compromising quality or productivity. This is particularly troublesome with law enforcement agencies, which are struggling to keep crime rates down with relatively scarce resources. And that’s why many agencies use big data analytics; the technology streamlines operations while giving the agency a more holistic view of criminal activity.

Retail

Customer service has evolved in the past several years, as savvier shoppers expect retailers to understand exactly what they need, when they need it. Big data analytics technology helps retailers meet those demands. Armed with endless amounts of data from customer loyalty programs, buying habits and other sources, retailers not only have an in-depth understanding of their customers, they can also predict trends, recommend new products – and boost profitability.

Source: https://www.sas.com/en_au/insights/analytics/big-data-analytics.html

Wrap up of Lecture 1.1

What can we do with the data?

Spam/False Information Detection

Credit card fraud detection

Recommendation systems

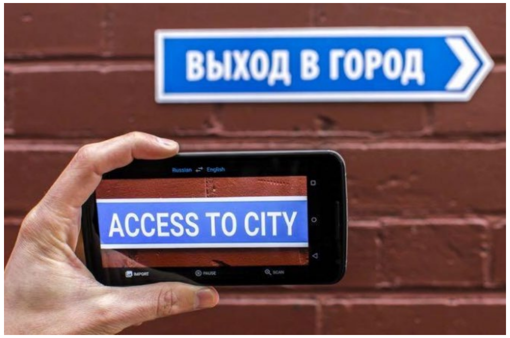
Human activity recognition/prediction

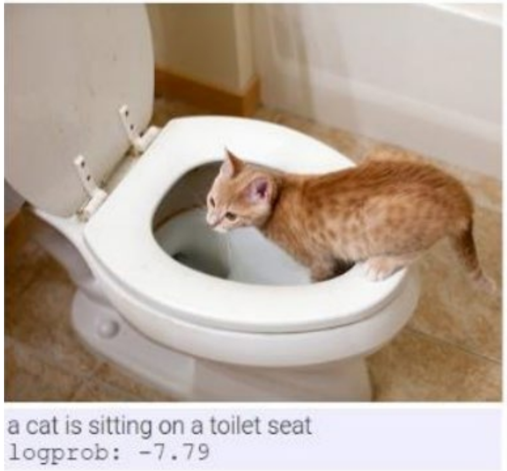
Machine translation

Face/Scene recognition

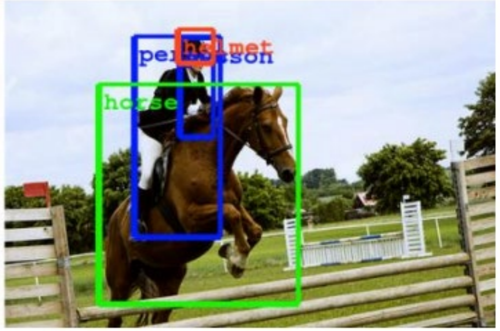
Image caption

Self-driving cars





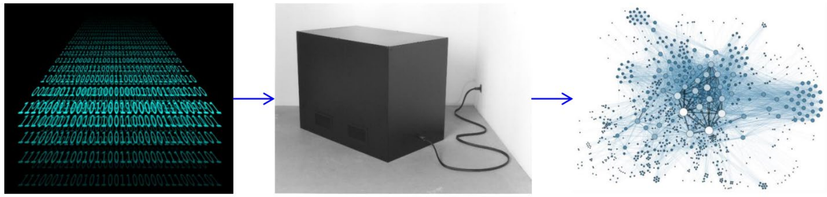
a cat is sitting on a toilet seat
logprob: -7.79

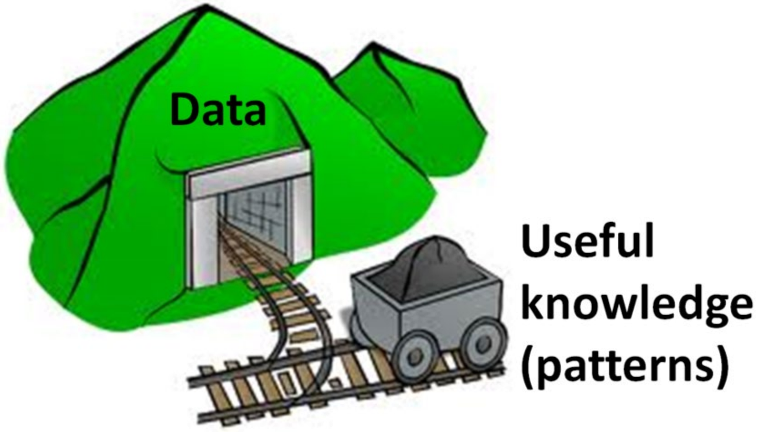


Wrap up of Lecture 1.1

Tool: Data Mining (DM)

Automatically extract useful knowledge from large datasets

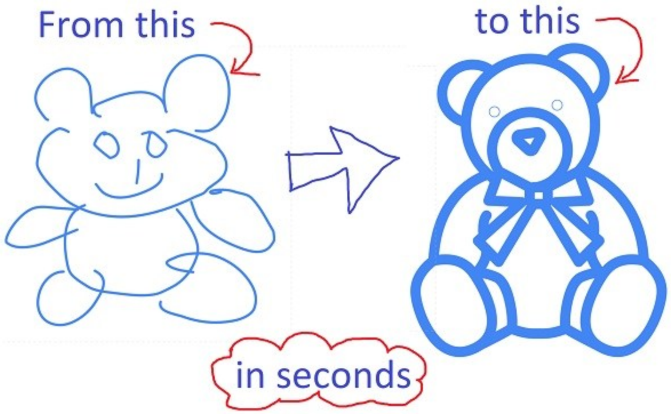




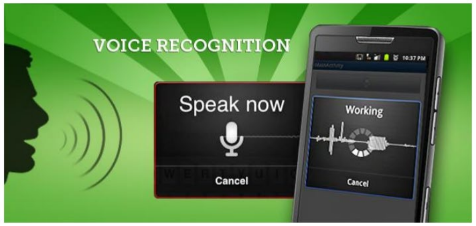
Usually, to help with human decision making

Wrap up of Lecture 1.1

Tool: Machine Learning (ML)



Using computer to automatically detect patterns in data and use these to make predictions or decisions.

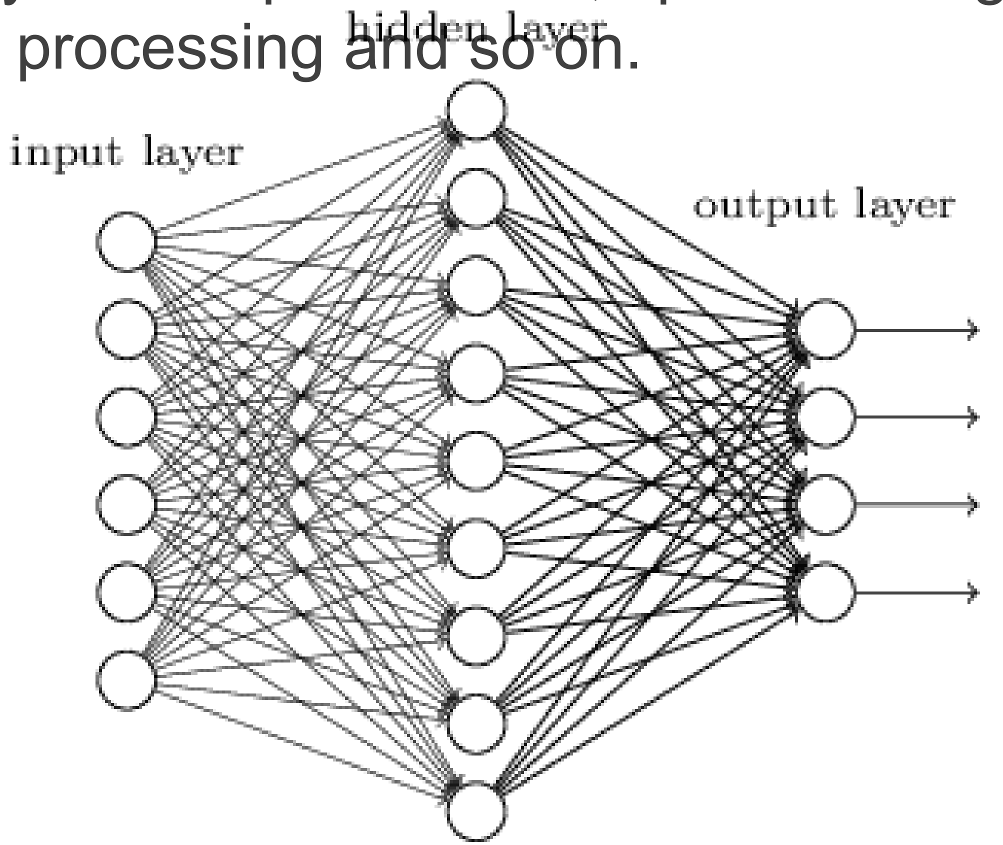


Most useful when: – We want to automate something a human can do. – We want to do things a human can’t do (look at 1 TB of data)

Wrap up of Lecture 1.1

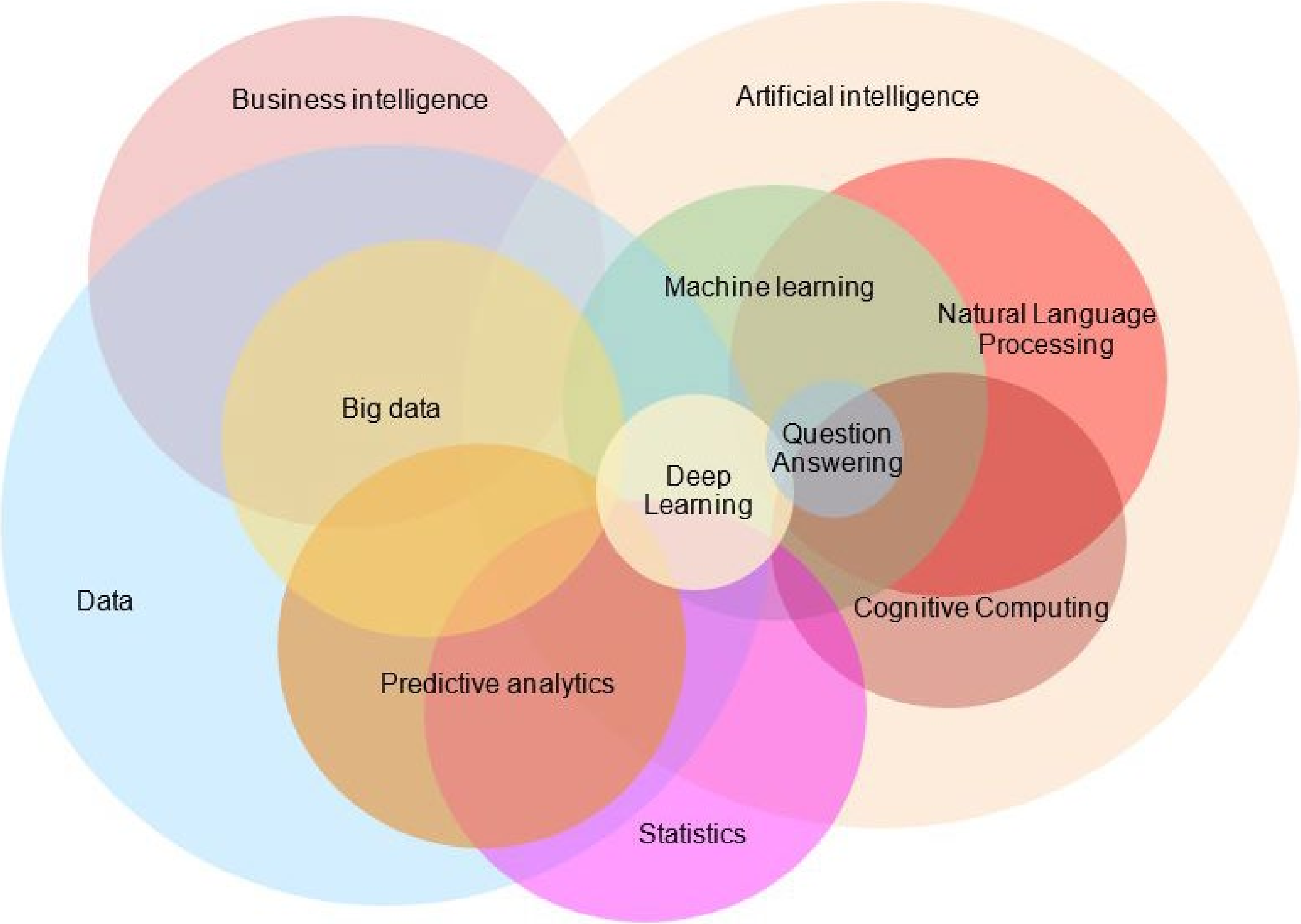
Tool: Deep Learning (DL)

Deep learning is part of machine learning. Deep learning use widely on computer vision, speech recognition, natural language processing and so on.



