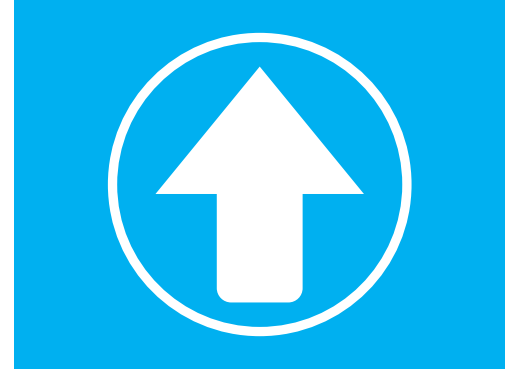


Hack Fest 2016 Big Data and Advanced Analytics Workshop

Oil & Gas companies

October, 2016



Agenda

Data Science Flow - Recap

Machine Learning – Introduction and Type

Regression

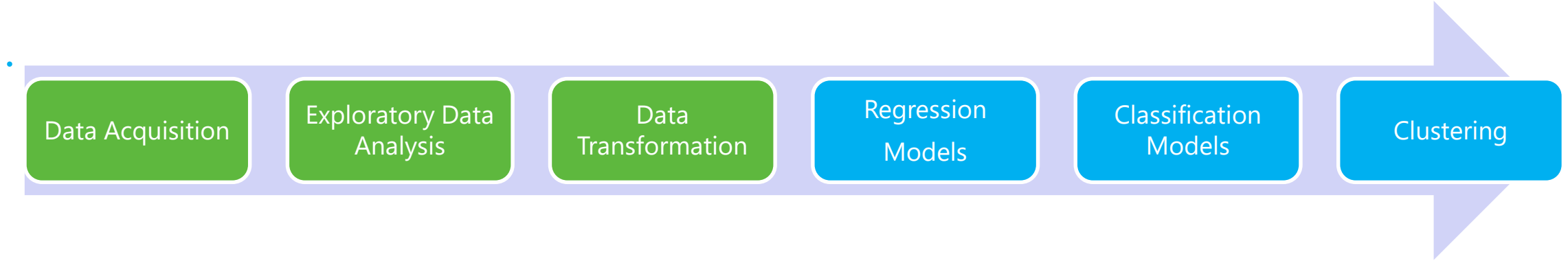
Classification

Clustering

Summary

Lab preview

Data Science Flow



Data Acquisition, EDA & Transformation

- Acquiring data from multiple sources
- Getting a feel for data
- Transforming data in context for prediction

Machine Learning

- Supervised vs Unsupervised
- Machine Learning Algorithms
- Diagnostics and Performance

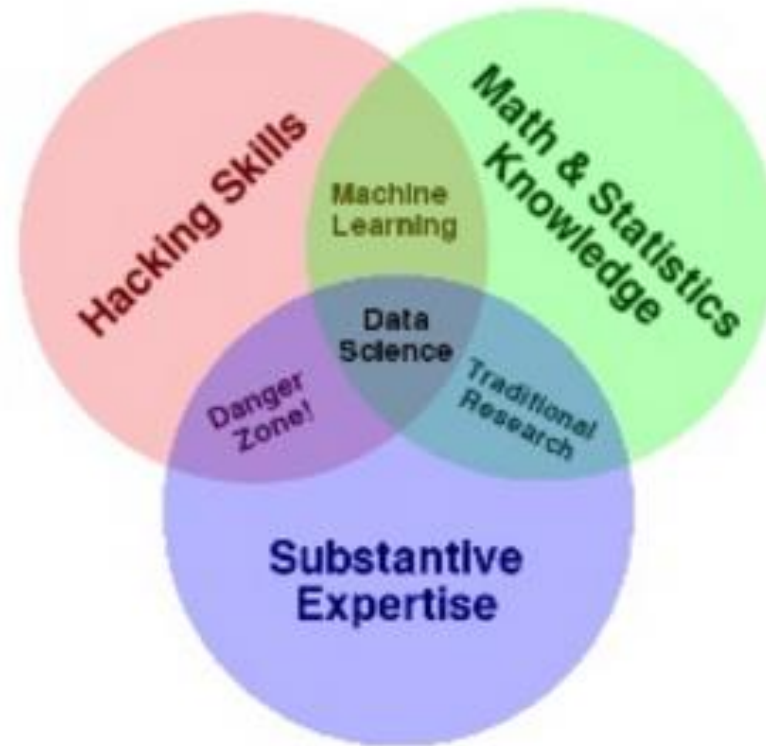
Machine Learning

"Field of study that gives computers the ability to learn without being explicitly programmed."

(Arthur Samuel, 1959)



By Philip Taylor [CC BY 2.0]



<http://drewcrumway.com/zia/2013/3/26/the-data-science-venn-diagram>

Examples of Machine Learning

Fraud Detection



SPAM filtering



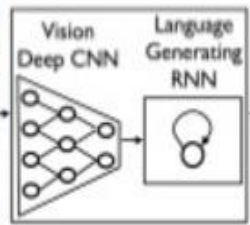
<https://doi.org/10.1515/SELVANGG> (CC BY 2.0)

Recommendation Systems



http://commons.wikimedia.org/wiki/File:Naflix_logo.svg (public)

Photo search



A group of people shopping at an outdoor market.

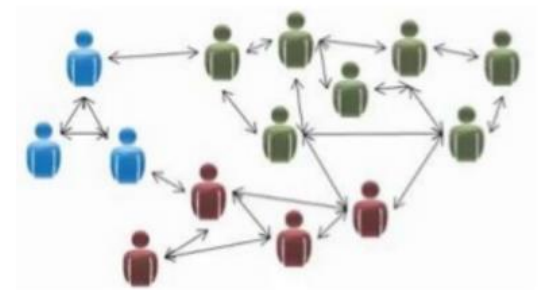
There are many vegetables at the fruit stand.

