Outline

- Explain and discuss the concepts of datamining.
- Explain and discuss the concepts of training versus testing data-sets.
- discuss descriptive models in SSAS.
- Create a descriptive data-mining solution using the Microsoft decision tree algorithm.

Data Mining: Overview

- What is Data Mining?
- Recently coined term for confluence of ideas from statistics and computer science (machine learning and database methods) applied to large databases in science, engineering and business.
- In a state of flux, many definitions, lot of debate about what it is and what it is not. Terminology not standard e.g. bias, classification, prediction, feature = independent variable, target = dependent variable, case = exemplar = row.

* First International workshop on Knowledge Discovery and Data M ining was in 1995

- 1

Broad and Narrow Definitions

- Broad Definition includes traditional statistical methods, Narrow Definition emphasizes automated and heuristic methods
- Data mining, data dredging, fishing expeditions
- Knowledge Discovery in Databases (KDD)

Another definition by Nitin Patel

- “Statistics at scale and speed” Darryl Pregibon
- My extension:
- – “ . . . And simplicity”

Gartner Group

- “Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques.”

Evolution of data

Evolutionary Step	Business Question	Enabling Technologies	Product Providers	Characteristics
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective, static data delivery
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), Structured Query Language (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record level
Data Warehousing & Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses	Pilot, Comshare, Arbor, Cognos, Microstrategy	Retrospective, dynamic data delivery at multiple levels
Data Mining (Emerging Today)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases	Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry)	Prospective, proactive information delivery

Drivers

- Market: From focus on product/service to focus on customer
- IT: From focus on up-to-date balances to focus on patterns in transactions - Data Warehouses -
- OLAP
- Dramatic drop in storage costs : Huge databases
  - e.g Walmart: 20 million transactions/day, 10 terabyte database, Blockbuster: 36
  - million households
- Automatic Data Capture of Transactions
  - e.g. Bar Codes , POS devices, Mouse clicks, Location data (GPS, cell phones)
- Internet: Personalized interactions, longitudinal data

Core Disciplines

- Statistics (adapted for 21st century data sizes and speed requirements). Examples:
  - Descriptive: Visualization
  - Models (DMD): Regression, Cluster Analysis
- Machine Learning: e.g. Neural Nets
- Data Base Retrieval: e.g. Association Rules
- Parallel developments: e.g. Tree methods, k Nearest Neighbors, OLAP-EDA

Typical characteristics of mining data

- “Standard” format is spreadsheet:
  - Row=observation unit, Column=variable
- Many rows, many columns
- Many rows moderate number of columns (e.g. tel. calls)
- Many columns, moderate number of rows (e.g. genomics)
- Opportunistic (often by-product of transactions)
  - Not from designed experiments
  - Often has outliers, missing data

Process – Data mining

1. Develop understanding of application, goals
2. Create dataset for study (often from Data Warehouse)
3. Data Cleaning and Preprocessing
4. Data Reduction and projection
5. Choose Data Mining task
6. Choose Data Mining algorithms
7. Use algorithms to perform task
8. Interpret and iterate thru 1-7 if necessary

Data Mining

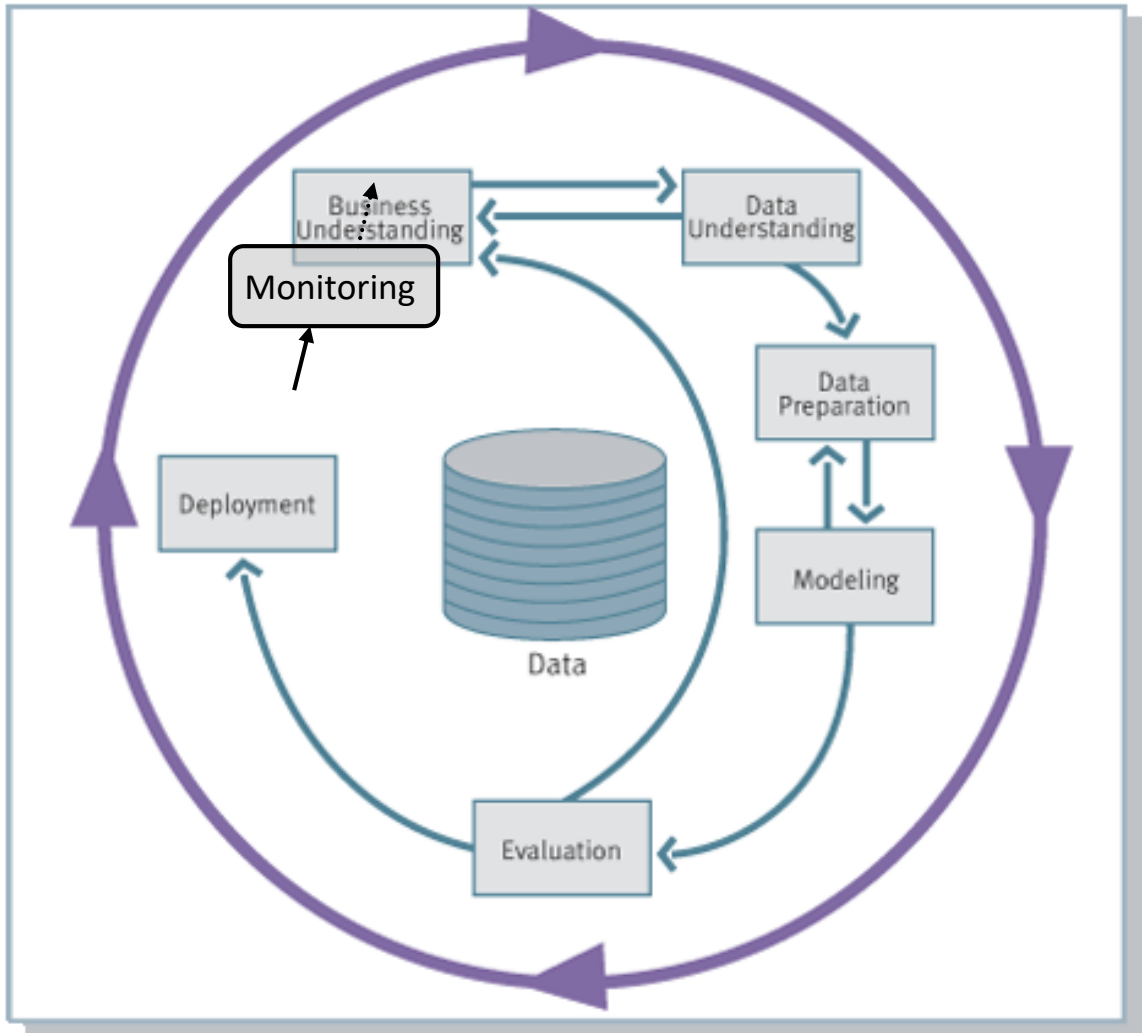
9. Deploy: integrate into operational systems.