Wrap up of Lecture 1.1

Deep Learning vs. ML vs. AI



Traditional we've viewed ML as a subset of AI.

» And "deep learning" as a subset of ML.

Overview of Data Mining

COMP9321 2019T1

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]

Basics of Probability

Probabilities

Important Terms

Probability – the chance that an uncertain event will occur (always between 0 and 1)

Event – Each possible type of occurrence or outcome

Simple Event – an event that can be described by a single characteristic

Sample Space – the collection of all possible events

Assessing Probability

There are three approaches to assessing the probability of un uncertain event:

1. *a priori* classical probability

$$\text{probability of occurrence} = \frac{X}{T} = \frac{\text{number of ways the event can occur}}{\text{total number of elementary outcomes}}$$

2. empirical classical probability

$$\text{probability of occurrence} = \frac{\text{number of favorable outcomes observed}}{\text{total number of outcomes observed}}$$

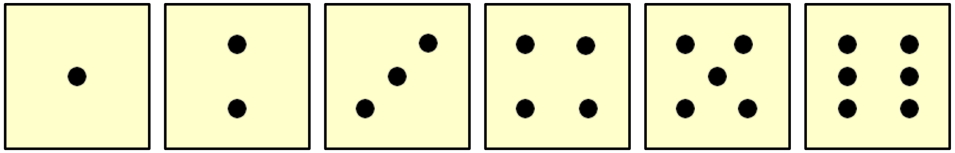
3. subjective probability

an individual judgment or opinion about the probability of occurrence

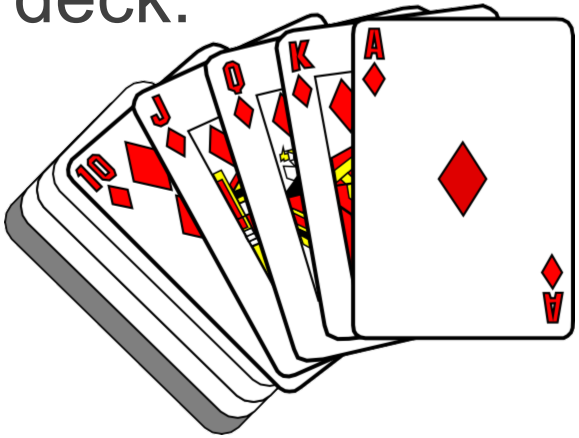
Sample Space

The Sample Space is the collection of all possible events

e.g. All 6 faces of a die:



e.g. All 52 cards of a bridge deck:



Events

Simple event

An outcome from a sample space with one characteristic

e.g., A red card from a deck of cards

Complement of an event A (denoted A')

All outcomes that are not part of event A

e.g., All cards that are not diamonds

Joint event

Involves two or more characteristics simultaneously

e.g., An ace that is also red from a deck of cards

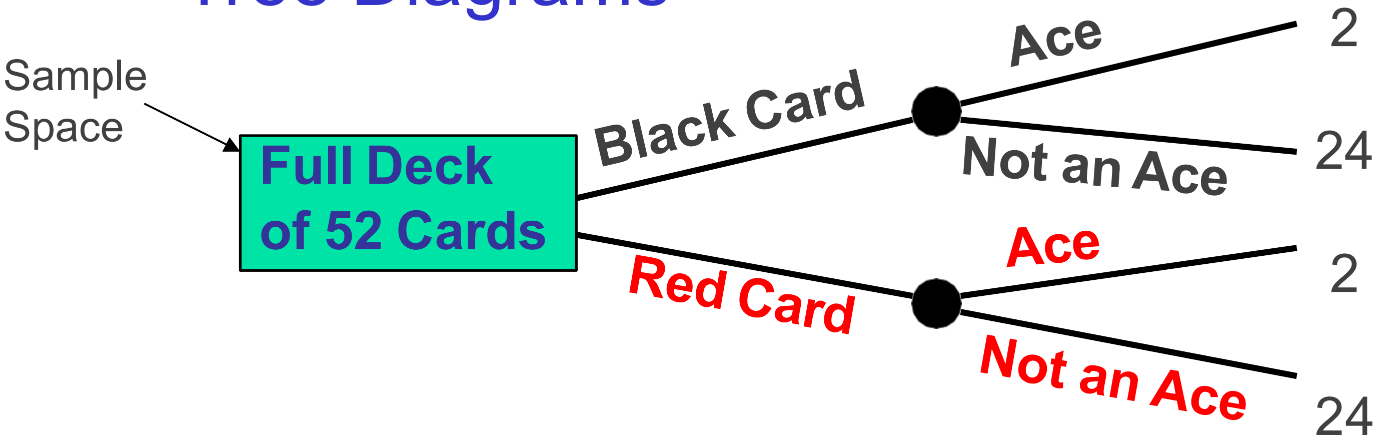
Visualizing Events

Contingency Tables

	Ace	Not Ace	Total
Black	2	24	26
Red	2	24	26
Total	4	48	52

Sample Space

Tree Diagrams



Mutually Exclusive Events

Mutually exclusive events

Events that cannot occur together

example:

A = queen of diamonds; B = queen of clubs

Events A and B are mutually exclusive

Collectively Exhaustive Events

Collectively exhaustive events

One of the events must occur

The set of events covers the entire sample space

example:

A = aces; B = black cards;
C = diamonds; D = hearts

Events A, B, C and D are collectively exhaustive (but not mutually exclusive – an ace may also be a heart)

Events B, C and D are collectively exhaustive and also mutually exclusive

Probability

Probability is the numerical measure of the likelihood that an event will occur

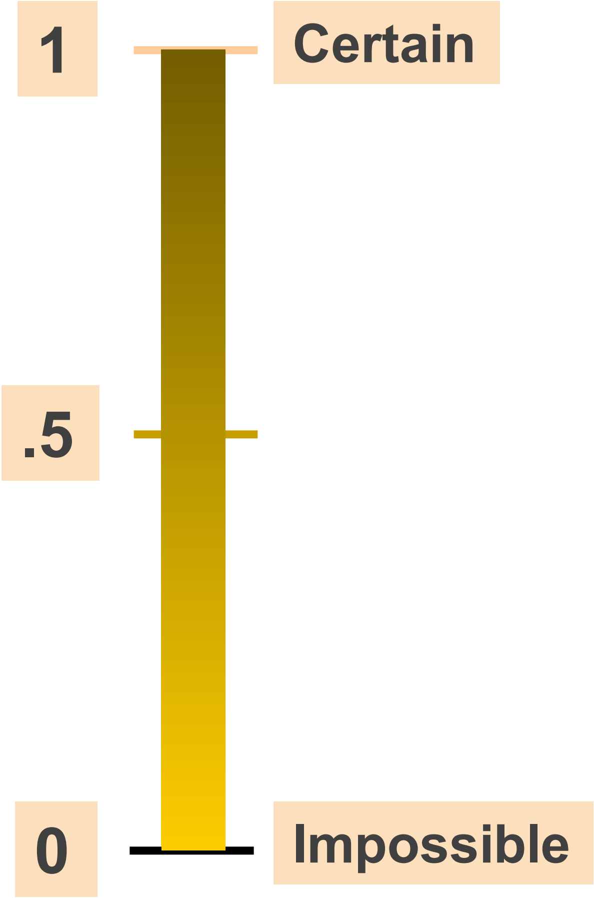
The probability of any event must be between 0 and 1, inclusively

$0 \le P(A) \le 1$ For any event A

The sum of the probabilities of all mutually exclusive and collectively exhaustive events is 1

$$P(A) + P(B) + P(C) = 1$$

If A, B, and C are mutually exclusive and collectively exhaustive



Computing Joint and Marginal Probabilities

The probability of a joint event, A and B:

$$\boxed{P(\text{A and B}) = \frac{\text{number of outcomes satisfying A and B}}{\text{total number of elementary outcomes}}}$$

Computing a marginal (or simple) probability:

$$P(A) = P(\text{A and } B_1) + P(\text{A and } B_2) + \cdots + P(\text{A and } B_k)$$

Where B_1 , B_2 , …, B_k are k mutually exclusive and collectively exhaustive events

Joint Probability Example

P(Red and Ace)

$$= \frac{\text{number of cards that are red and ace}}{\text{total number of cards}} = \frac{2}{52}$$

Type	Color		Total
	Red	Black	
Ace	2	2	4
Non-Ace	24	24	48
Total	26	26	52

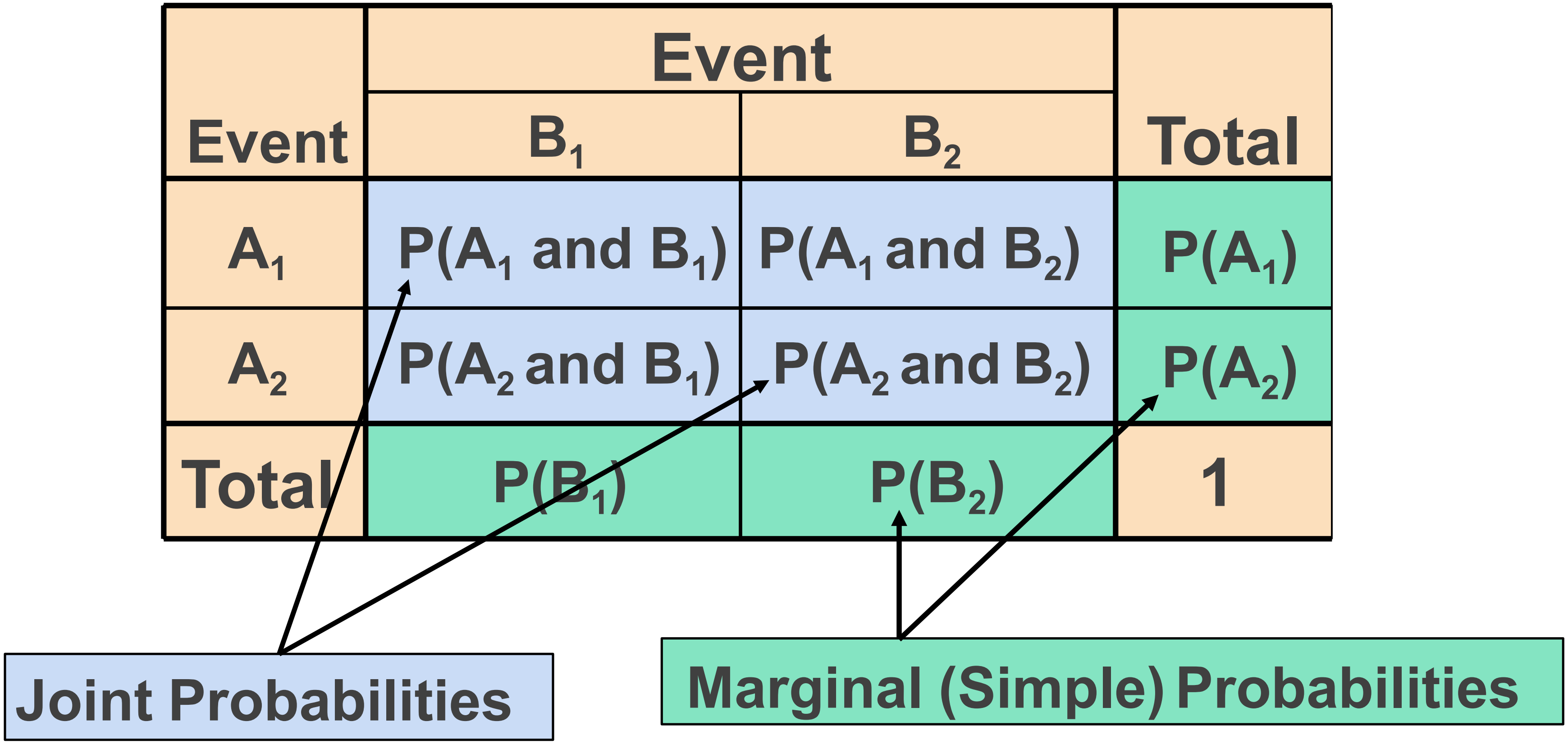
Marginal Probability Example

P(Ace)

$$= \mathrm{P(Ace\ and\ Red)} + \mathrm{P(Ace\ and\ Black)} = \frac{2}{52} + \frac{2}{52} = \frac{4}{52}$$