<http://googlesearch.blogspot.com/2014/11/a-picture-is-worth-thousand-coherent.html>

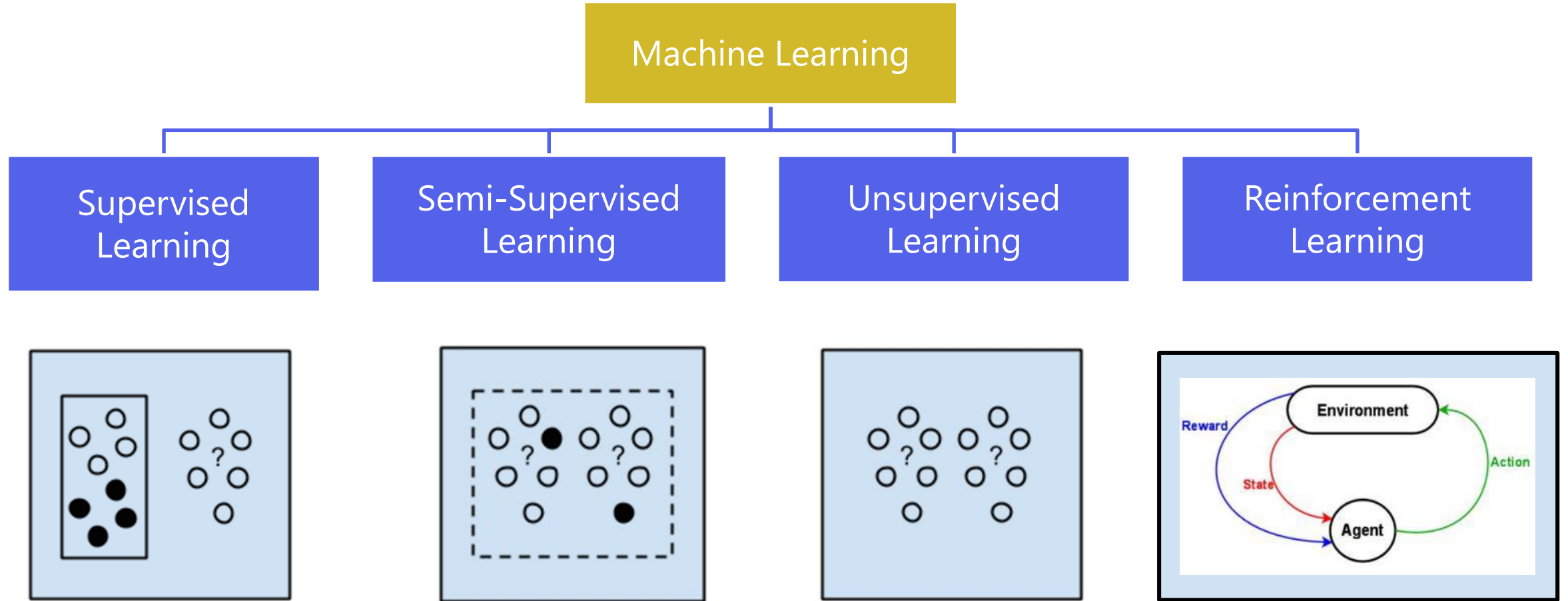
Predictive maintenance



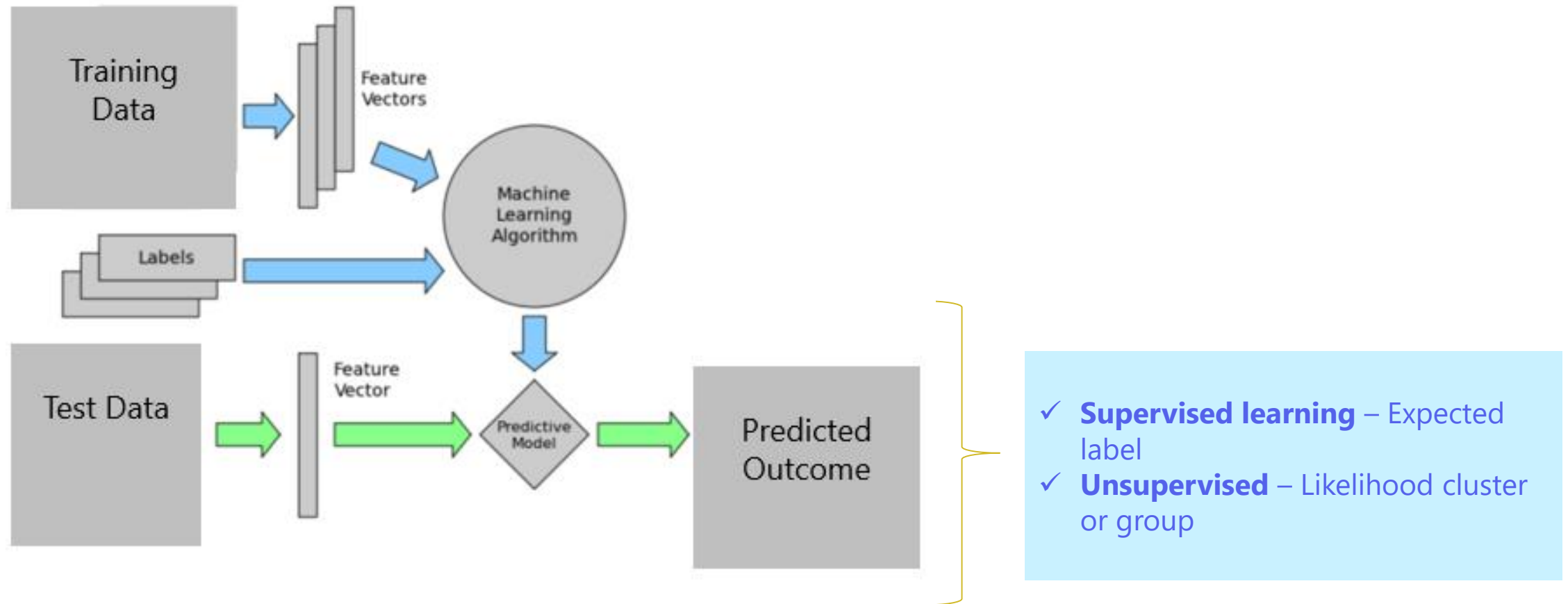
Social Network Analysis



Types of Machine Learning

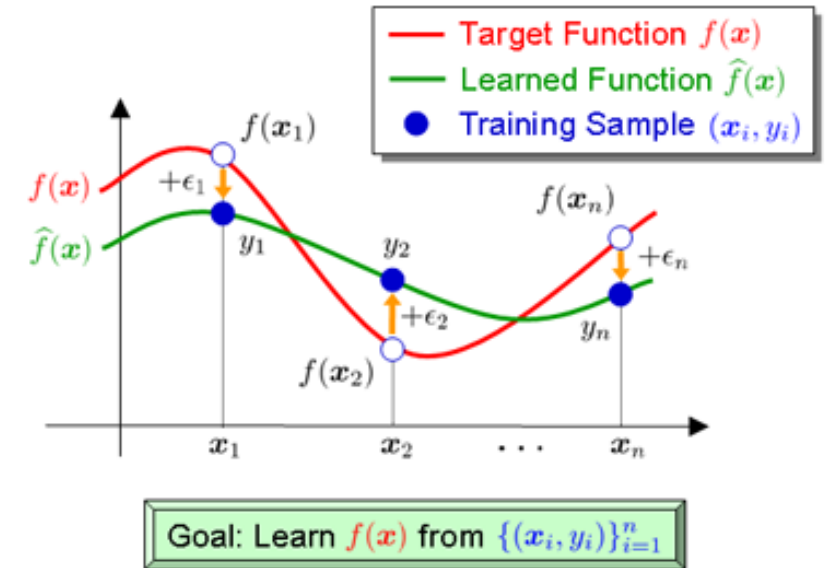


Machine Learning Workflow



Supervised Learning ~ Predictive Methods

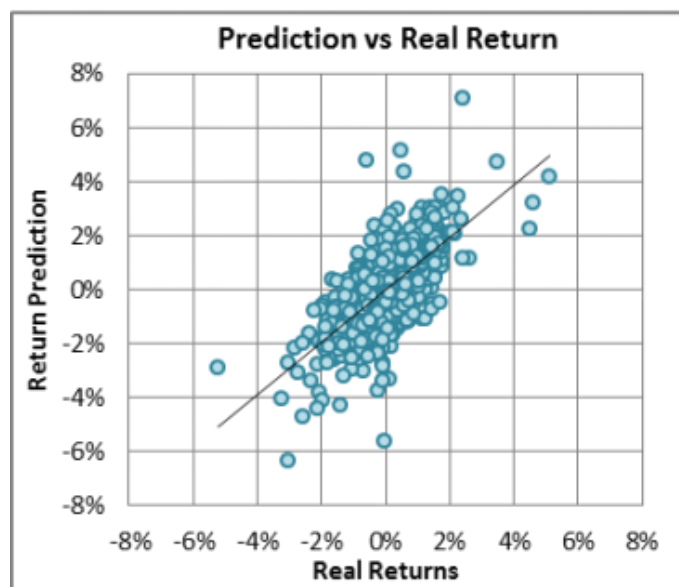