Data mining as a process of knowledge discovery

- Data cleaning, a process that removes or transforms noise and inconsistent data
- Data integration, where multiple data sources may be combined
- Data selection, where data relevant to the analysis task are retrieved from the database
- Data transformation, where data are transformed or consolidated into forms appropriate for mining
- Data mining, an essential process where intelligent and efficient methods are applied in order to extract patterns
- Pattern evaluation, a process that identifies the truly interesting patterns representing knowledge based on some interestingness measures
- Knowledge presentation, where visualization and knowledge representation techniques are used to present the mined knowledge to the user

Knowledge Discovery Process flow, according to CRISP-DM

Cross-industry standard process for data mining, known as CRISP-DM, is an open standard process model that describes common approaches used by data mining experts. It is the most widely-used analytics model.



see
www.crisp-dm.org
for more
information

Continuous
monitoring and
improvement is
an addition to CRISP

Training

- Training Partition

- The training partition is typically the largest partition, and contains the data used to build the various models we are examining. The same training partition is generally used to develop multiple models.

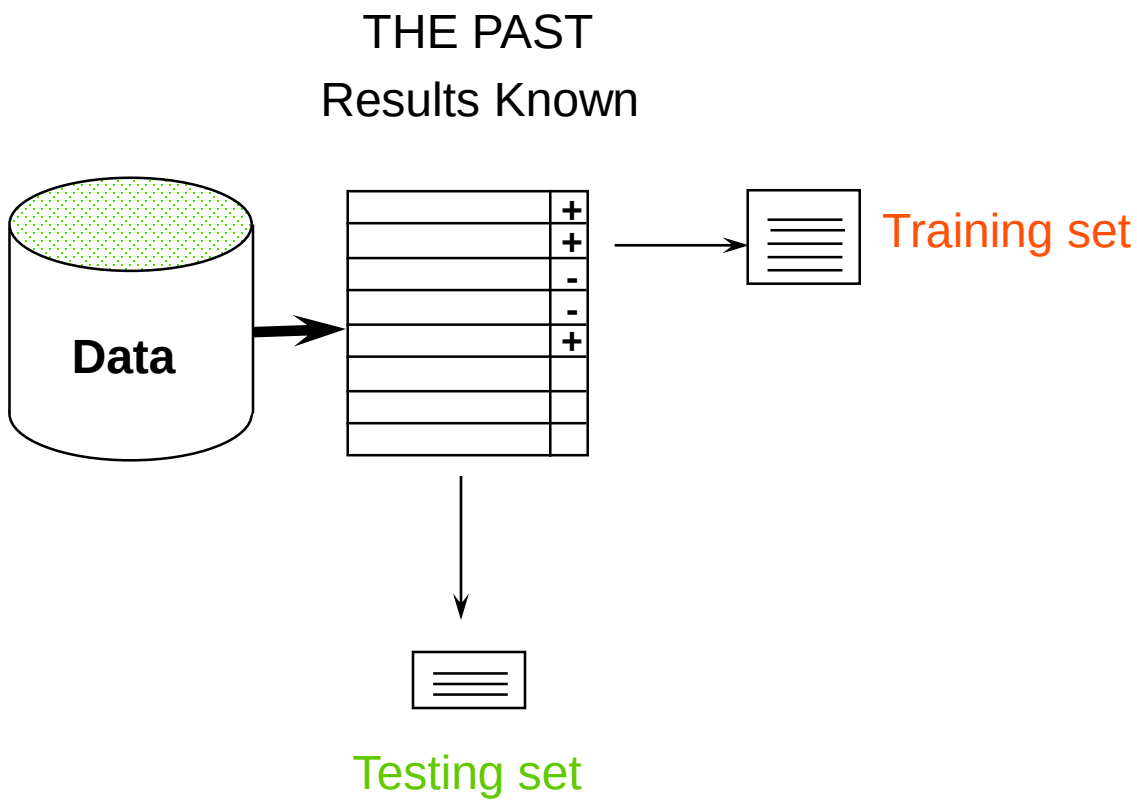
Validation

This partition (sometimes called the “test” partition) is used to assess the performance of each model,so that you can compare models and pick the best one. In some algorithms (e.g. classification and regression trees), the validation partition may be used in automated fashion to tune and improve the model.