Type	Color		Total
	Red	Black	
Ace	2	2	4
Non-Ace	24	24	48
Total	26	26	52

Joint Probabilities Using Contingency Table



	Event		
Event	B_1	B_2	Total
A_1	P(A_1 and B_1)	P(A_1 and B_2)	P(A_1)
A_2	P(A_2 and B_1)	P(A_2 and B_2)	P(A_2)
Total	P(B_1)	P(B_2)	1

Joint Probabilities

Marginal (Simple) Probabilities

General Addition Rule

General Addition Rule:

P(A or B) = P(A) + P(B) - P(A and B)

If A and B are mutually exclusive, then

P(A and B) = 0, so the rule can be simplified:

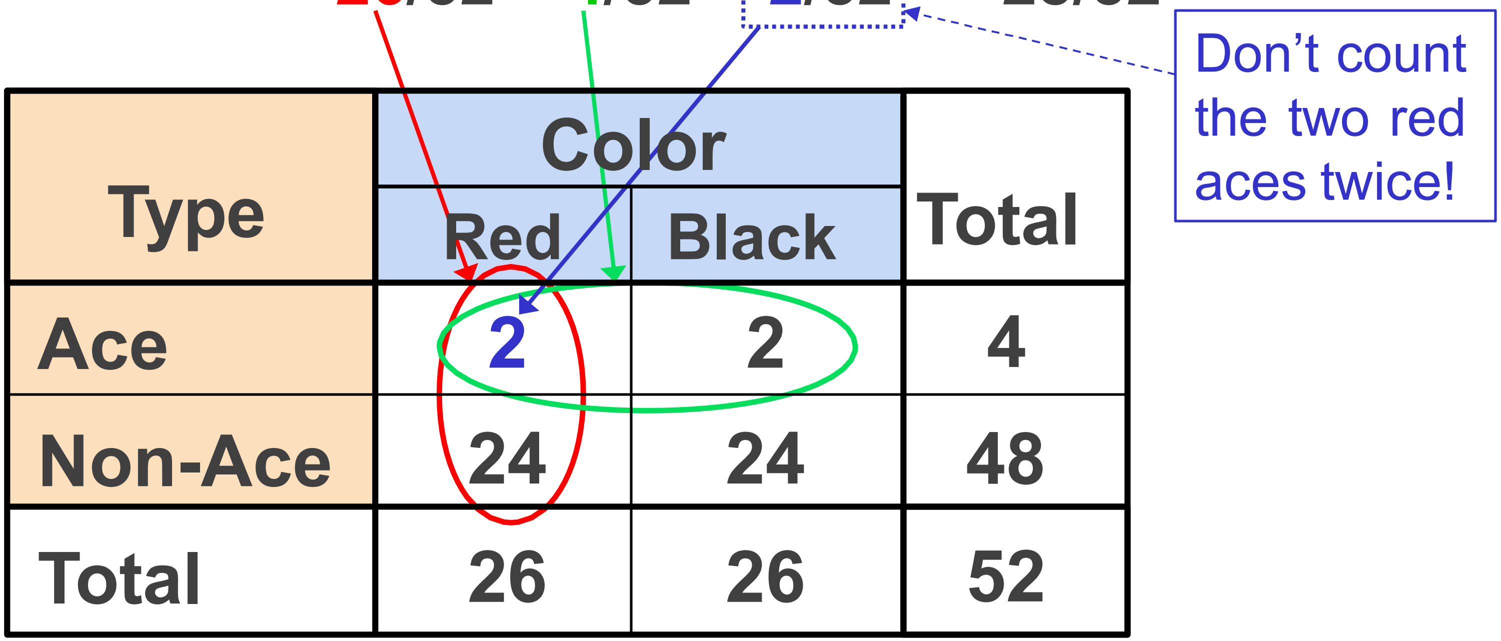
P(A or B) = P(A) + P(B)

For mutually exclusive events A and B

General Addition Rule Example

$$P(\text{Red or Ace}) = P(\text{Red}) + P(\text{Ace}) - P(\text{Red and Ace})$$

$$= 26/52 + 4/52 - \boxed{2/52} = 28/52$$



Type	Color		Total
	Red	Black	
Ace	2	2	4
Non-Ace	24	24	48
Total	26	26	52

Don't count the two red aces twice!

Computing Conditional Probabilities

A conditional probability is the probability of one event, given that another event has occurred:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

→ The conditional probability of A given that B has occurred

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

→ The conditional probability of B given that A has occurred

Where P(A and B) = joint probability of A and B
P(A) = marginal probability of A
P(B) = marginal probability of B

Conditional Probability Example

Of the cars on a used car lot, 70% have air conditioning (AC) and 40% have a CD player (CD). 20% of the cars have both.

What is the probability that a car has a CD player, given that it has AC ?

i.e., we want to find $P(CD \mid AC)$

Conditional Probability Example

(continued)

Of the cars on a used car lot, **70%** have air conditioning (AC) and **40%** have a CD player (CD). **20%** of the cars have both.

	CD	No CD	Total
AC	**.2**	.5	**.7**
No AC	.2	.1	.3
Total	**.4**	.6	1.0

$$P(\text{CD} \mid \text{AC}) = \frac{P(\text{CD and AC})}{P(\text{AC})} = \frac{.2}{.7} = .2857$$

Conditional Probability Example

(continued)

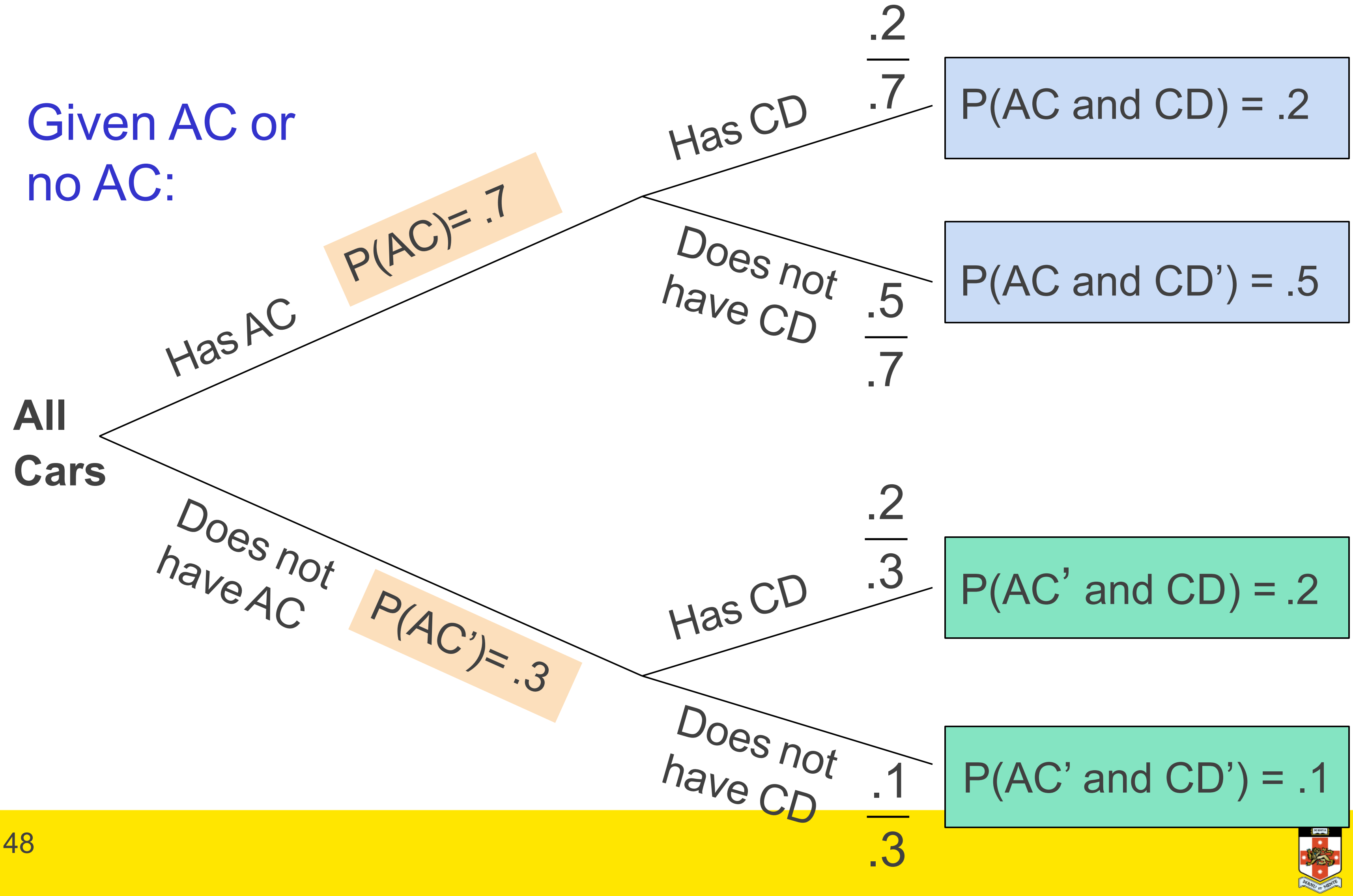
Given AC, we only consider the top row (70% of the cars). Of these, 20% have a CD player. 20% of 70% is about 28.57%.