- **Supervised learning** is the task of inferring a target function (training a model) that maps a set of data to a set of target values.,^[1]
- Predictions and targets can be categorical (classification) or numeric (regression)
- Supervised Learning models account for nearly 70% of models
- Common algorithms - Linear Regression, Logistic Regression, Decision Trees, Neural Networks



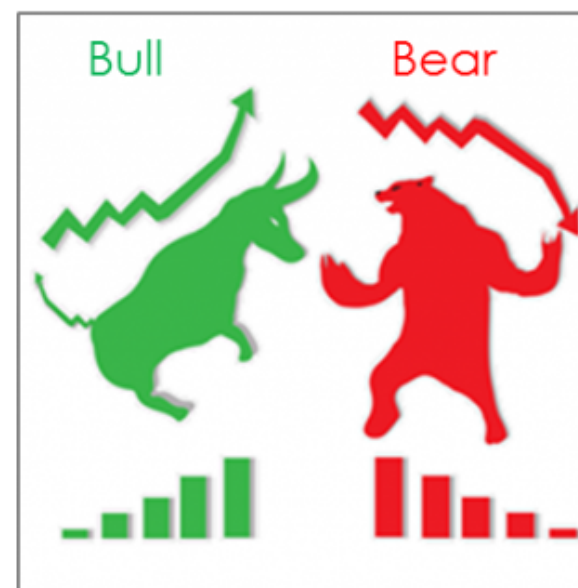
Supervised Learning - Classifiers and Regressors

What question are you trying to answer?