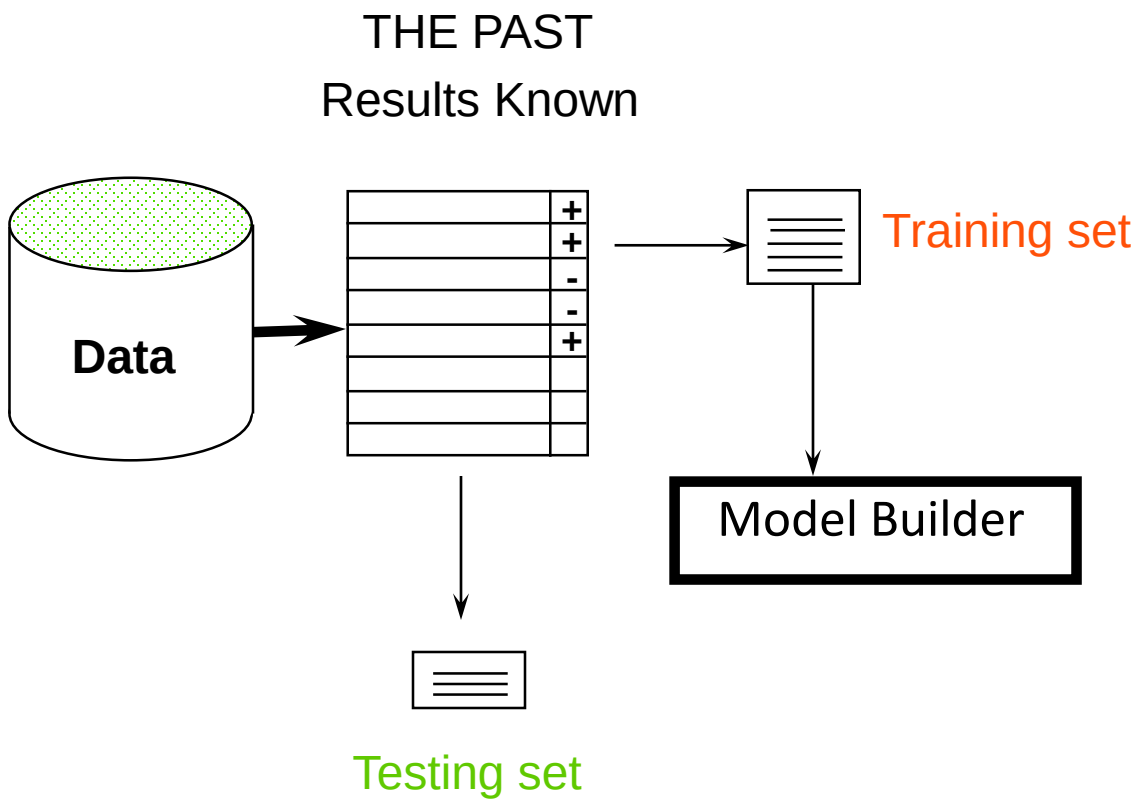
Test

- Test Partition
- This partition (sometimes called the “holdout” or “evaluation” partition) is used if we need to assess the performance of the chosen model with new data.
- Why have both a validation and a test partition? When we use the validation data to assess multiple models and then pick the model that does best with the validation data, we again encounter another (lesser) facet of the overfitting problem — chance aspects of the validation data that happen to match the chosen model better than other models.