	CD	No CD	Total
AC	.2	.5	.7
No AC	.2	.1	.3
Total	.4	.6	1.0

$$P(\text{CD} \mid \text{AC}) = \frac{P(\text{CD and AC})}{P(\text{AC})} = \frac{.2}{.7} = .2857$$

Using Decision Trees

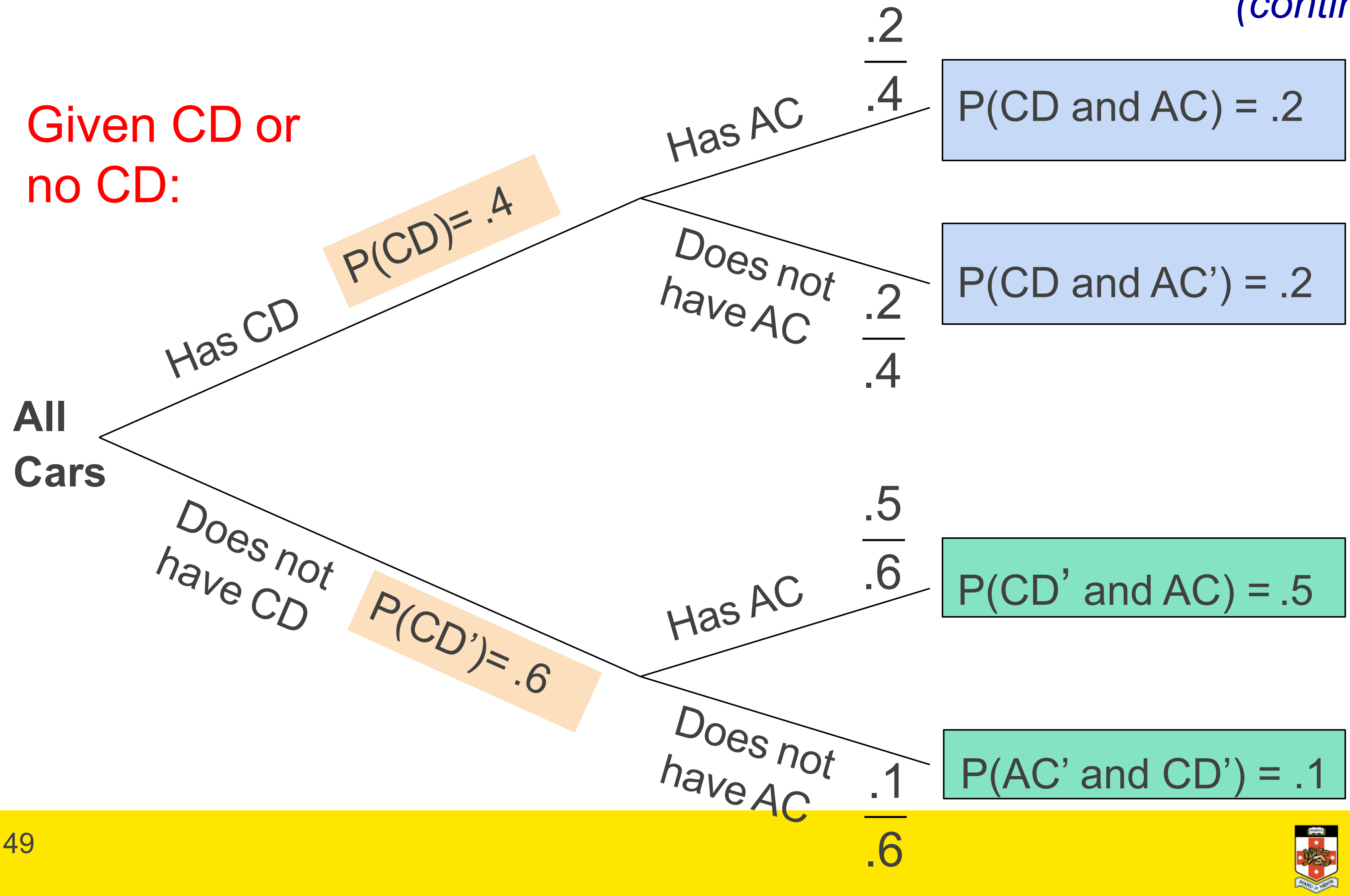
Given AC or no AC:



Using Decision Trees

(continued)

Given CD or no CD:



All Cars

Has CD

P(CD)= .4

Does not have CD

P(CD’)= .6

Has AC

$\frac{.2}{.4}$

P(CD and AC) = .2

Does not have AC

$\frac{.2}{.4}$

P(CD and AC’) = .2

Has AC

$\frac{.5}{.6}$

P(CD’ and AC) = .5

Does not have AC

$\frac{.1}{.6}$

P(AC’ and CD’) = .1

Statistical Independence

Two events are independent if and only if:

$$P(A|B) = P(A)$$

Events A and B are independent when the probability of one event is not affected by the other event

Multiplication Rules

Multiplication rule for two events A and B:

$$P(A \text{ and } B) = P(A \mid B)P(B)$$

Note: If A and B are independent, then $P(A \mid B) = P(A)$ and the multiplication rule simplifies to

$$P(A \text{ and } B) = P(A)P(B)$$

Marginal Probability

Marginal probability for event A:

$$P(A) = P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \cdots + P(A \mid B_k)P(B_k)$$

Where B_1, B_2, …, B_k are k mutually exclusive and collectively exhaustive events

Bayes' Theorem

$$P(B_i | A) = \frac{P(A | B_i)P(B_i)}{P(A | B_1)P(B_1) + P(A | B_2)P(B_2) + \ldots + P(A | B_k)P(B_k)}$$

where:

B_i = ith event of k mutually exclusive and collectively exhaustive events

A = new event that might impact $P(B_i)$

Bayes' Theorem Example

- A drilling company has estimated a 40% chance of striking oil for their new well.
- A detailed test has been scheduled for more information. Historically, 60% of successful wells have had detailed tests, and 20% of unsuccessful wells have had detailed tests.
- Given that this well has been scheduled for a detailed test, what is the probability that the well will be successful?

Bayes' Theorem Example

Let S = successful well

U = unsuccessful well

P(S) = .4 , P(U) = .6 (prior probabilities)

Define the detailed test event as D

Conditional probabilities:

P(D|S) = .6 P(D|U) = .2

Goal is to find P(S|D)

Bayes' Theorem Example

Apply Bayes' Theorem:

$$P(S \mid D) = \frac{P(D \mid S)P(S)}{P(D \mid S)P(S) + P(D \mid U)P(U)}$$

$$= \frac{(.6)(.4)}{(.6)(.4) + (.2)(.6)}$$

$$= \frac{.24}{.24 + .12} = \boxed{.667}$$

So the revised probability of success, given that this well has been scheduled for a detailed test, is .667

Bayes' Theorem Example

(continued)

Given the detailed test, the revised probability of a successful well has risen to .667 from the original estimate of .4

Event	Prior Prob.	Conditional Prob.	Joint Prob.	Revised Prob.
S (successful)	.4	.6	.4*.6 = .24	.24/.36 = .667
U (unsuccessful)	.6	.2	.6*.2 = .12	.12/.36 = .333

Sum = .36

Overview of Data Mining

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

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

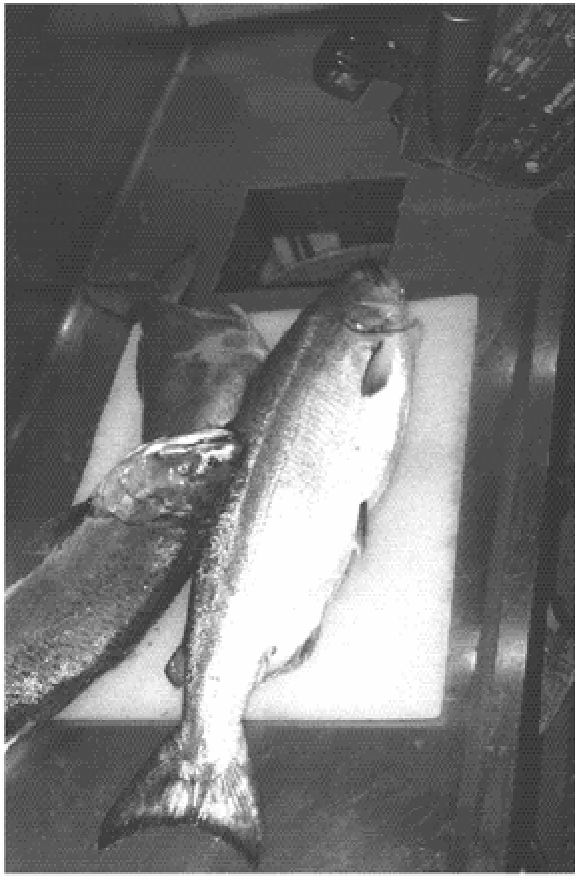
As a pilot project, it is decided to try to separate sea bass from salmon using optical sensing

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

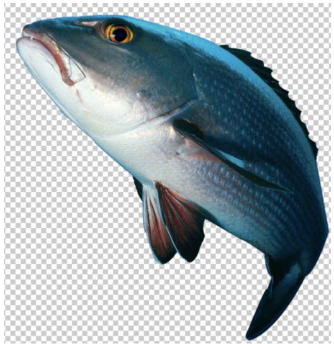
An Example (continued)

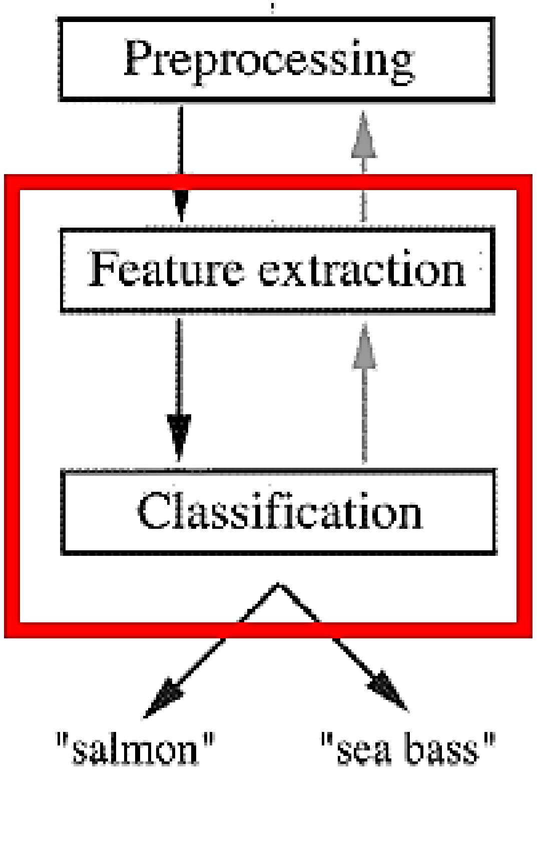
Features (to distinguish):

- 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:

- E.g., 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):

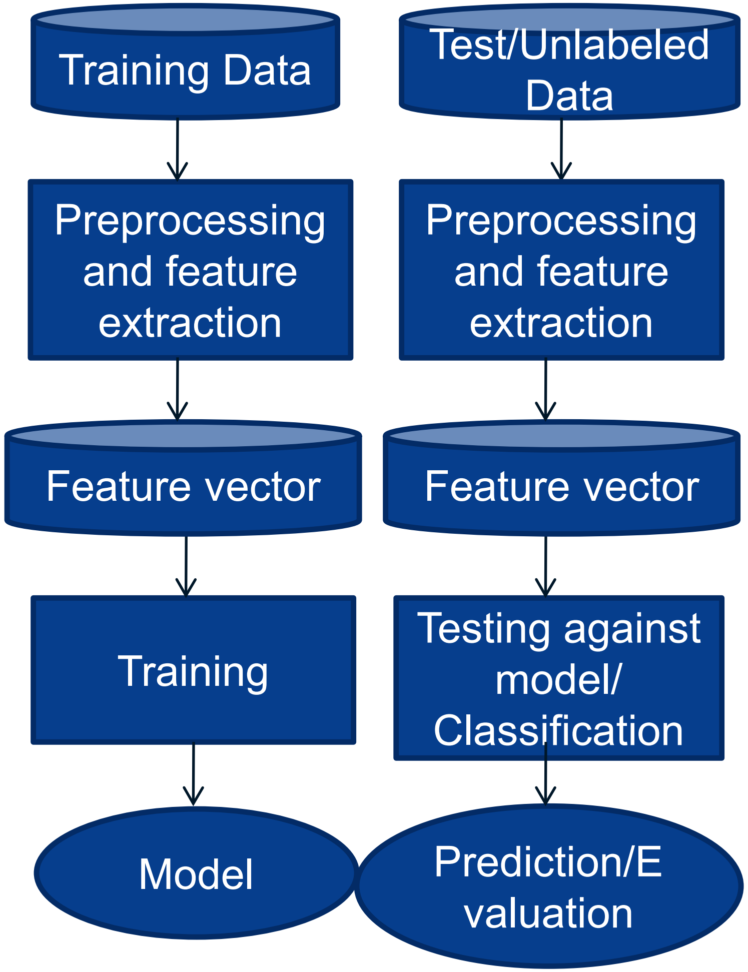
- The classifier generates a model from the training data to classify future examples (test examples)
- An example of the model is a rule like this:
- If Length >= l* then sea bass otherwise salmon
- Here the value of l* determined by the classifier