Regressors predict values



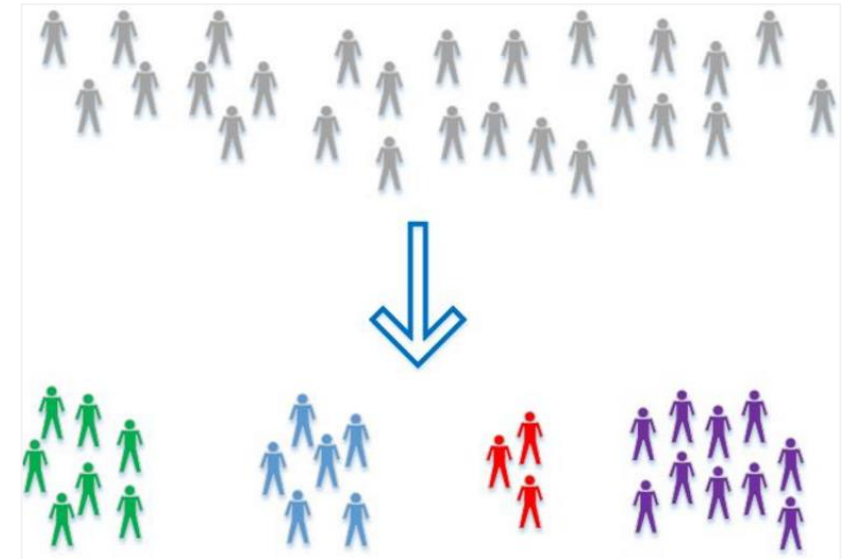
Classifiers predict classes



- ✓ The same type of model can be used to do either task

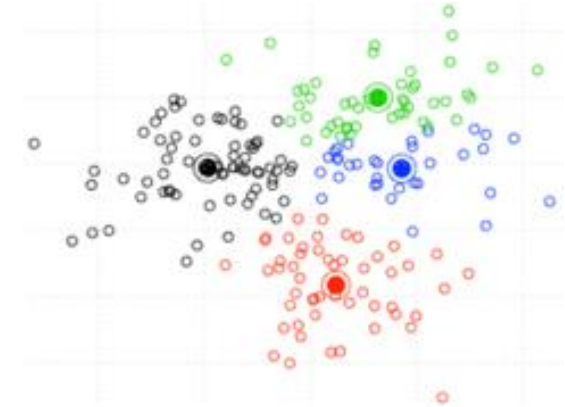
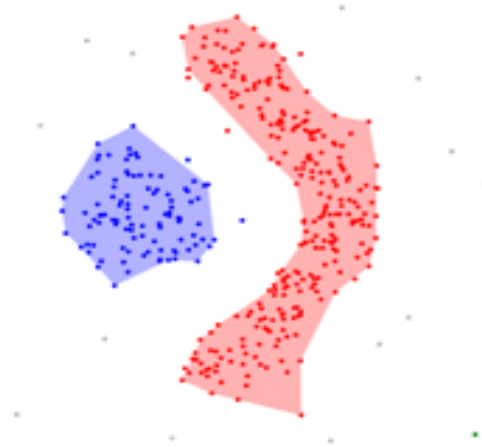
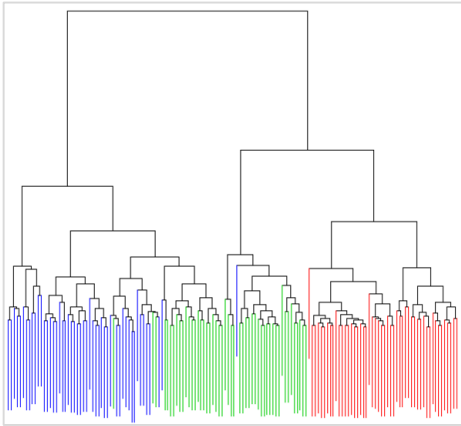
Unsupervised Learning ~ Descriptive Methods

- **Unsupervised Learning** provides insights into the data by discovering hidden structure from unlabeled data
- Unsupervised learning is kind of doing EDA as we are exploring data to uncover patterns or associations
- Evaluation of the models is indirect as there is no pre-defined label
- Common Algorithms – k-Means Clustering, Anomaly detection, Dimension Reduction, Hierarchical Clustering



Clustering Analysis

Cluster analysis or clustering is the task of **grouping** a set of objects in such a way that objects in the **same** group (called a cluster) are more **similar** (in some sense or another) to each other than to those in **other** groups (clusters)



Machine Learning- Pros and Cons

Higher Accuracy with
compute power

Data driven with multiple
data sources

Uncovers relationships from
disordered datasets

Automated process that can
be repeatable

Very efficient for Big Data
scenarios

High reliance on Labelled
Data

Needs fine-tuning with
business sense

Limitations in performance
based on assumptions

Parametric vs Non Parametric Machine Learning Models

Parametric Models

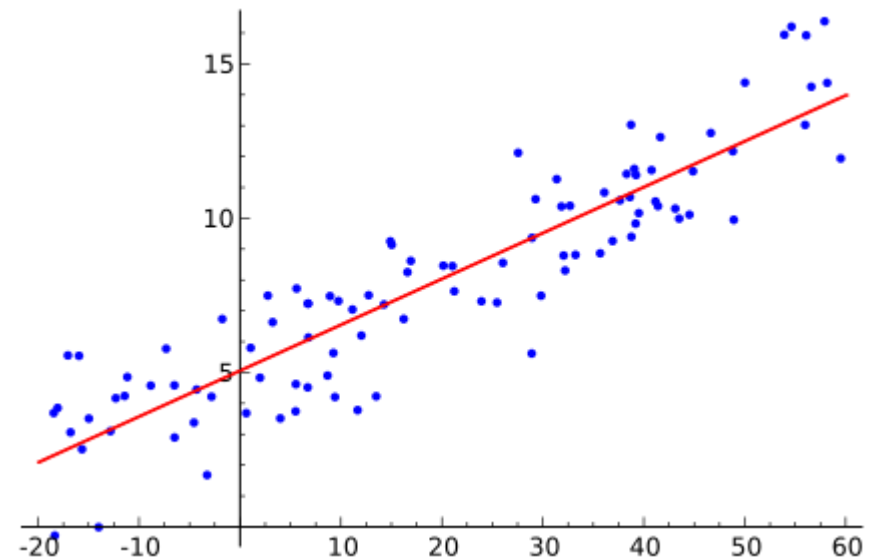
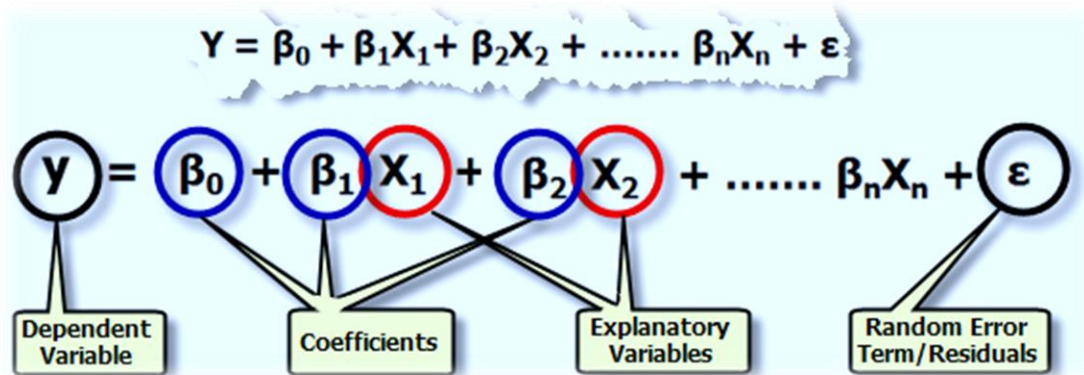
- Parametric algorithm has a fixed number of parameters.
- The algorithm assumes underlying linear function and will hence yield bad results if the assumptions is wrong
- Advantage is models are simpler to understand, computationally faster and requires less data
- Disadvantage is models are constrained to the function chosen and may lead to poor fit.
- Common example of a parametric algorithm is linear regression, logistic regression.