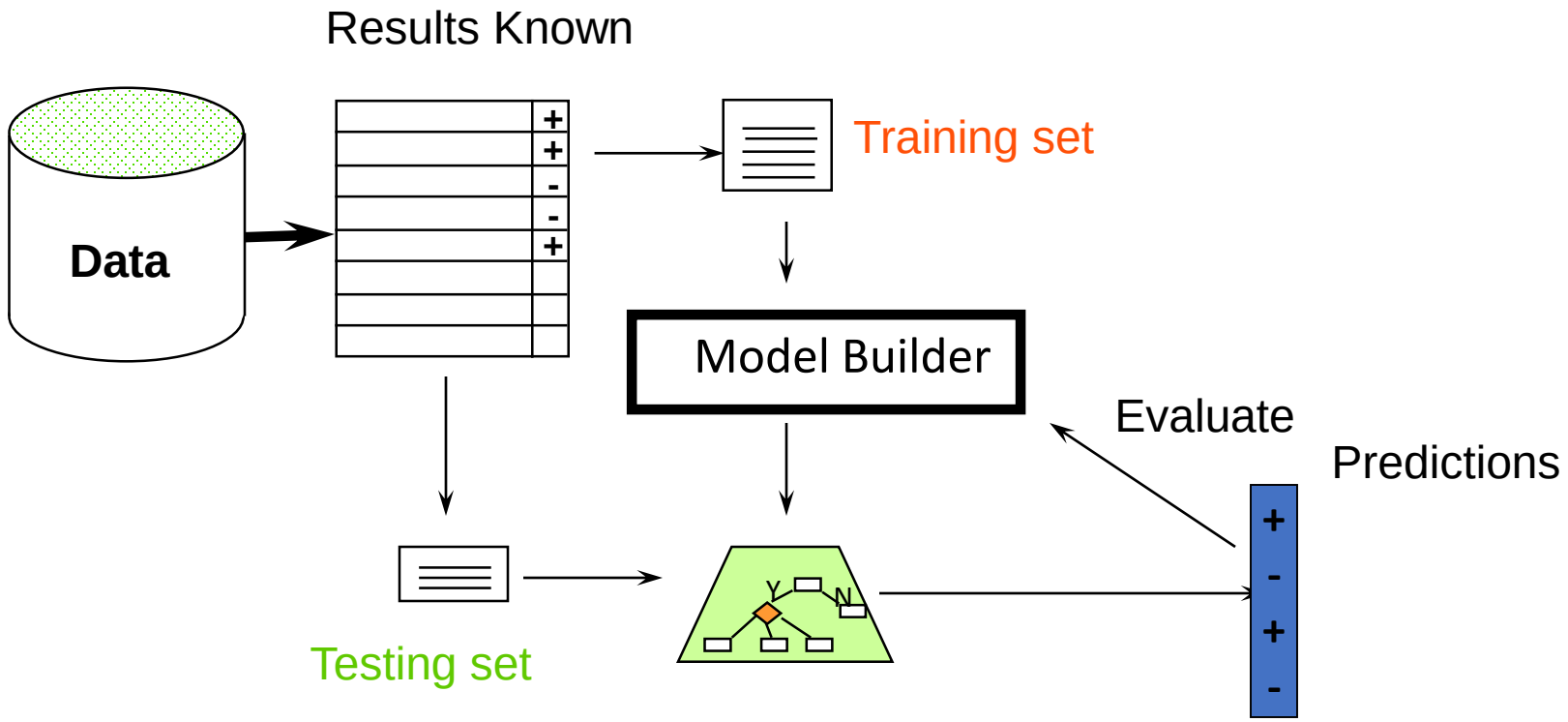
Classification Step 1:
Split data into train and test sets



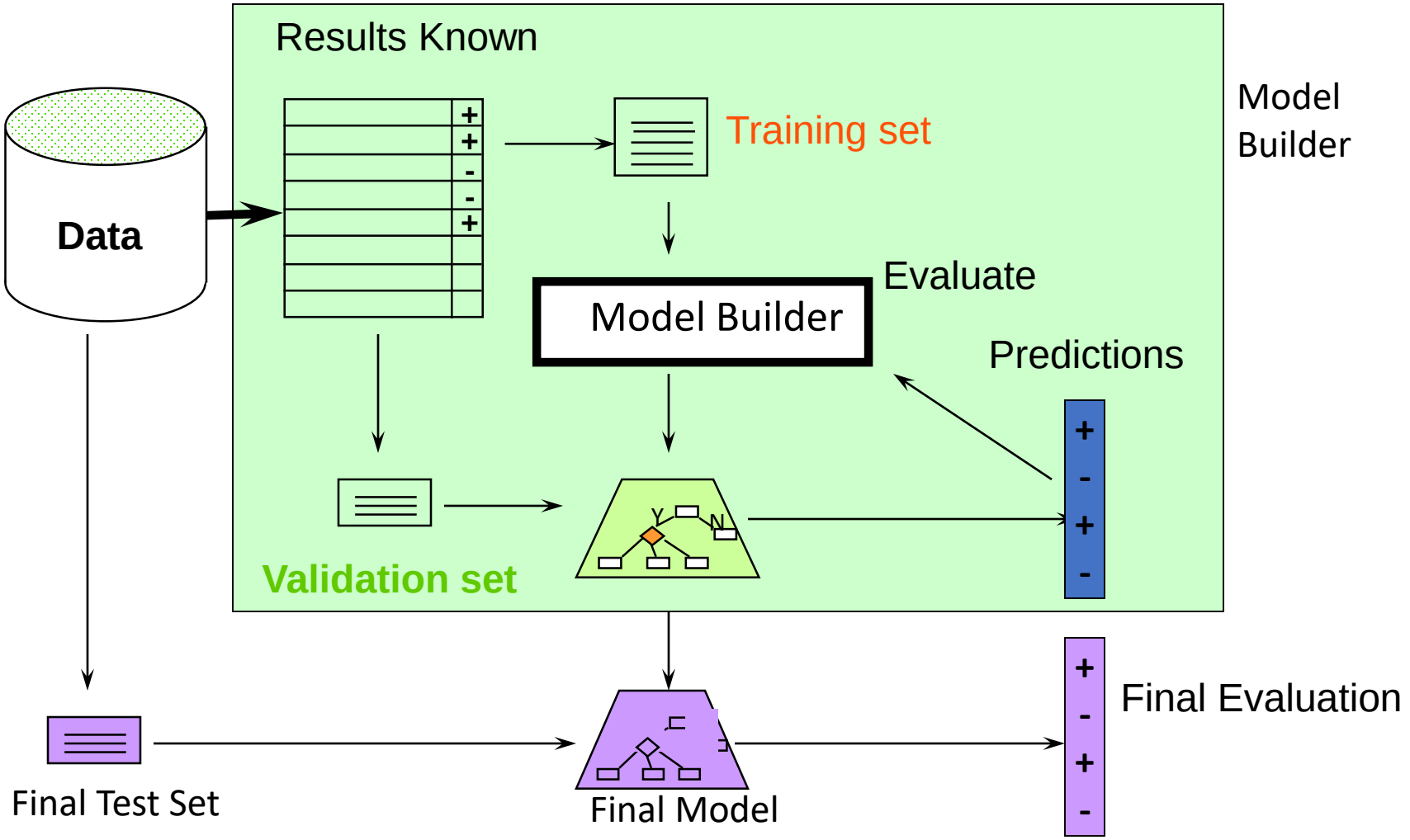
Classification Step 2:
Build a model on a training set