Testing the model

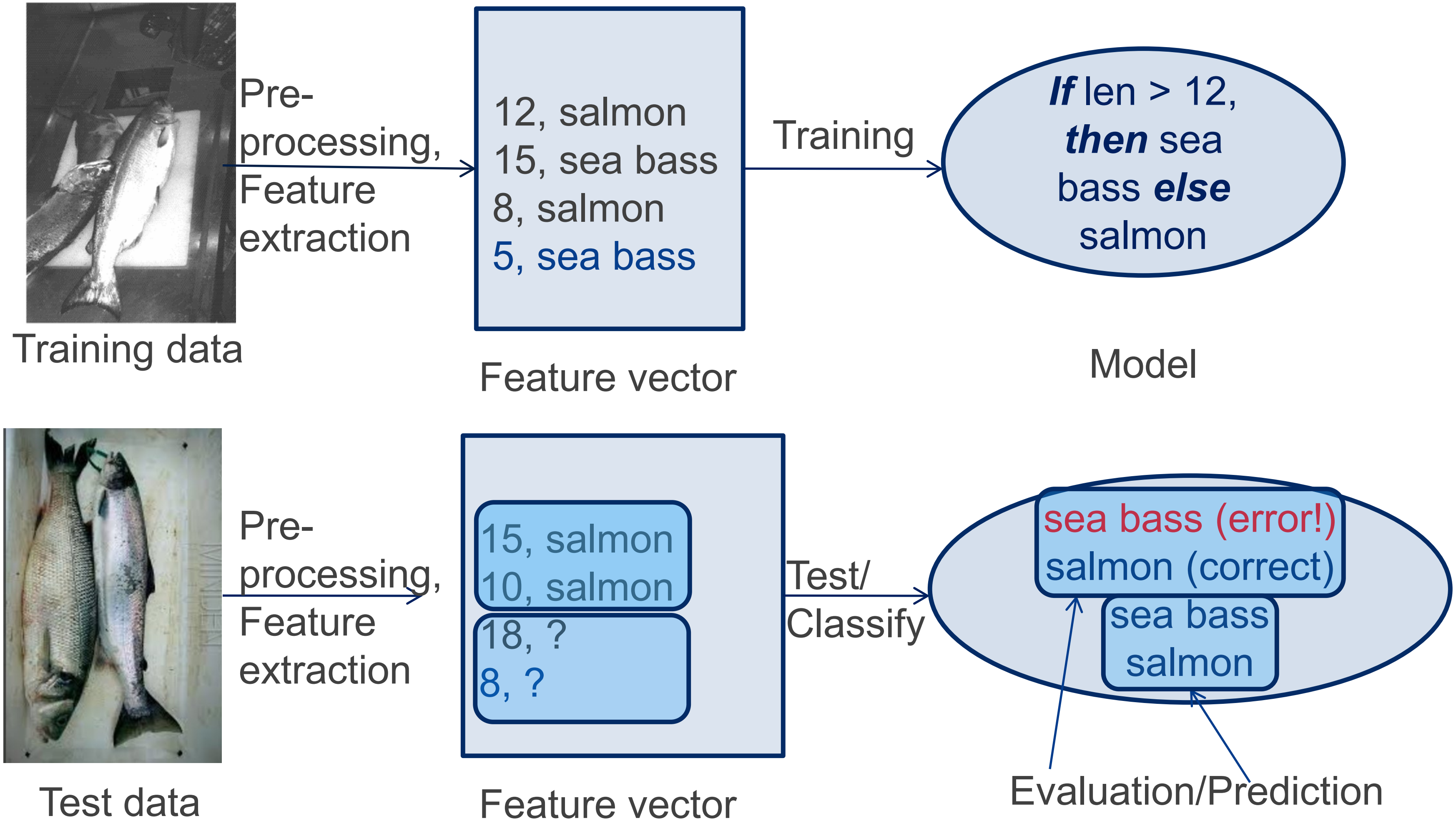
- Once we get a model out of the classifier, we may use the classifier to test future examples
- The test data is provided in the form <fish_length>
- The classifier outputs <fish_type> by checking fish_length against the model

An Example (continued)

So the overall classification process goes like this →


Training Data
Test/Unlabeled Data
Preprocessing and feature extraction
Preprocessing and feature extraction
Feature vector
Feature vector
Training
Testing against model/ Classification
Model
Prediction/E valuation

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

15, salmon
10, salmon
18, ?
8, ?

Feature vector

Test/ Classify

sea bass (error!)
salmon (correct)
sea bass
salmon

Evaluation/Prediction

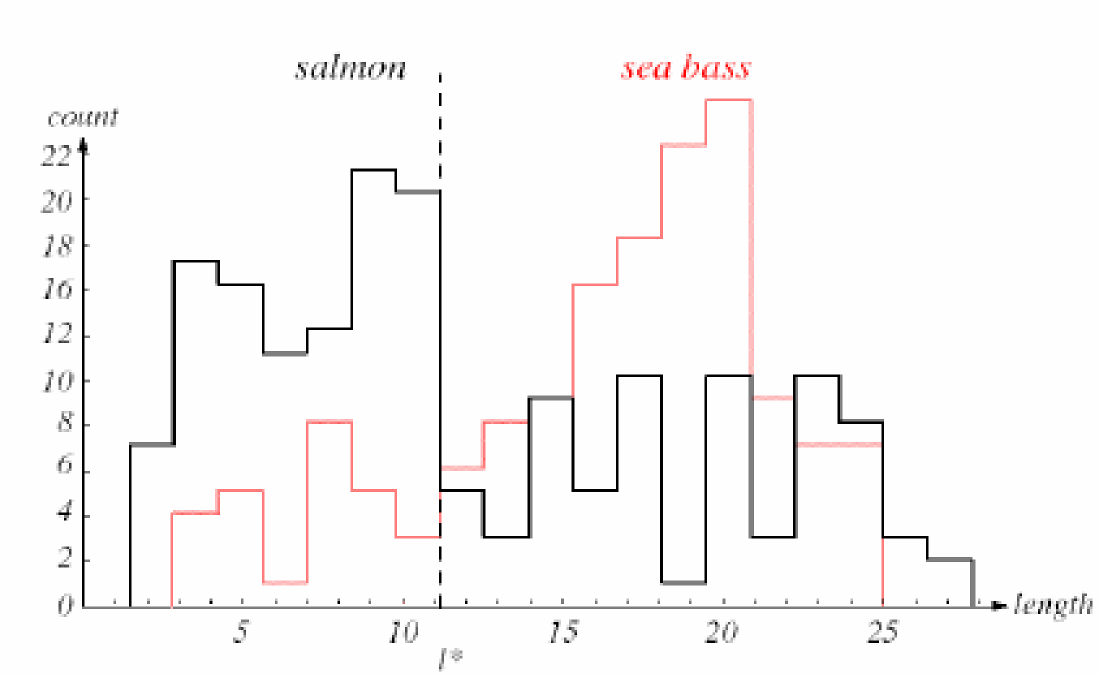
An Example (continued)

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

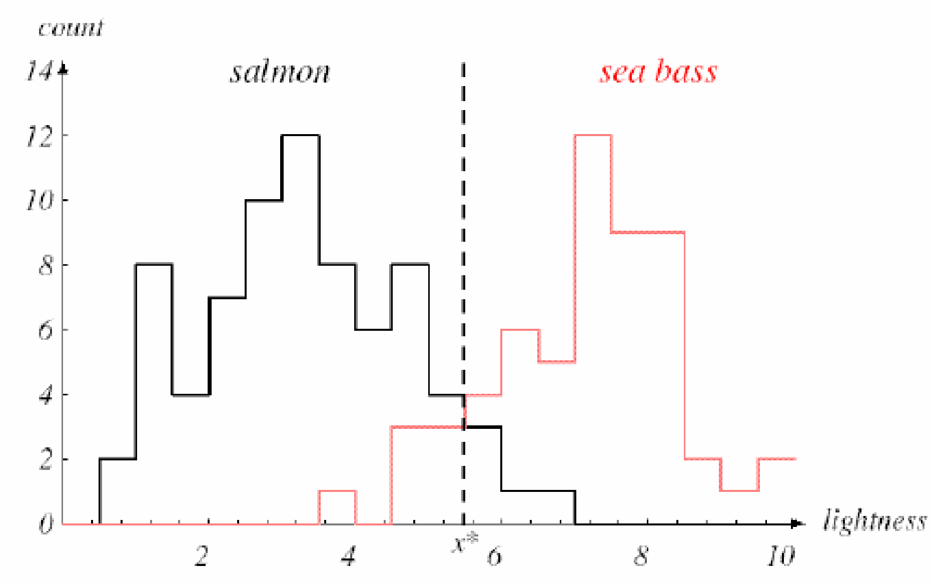
We cannot reliably separate sea bass from salmon by length alone!

An Example (continued)

New Feature:

- Average lightness of the fish scales

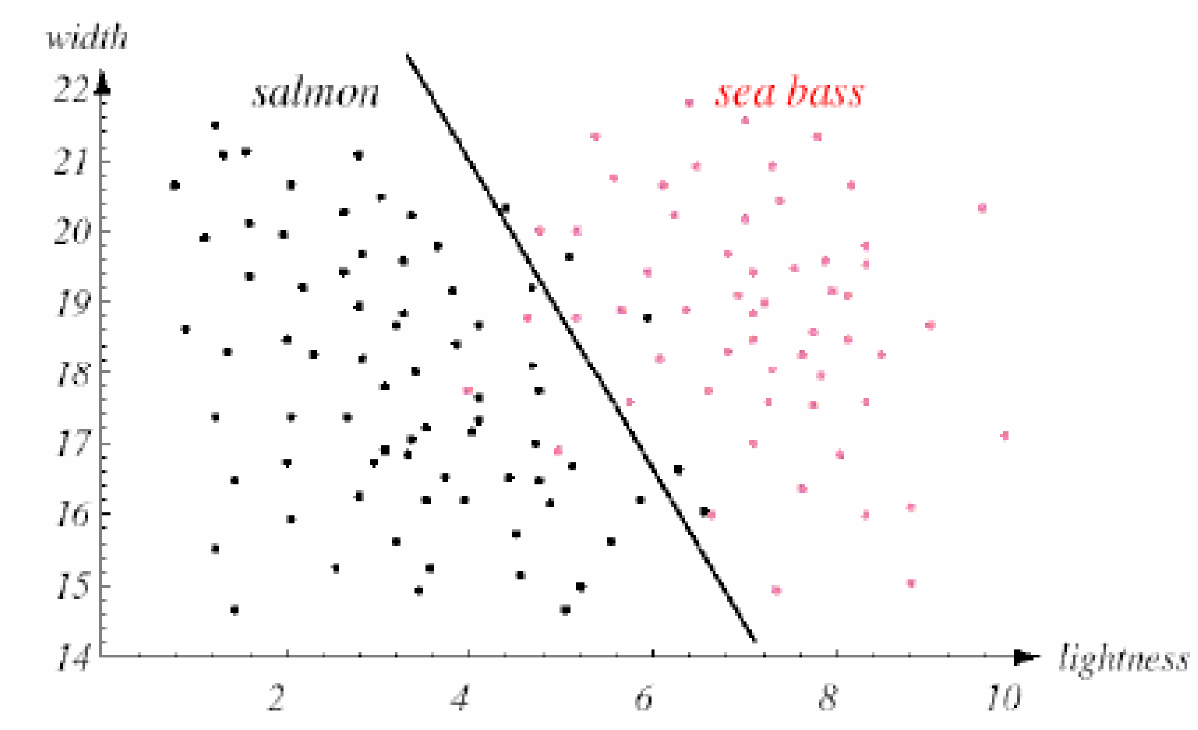
Histograms of the lightness feature for the two categories



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

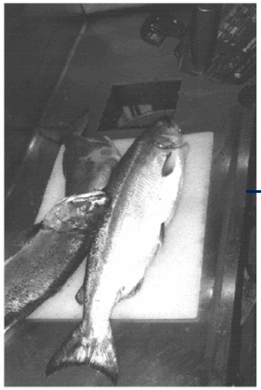
12, 4, salmon
15, 8, sea bass
8, 2, salmon
5, 10, sea bass