Non Parametric Models

- Non-parametric algorithm uses a flexible number of parameters, and the number of parameters often grows as it learns from more data.
- Advantage of non parametric models is flexibility to different forms, can handle complex problems, makes less assumptions of data
- Disadvantage is needs more training data, computationally slower, difficult to explain and higher chances of overfitting
- Common example -K-nearest neighbor , Decision Tree, SVM

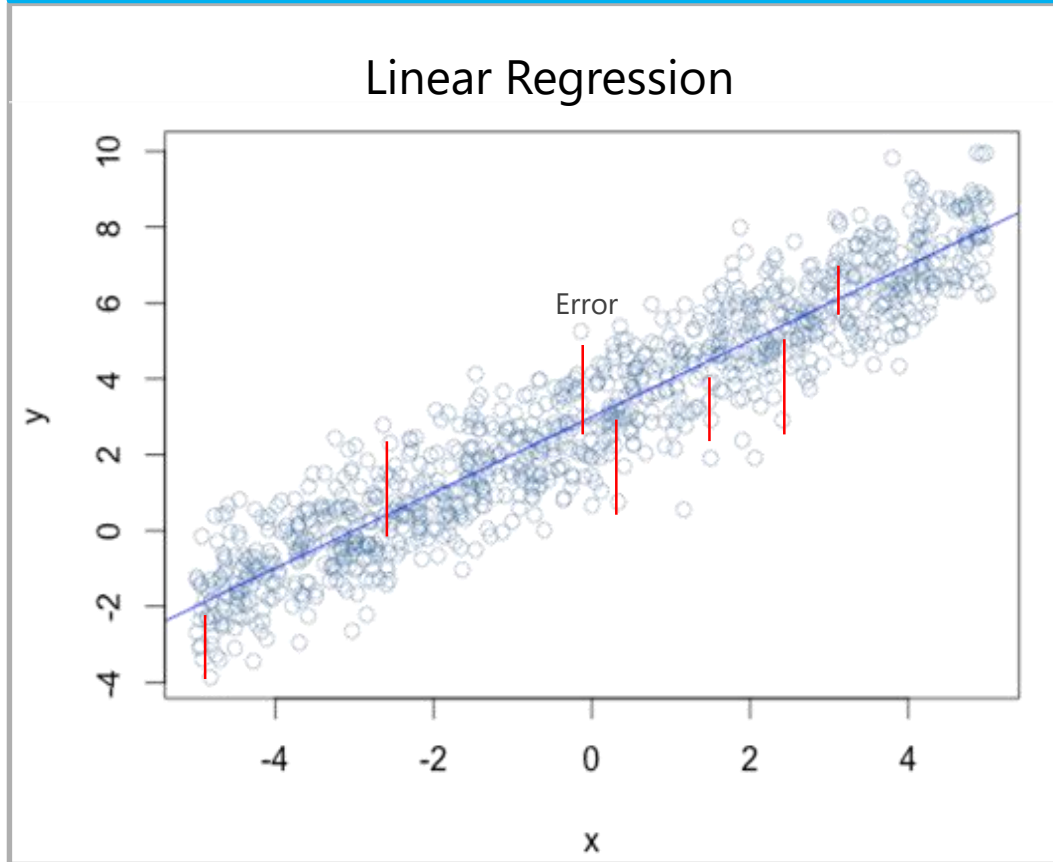
Simplest Case: Linear Regression, a parametric model

- In statistics, linear regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted X . The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.
- Assumptions:
 - Target function is linear
 - Residuals are homoscedastic and normally distributed
 - Independent variables are uncorrelated



Regression Performance Metrics

Regression Model Predictions VS Actual



Error Metrics

- Root Mean Square Deviation (RMSD)

$$\text{RMSD} = \sqrt{\frac{\sum_{t=1}^n (\hat{y}_t - y_t)^2}{n}}$$

- Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |f_i - y_i|$$

- Coefficient of Determination

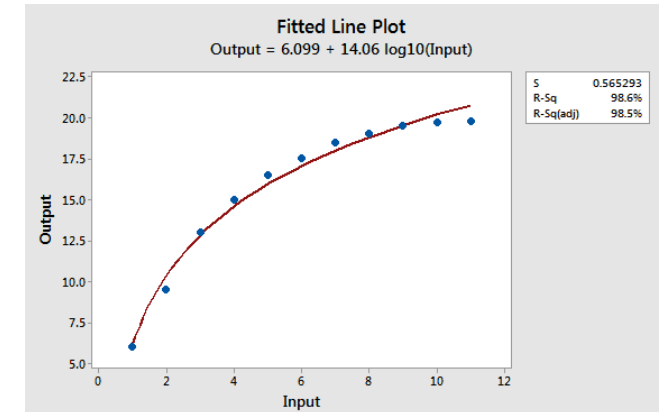
$$R^2 = 1.0 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

Performance Metrics

- **Regression coefficients** represent the mean change in the response variable for one unit of change in the predictor variable while holding other predictors in the model constant.
- The **p-value** for each term tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis.
- **R-squared** is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination
- The **t statistic** is the coefficient divided by its standard error.
- The **standard error** is an estimate of the standard deviation of the coefficient. It can be thought of as a measure of the precision with which the regression coefficient is measured.