Classification Step 3: Evaluate on test set (Re-train?)



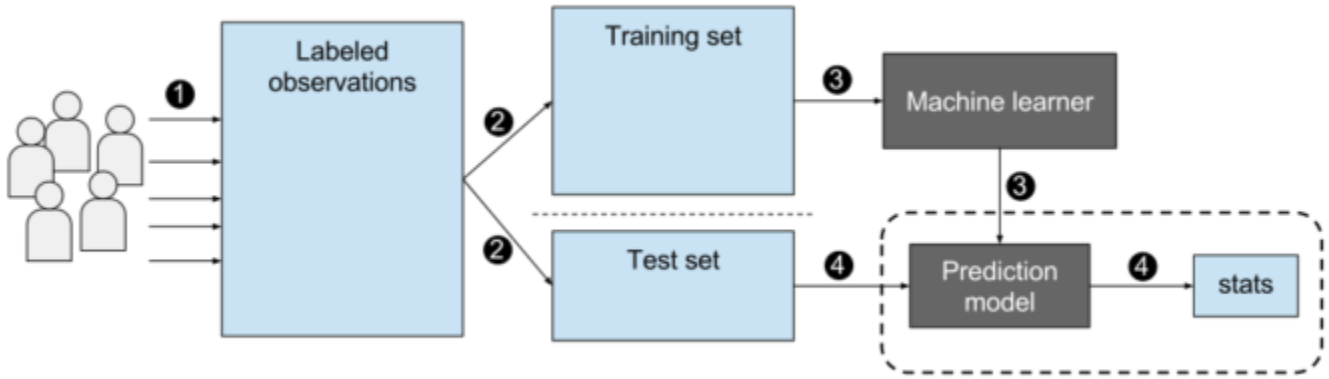
Classification:
Train, Validation, Test split



Supervised Learning

- Process of learning algorithm from the training dataset.
- You have input variables and an output variable.
- Use an algorithm to learn the mapping function from the input to the output.
- Aim is to approximate the mapping function so that when we have new input data we can predict the output variables for that data.

Supervised Learning

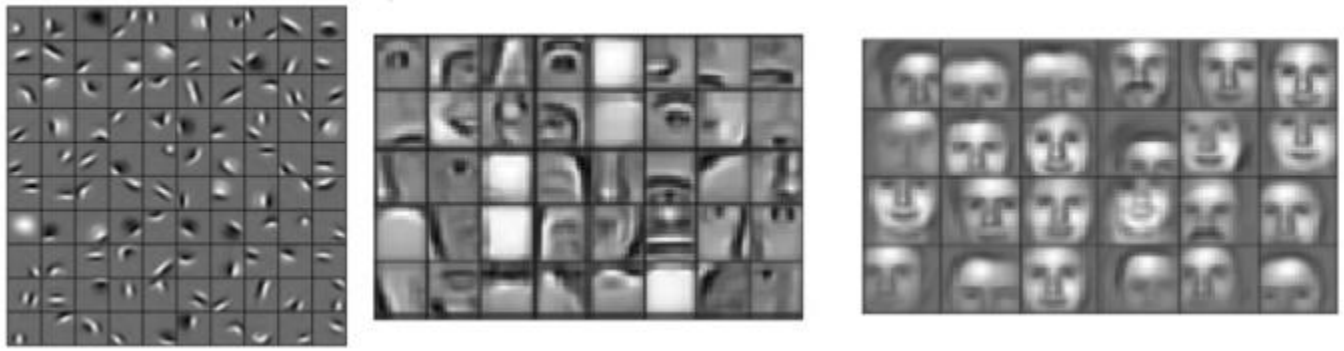


With supervised machine learning, the algorithm learns from labeled data.

Unsupervised Learning

- Modeling the underlying or hidden structure of the data
- You only have input data and no corresponding output variables.
- The training dataset is a collection of examples without a specific desired outcome or correct answer.