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

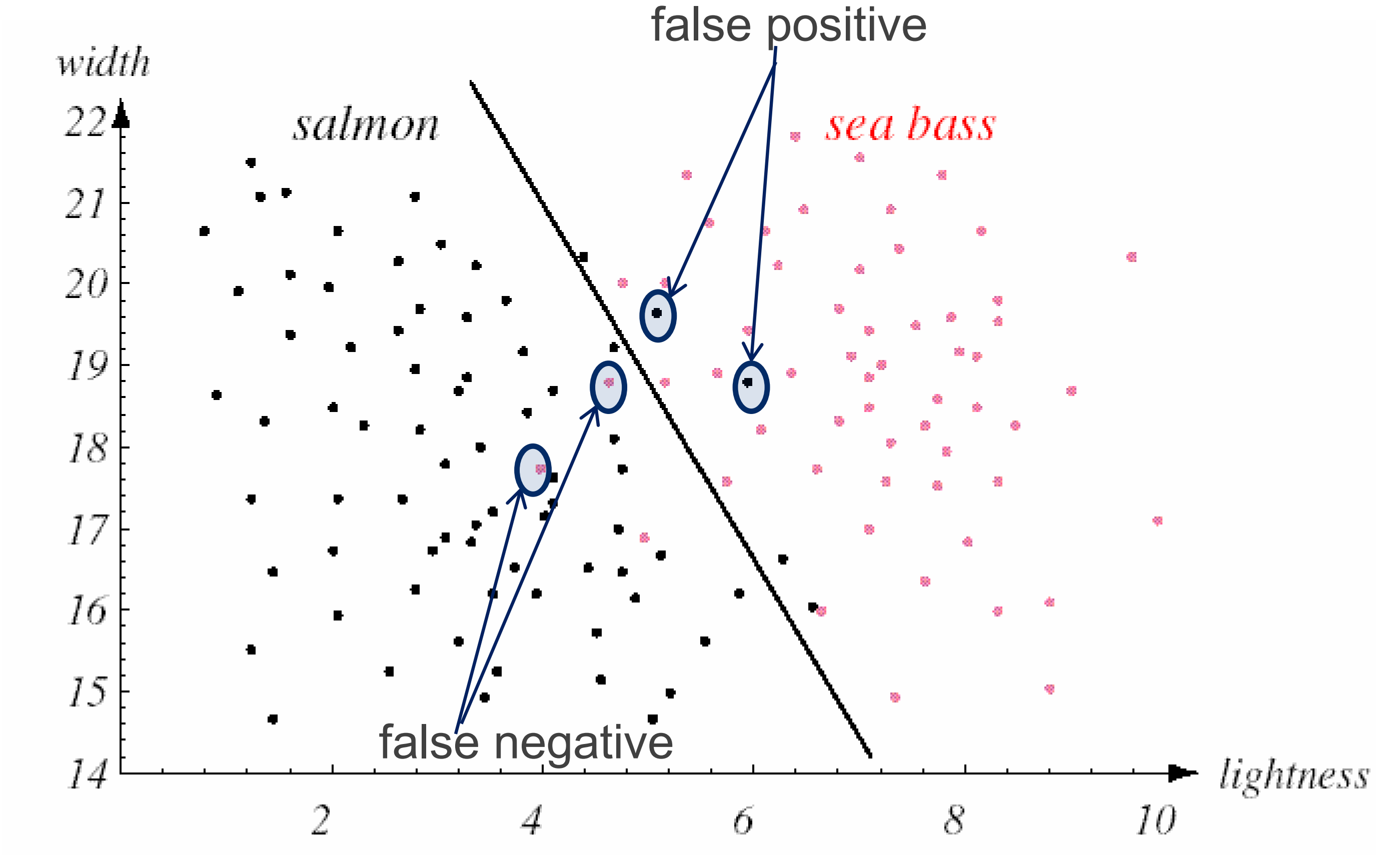
Evaluation/Prediction

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

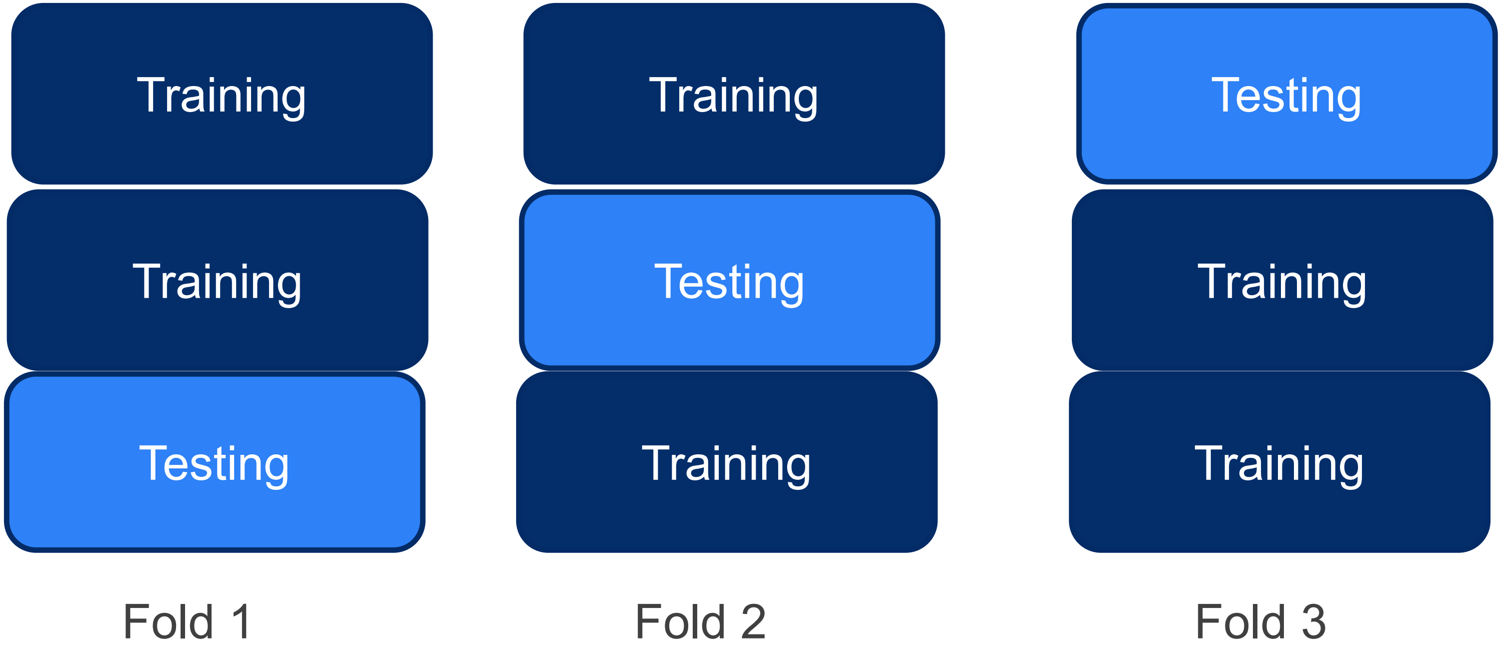


false positive

false negative

Terms

- Cross validation (3 fold)



Training Training Testing

Training Testing Training

Testing Training Training

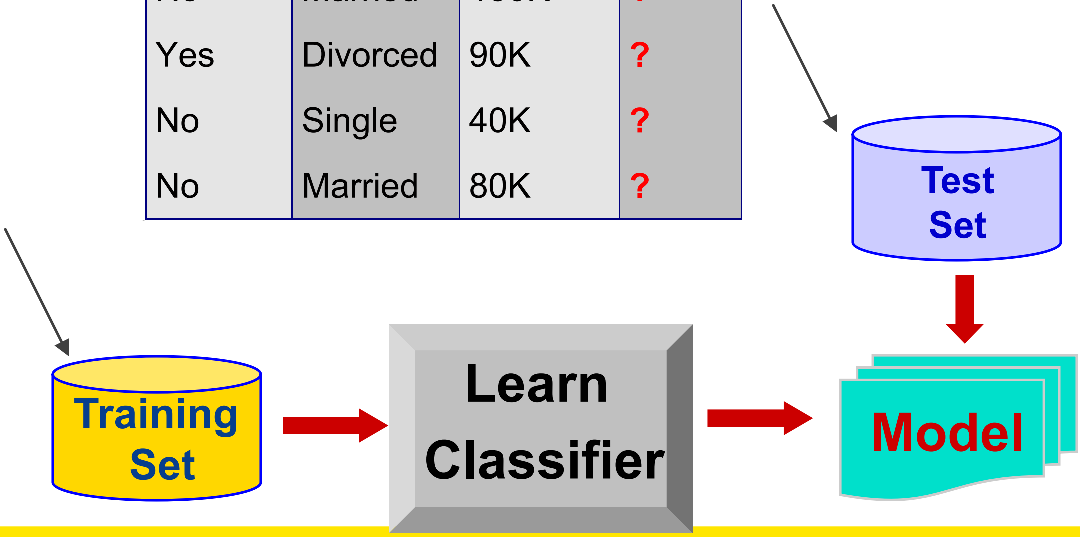
Fold 1 Fold 2 Fold 3

Classification Example 2

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

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

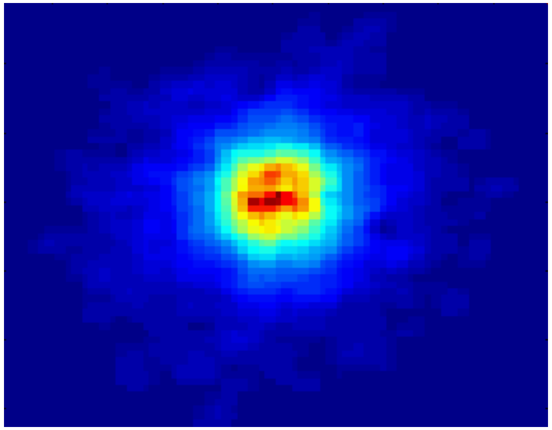
Sky Survey Cataloging

- 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!

Application 5: Classifying Galaxies

Early



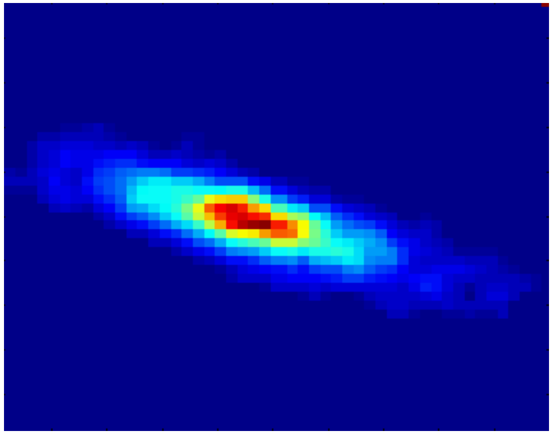
Class:

- Stages of Formation

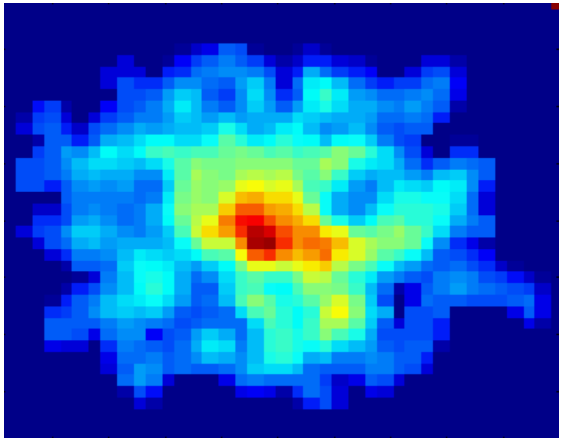
Attributes:

- 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

Overview of Data Mining

Clustering

Credit to Mehedy Masud, **UTDallas**

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:

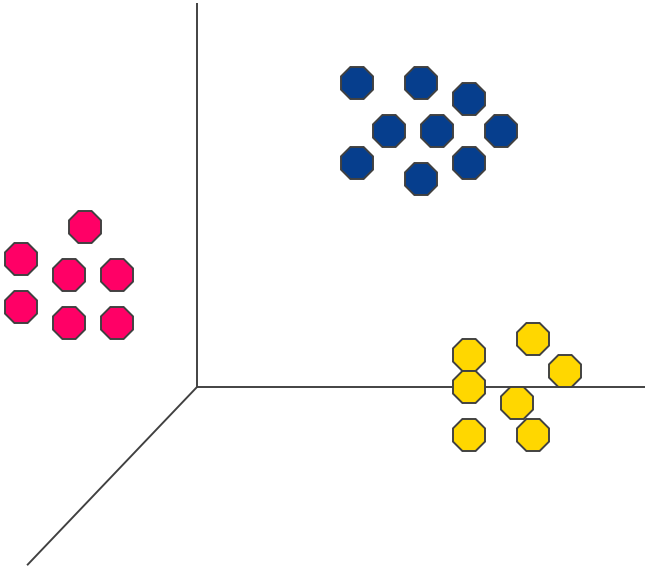
- Euclidean Distance if attributes are continuous.
- Other Problem-specific Measures.