Coefficients

Term	Coef	SE Coef	T	P
Constant	389.166	66.0937	5.8881	0.000
East	2.125	1.2145	1.7495	0.092
South	5.318	0.9629	5.5232	0.000
North	-24.132	1.8685	-12.9153	0.000



Model Comparison Metrics

- **Adjusted R square** – Adjusted R^2 also provides for the overall effectiveness of model. It compares how well the model is fitting the data.
 - Adjusted R^2 provides for evaluation among a set of models by penalizing addition of variables if the overall fit improved is not more than what would have been by chance
 - Helps reduce issue of Overfitting and keeping model variables to optimal levels
- The **Akaike information criterion (AIC)** is a measure of the relative quality of statistical models for a given set of data.
 - AIC deals with the trade-off between the goodness of fit of the model and the complexity of the model
 - AIC does not provide information for quality of model but given a set of models can help in choosing one over other based on model with minimum AIC
- **Bayesian information criterion (BIC)** or **Schwarz criterion** provides for relative comparison among a set of finite models.
 - BIC is based on likelihood function and assigns a penalty for adding additional parameters leading to overfitting
 - BIC is similar to AIC but assigns harsher penalty for overfitting

Generalization

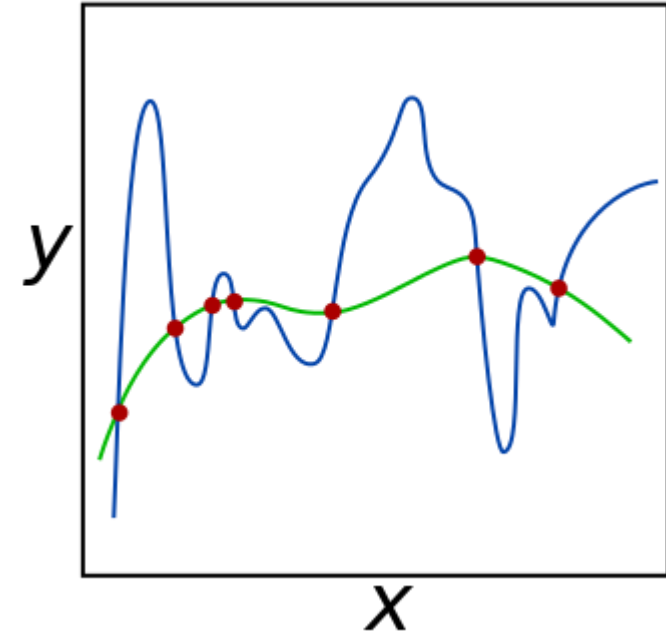
Generalization is the ability of a supervised algorithm to maintain performance when applied to a new dataset



Why doesn't my model work on new data?

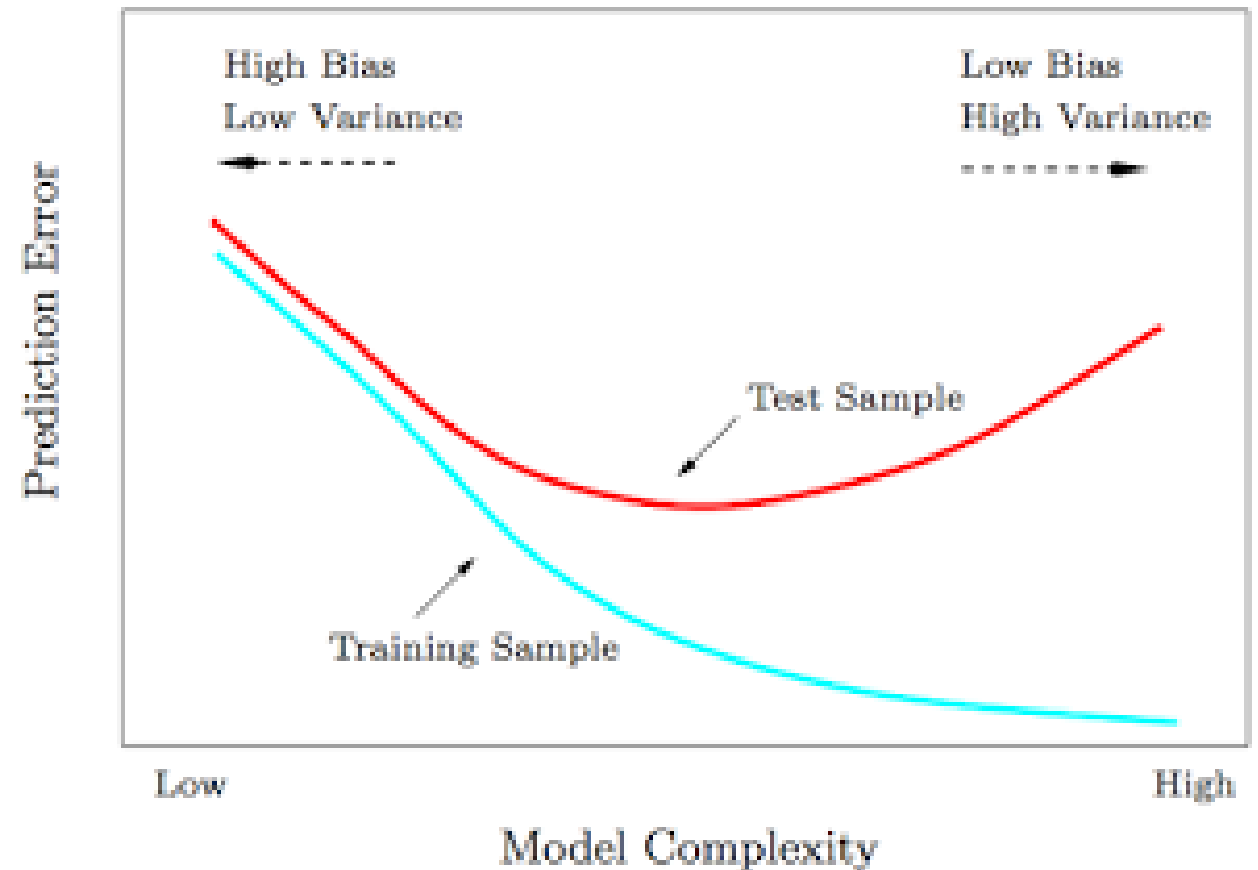
How to avoid extremes

- Overfitting can be avoided by reducing the complexity of the model
 - Regularization: Prevent the model from using too many dimensions (dependent variables)
 - OLS
 - Ridge
 - Random Forest
 - Tree Depth
 - Pruning
 - Naïve Bayes
 - Laplace Smoothing
 - Dimensionality Reduction before modeling
 - Principal Component Analysis
 - SVD
 - Mutual Information