Unsupervised Learning



Automatically extract features and find patterns.

SEMMA Methodology (SAS)

- **S**ample from data sets, Partition into
- Training, Validation and Test datasets
- **E**xplore data set statistically and graphically
- **M**odify: Transform variables, Impute missing values
- **M**odel: fit models e.g. regression, classfication tree, neural net
- **A**ssess: Compare models using Partition, Test datasets

Core Concepts

- Types of Data:
- – Numeric
- Continuous – ratio and interval
- Discrete
- Need for Binning
- Categorical – order and unordered
- Binary
- Overfitting and Generalization
- Distance
- Curse of Dimensionality

ordinal = categorical with an order e.g. age

Binning

- **Data binning** or **bucketing** is a data pre-processing technique used to reduce the effects of minor observation errors.
- The original data values which fall in a given small interval, a bin, are replaced by a value representative of that interval, often the central value.

Categorical, Ordered

- A categorical variable (sometimes called a nominal variable) is one that has two or more categories, but there is no ordering to the categories.
- E.g. Hair color is a categorical variable having a number of categories (blonde, grey, black)
- Ordered - suppose you have a variable, economic status, with three categories (low, medium and high). Measures some “degree” within the category.

Overfitting , Generalization

Overfitting is a general problem that plagues all machine learning methods.

A learning algorithm is said to overfit if it is:

- more accurate in fitting known data (i.e. training data) (hindsight)
- less accurate in predicting new data (i.e. test data) (foresight)
- Occurs when a classifier fits the training data too tightly.
- Works well on the training data but not on independent test data.

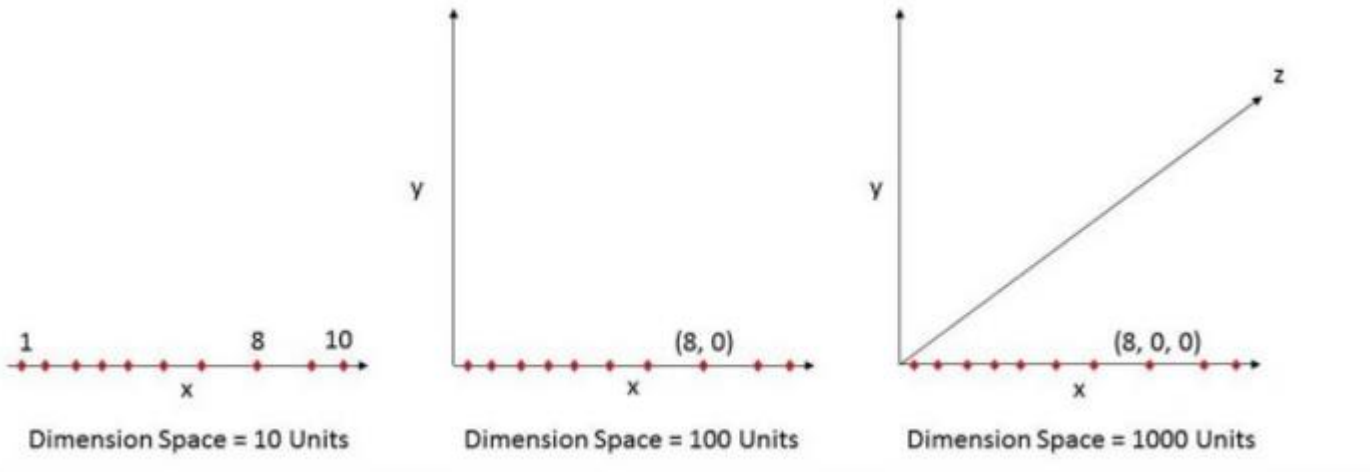
Overfitting cont’d

- Overfitting occurs when a model begins to memorize training data rather than learning to generalize from trend.
- The more difficult a criterion is to predict (i.e., the higher its uncertainty), the more noise exists in past information that need to be ignored. The problem is determining which part to ignore.
- Overfitting generally occurs when a model is excessively complex, such as having too many parameters relative to the number of observations.

Distance

- Used in classification type algorithms
- Measuring the closeness or similarity of objects with a notion of distance.

Dimensionality – more variables greater performance penalty.



Dimensional Reduction

- It reduces the time and storage space required.
- Removal of multi-collinearity (variables that are related) improves the performance of the machine learning model.
- It becomes easier to visualize the data when reduced to very low dimensions such as 2D or 3D.

Types of models

- **Descriptive**
  Provide insight into what has happened.
- **Prescriptive**
  Advice on how to optimize
- **Predictive**
  Forecasts provide insight on what could happen in the future.

Descriptive Models

- **Decision tree**:

  creates a tree based on the attribute values that play an important role in segmentation.

- **Association rule**:

  identifies the association between attributes.

  One of the most common usages of this is to perform a market-basket analysis.

- **Clustering**:

  categorizes items in groups with similar attribute values.

- **Naïve Bayes**:

  Uses the Bayesian technique for categorization of elements. It is useful for finding attributes that affect the generation of results; for example, finding the prospective buyers of a product.

Descriptive models cont’d

- **Linear regression**:

  Part decision tree that finds linear relationships between variables.