Illustrating Clustering

Euclidean Distance Based Clustering in 3-D space.

Intra-cluster distances are minimized

Inter-cluster 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.

Overview of Data Mining

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 1

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:

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

Overview of Data Mining

Some Classification techniques

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
- c_2=15.55 and therefore significant

Regression

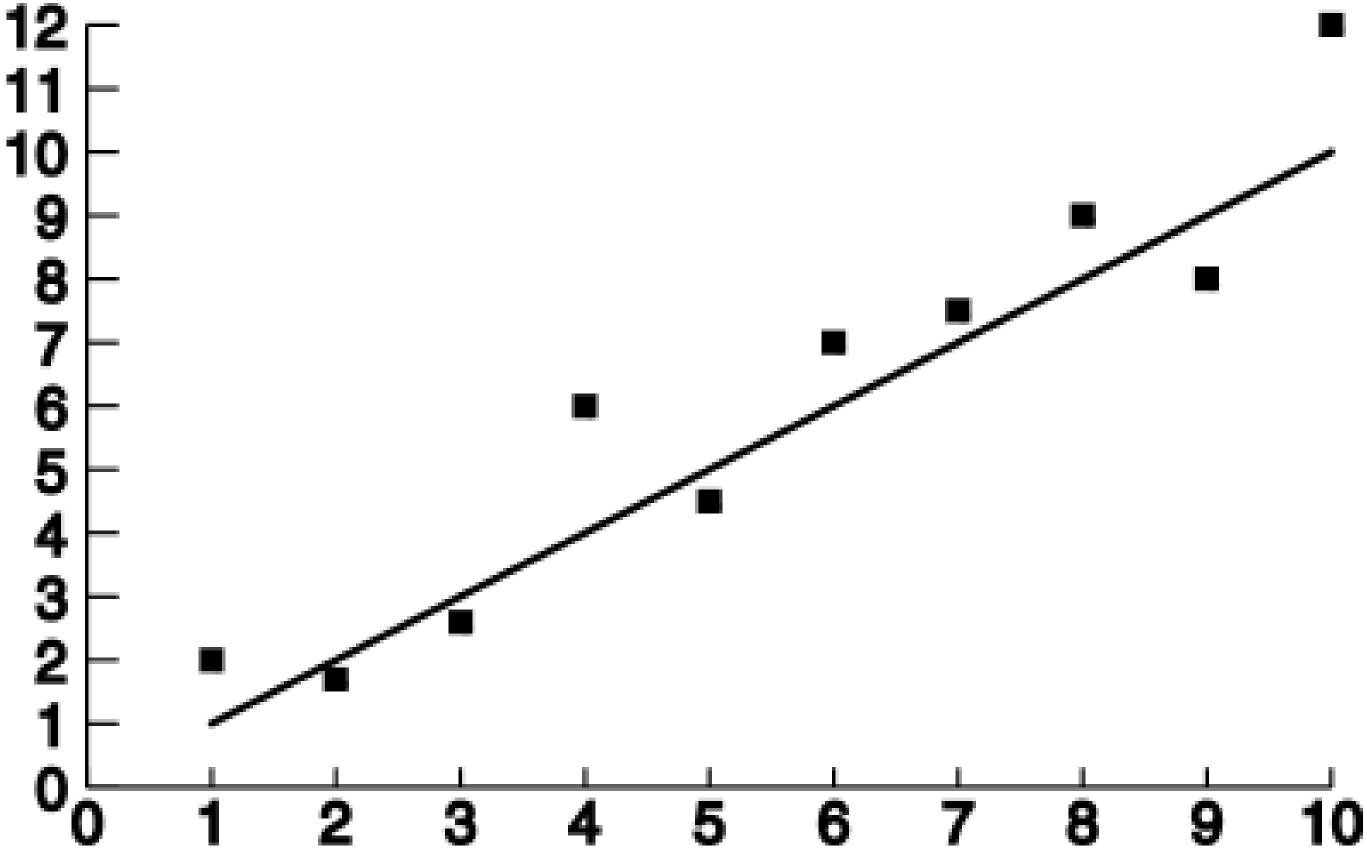
Predict future values based on past values

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:

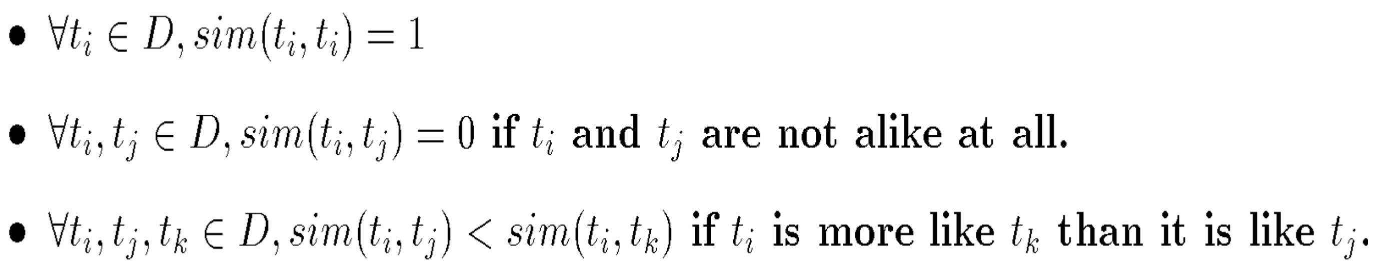
- 1 = perfect correlation
- -1 = perfect but opposite correlation
- 0 = no correlation