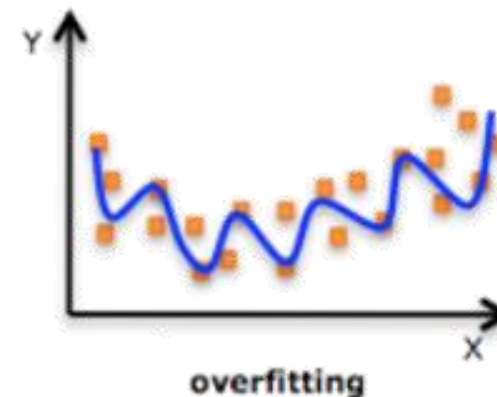
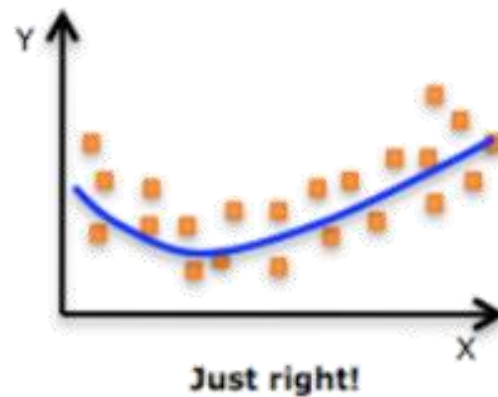
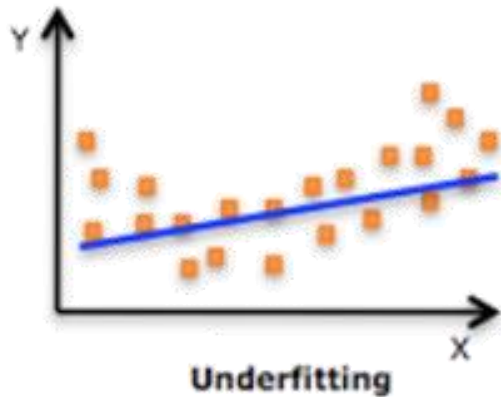
Bias / Variance Trade-Off

- In statistics and machine learning, the bias–variance tradeoff (or dilemma) is the problem of simultaneously minimizing two sources of error that prevent supervised learning algorithms from generalizing beyond their training set:
- The bias is error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (under-fitting).
- The variance is error from sensitivity to small fluctuations in the training set. High variance can cause overfitting: modeling the random noise in the training data, rather than the intended output



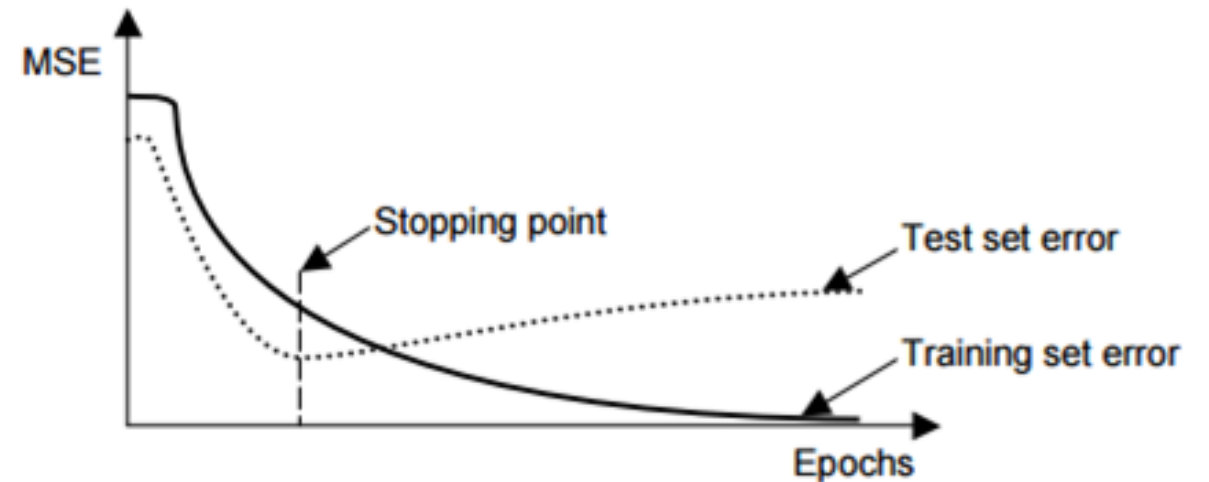
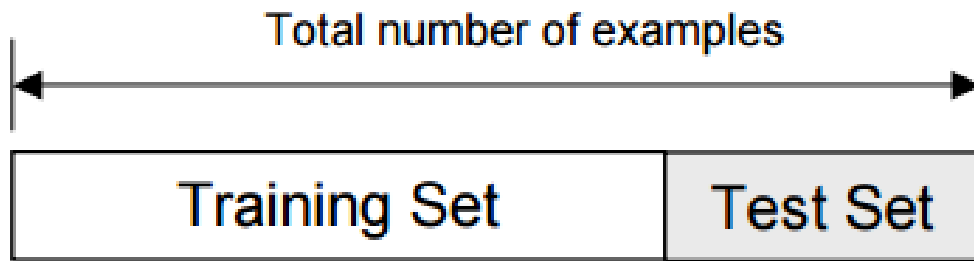
Two Extremes a good model must avoid

- The goal is to learn the relationships in the data and not the underlying variance in the data
- Overfitting occurs when the algorithm performs well on the training data set but performs poorly on any new data (validation, testing, real world)
- Underfitting occurs when the algorithm lacks the complexity to explain relationships in data (predicting the mean)