  Good option for figuring out the trend between continuous variables, for example, marketing costs and sales.
- **Neural network**:

  Works with the state of the input and the predictable variables and generates the possibility of the state’s relationships.

  Good candidate to answer text mining questions.

- **Logistic regression**:

  Version of the neural network algorithm; it calculates the effect of input variables on outputs and generates weights based on calculations. This algorithm can be used to find the weight factor of different inputs to generate the result.
- **Sequence clustering**:
- Identifies the sequence of variables. It can be used to answer the work order or the clicking path on a website.
- **Time Series**:

  Used for time-based analysis, for example, predicting sales for the next couple of months.

Decision Trees in industry

How well does it work?

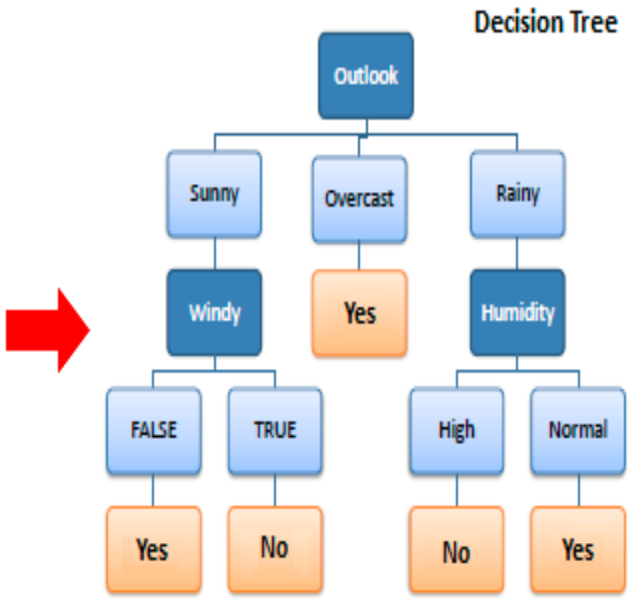
Many case studies have shown that decision trees are at least as accurate as human experts.

- A study for diagnosing breast cancer had humans correctly classifying the examples 65% of the time; the decision tree classified 72% correct
- British Petroleum designed a decision tree for gas-oil separation for offshore oil platforms that replaced an earlier rule-based expert system
- Cessna designed an airplane flight controller using 90,000 examples and 20 attributes per example

Decision Trees

trees must be balanced
challenge is getting the data from the tables into the trees and having them balanced

Predictors				Target
Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No



How does one build a decision tree

- For the root node, select the attribute with the highest entropy
- This is calculated by frequency using the following formula

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

we want randomness here...?
pbhola drew on board and said if all yes, we dont get information, we want y and n's

- e.g:

Play Golf	
Yes	No
9	5

Entropy(PlayGolf) = Entropy (5,9)
= Entropy (0.36, 0.64)
= - (0.36 $\log_2$ 0.36) - (0.64 $\log_2$ 0.64)
= 0.94

information gain is moving from more random to less random!

- The #'s coming from the frequency counts

Entropy using the frequency table of two attributes:

$$E(T,X) = \sum_{c \in X} P(c)E(c)$$

		Play Golf		
		Yes	No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5
				14

E(PlayGolf, Outlook) = **P**(Sunny)*__E__(3,2) + **P**(Overcast)*__E__(4,0) + **P**(Rainy)*__E__(2,3)

= (5/14)*0.971 + (4/14)*0.0 + (5/14)*0.971

= 0.693

Find the next attribute based on the information gain computed against the target attribute

Step 1: Calculate entropy of the target.

Entropy(PlayGolf) = Entropy (5,9)

= Entropy (0.36, 0.64)

= - (0.36 $\log_2$ 0.36) - (0.64 $\log_2$ 0.64)

= 0.94

For each candidate attribute, calculate the information gain as below

		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3
Gain = 0.247			

		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1
Gain = 0.029			

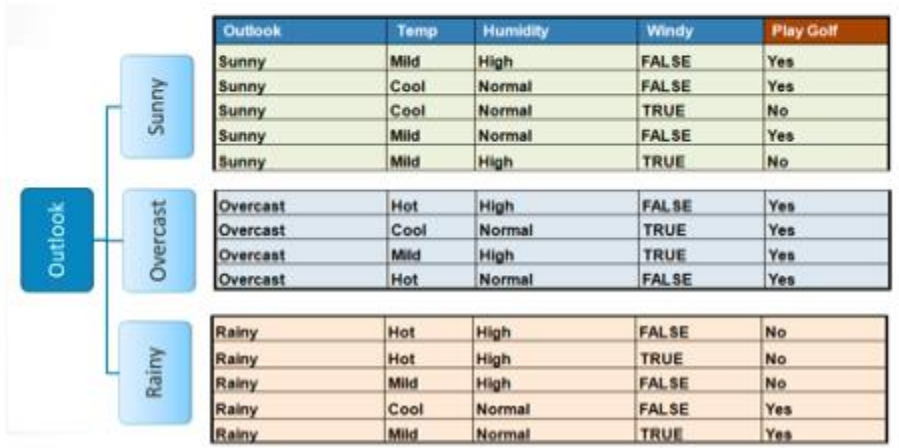
		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1
Gain = 0.152			