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

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

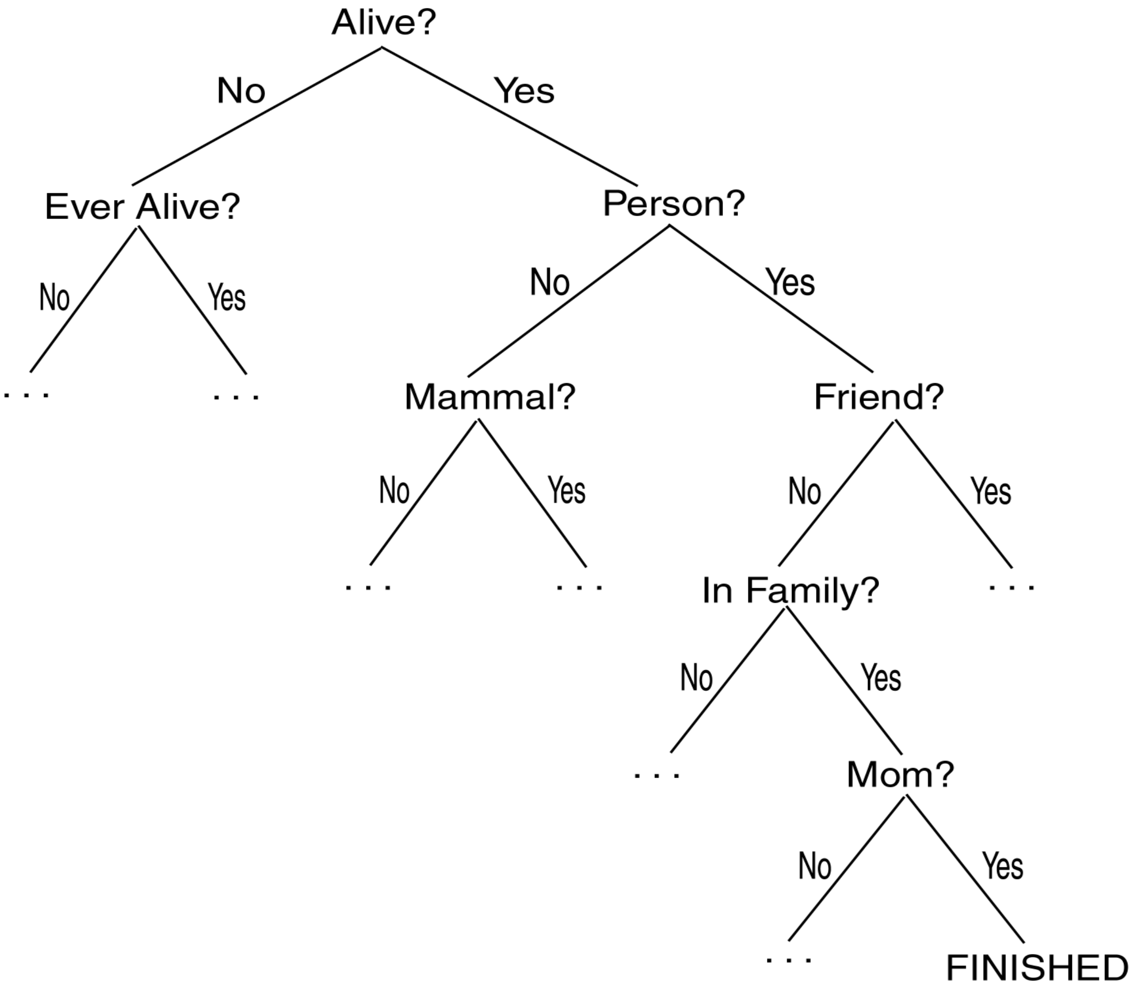
Distance Measures

Measure dissimilarity between objects

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

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

Twenty Questions Game



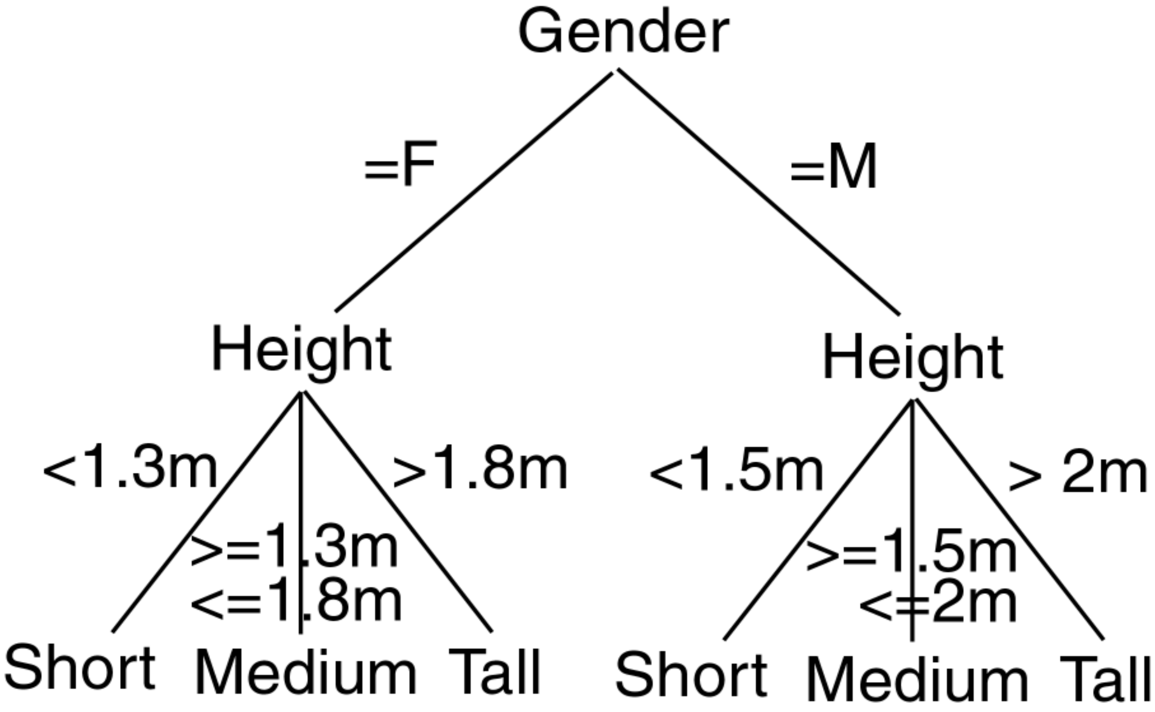
Decision Trees

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

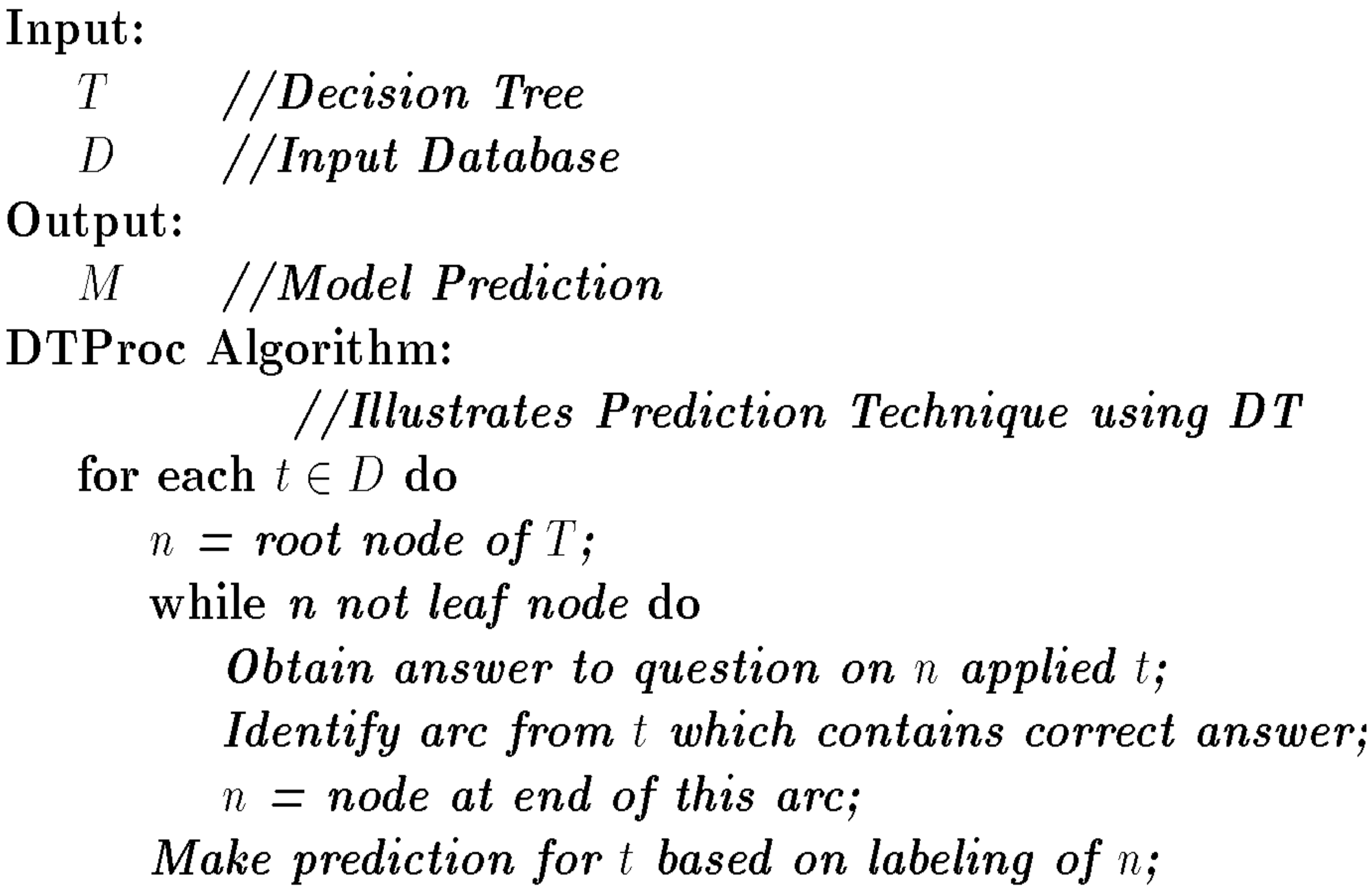
A Decision Tree Model is a computational model consisting of three parts:

- 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 $t \in D$ 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.

Neural Networks

Based on observed functioning of human brain.

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 etc..

Neural Networks

Neural Network (NN) is a directed graph

F=<V,A>

with vertices

V={1,2,…,n}

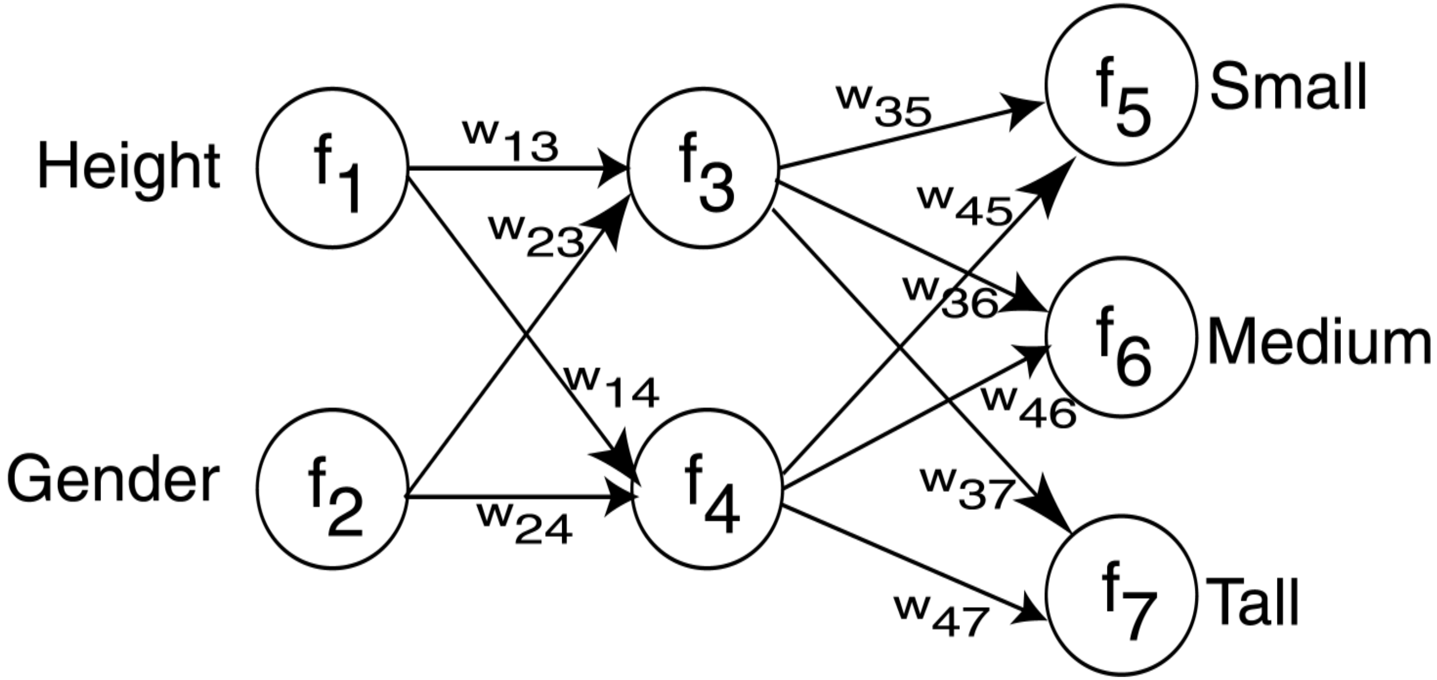
and arcs

A={<i,j>|1<=i,j<=n},

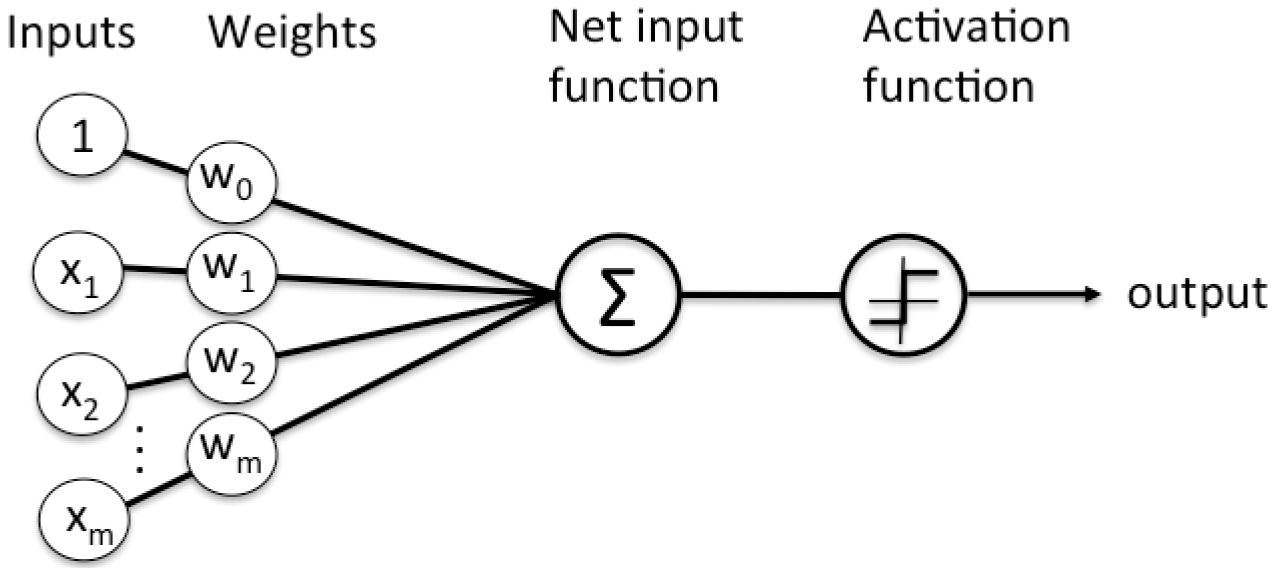
with the following restrictions:

- 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

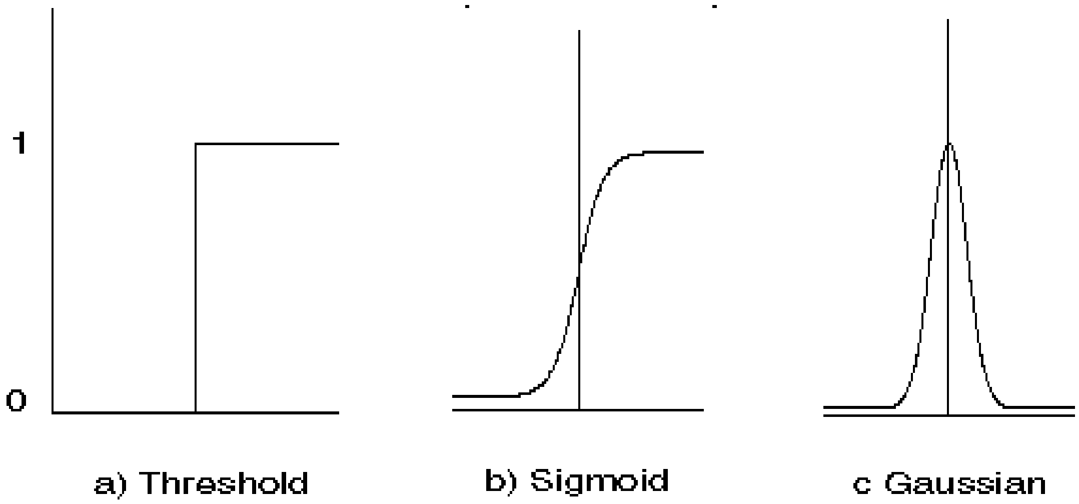


$$y_i = f_i(\sum_{j=1}^{k} w_{ji}\, x_{ji}) = f_i([w_{1i}...w_{ki}] \begin{bmatrix} x_{1i} \\ \cdots \\ x_{ki} \end{bmatrix})$$

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.