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3
Gain = 0.048			

$$Gain(T,X) = Entropy(T) - Entropy(T,X)$$

Select the attribute (outlook) with the largest information gain

★		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3
Gain = 0.247			



Outlook	Temp	Humidity	Windy	Play Golf
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Sunny	Mild	Normal	FALSE	Yes
Sunny	Mild	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Overcast	Cool	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes

Evaluating classifiers – confusion matrix

- Calculating the accuracy of the classifier

		Predicted Class	
		C_0	C_1
Actual Class	C_0	$n_{0,0}$ = Number of correctly classified C_0 cases	$n_{0,1}$ = Number of C_0 cases incorrectly classified as C_1
	C_1	$n_{1,0}$ = Number of C_1 cases incorrectly classified as C_0	$n_{1,1}$ = Number of correctly classified C_1 cases

Table 4.1: Classification Matrix: Meaning of Each Cell

Confusion matrix

- Summarizes the correct and incorrect classification
- Rows and Columns correspond to the true and predicted classes
- Upper right and lower left diagonal cells give the correct classification
- The other cells give the misclassification
- Total observations $n = n_{0,0} + n_{0,1} + n_{1,0} + n_{1,1}$
- Accuracy $= ((n_{0,0} + n_{1,1})/n) \times 100$
- Error $= ((n_{0,1} + n_{1,0}) /n)$

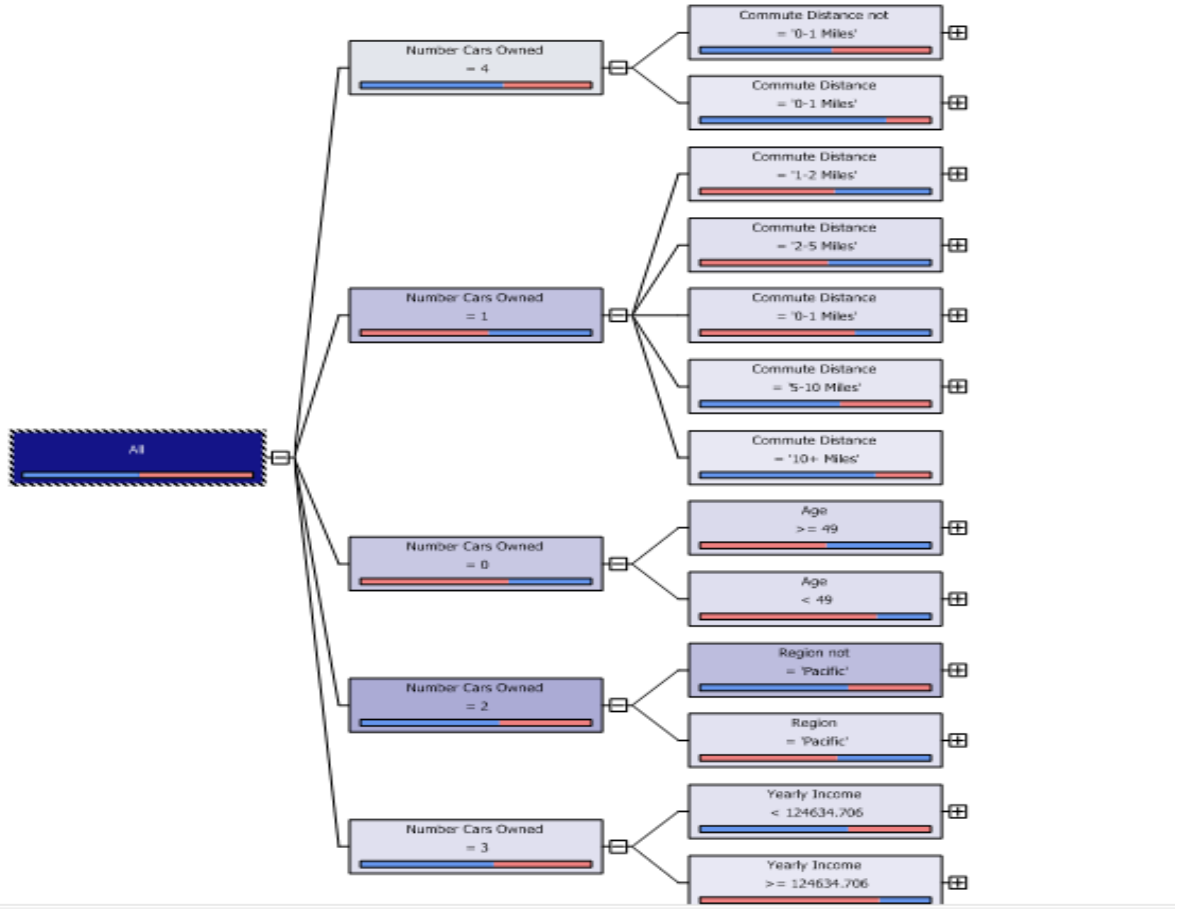
An example

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	201	85
0	25	2689

Figure 4.2: Classification Matrix Based on 3000 Observations and Two Classes

- Err = (25+85)/3000 = 3.67%
- Accuracy = (201 + 2689) / 3000 = 96.3%

Decision Tree SSAS



Dependency Network – Contribution of variables to decision

Strength of contribution depends on the distance from the target node.