Assessing model performance: The Holdout Method

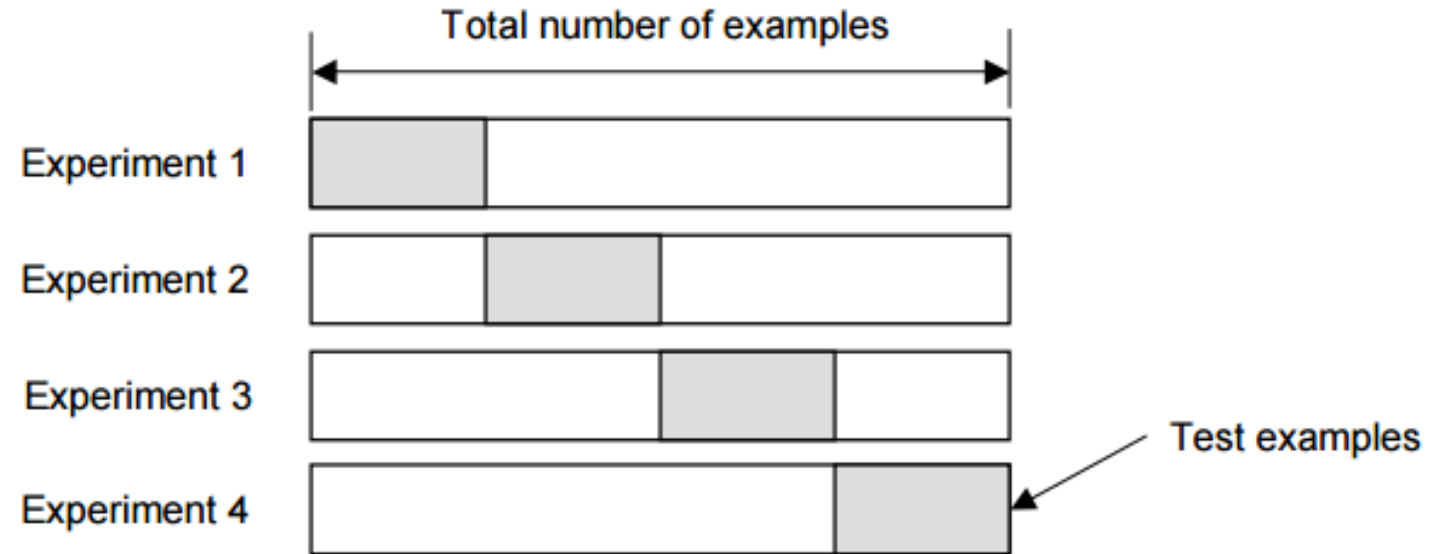
- Split dataset into two groups n
 - Training set: used to train the classifier n
 - Test set: used to estimate the error rate of the trained classifier
- A typical application the holdout method is determining a stopping point for the back propagation error



Assessing Model Performance: K-Fold Cross Validation

- Create a K-fold partition of the dataset n For each of K experiments, use K-1 folds for training and the remaining one for testing
- K-Fold Cross validation is similar to Random Subsampling n
 - The advantage of K-Fold Cross validation is that all the examples in the dataset are eventually used for both training and testing
- The true error is estimated as the average error rate

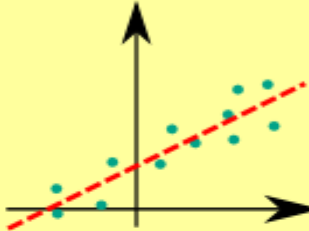
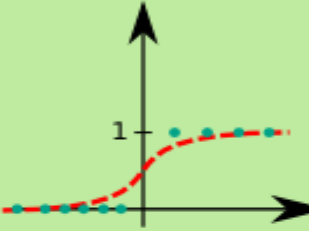
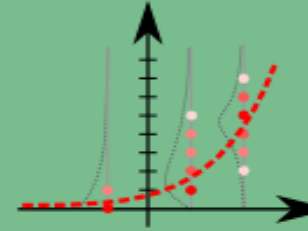
$$E = \frac{1}{K} \sum_{i=1}^K$$



Generalized Linear Models (GLM)

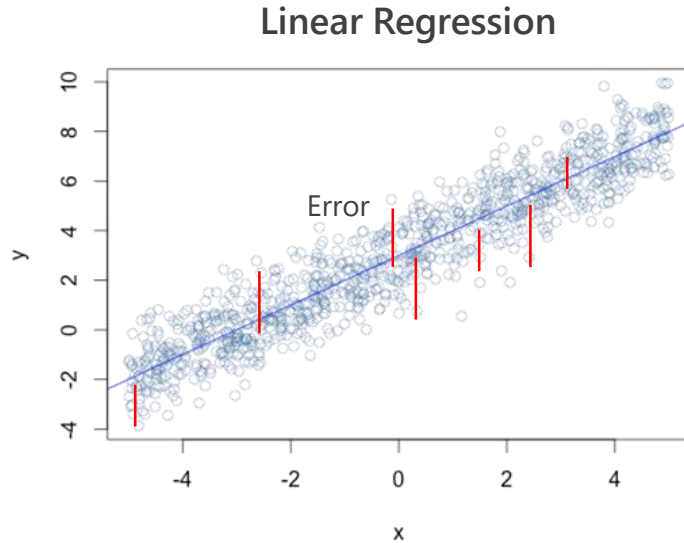
GLM extend linear regression and allow the response variable to have distributions other than normal.

- Logistic Regression Models can be used to predict a binary: pass/fail, win/loss probabilities.
- Poisson Regression used to predict a count or rate: How many failures this season?, How many months until failure (RUL)

LINEAR REGRESSION	LOGISTIC REGRESSION	POISSON REGRESSION
<ul style="list-style-type: none">① Econometric modelling② Marketing Mix Model③ Customer Lifetime Value	<ul style="list-style-type: none">① Customer Choice Model② Click-through Rate③ Conversion Rate④ Credit Scoring	<ul style="list-style-type: none">① Number of orders in lifetime② Number of visits per user
		
Continuous \Rightarrow Continuous	Continuous \Rightarrow True/False	Continuous \Rightarrow 0,1,2,...
$y = \alpha_0 + \sum_{i=1}^N \alpha_i x_i$	$y = \frac{1}{1 + e^{-z}}$ $z = \alpha_0 + \sum_{i=1}^N \alpha_i x_i$	$y \sim \text{Poisson}(\lambda)$ $\ln \lambda = \alpha_0 + \sum_{i=1}^N \alpha_i x_i$
<code>lm(y ~ x1 + x2, data)</code>	<code>glm(y ~ x1 + x2, data, family=binomial())</code>	<code>glm(y ~ x1 + x2, data, family=poisson())</code>
1 unit increase in x increases y by α	1 unit increase in x increases log odds by α	1 unit increase in x multiplies y by e^{α}

Ordinary Least Square

Ordinary Least Squares Regression



Line Equation: $y_i = h(\mathbf{x}_i, \mathbf{w}) = \mathbf{w}^T \mathbf{x}_i$

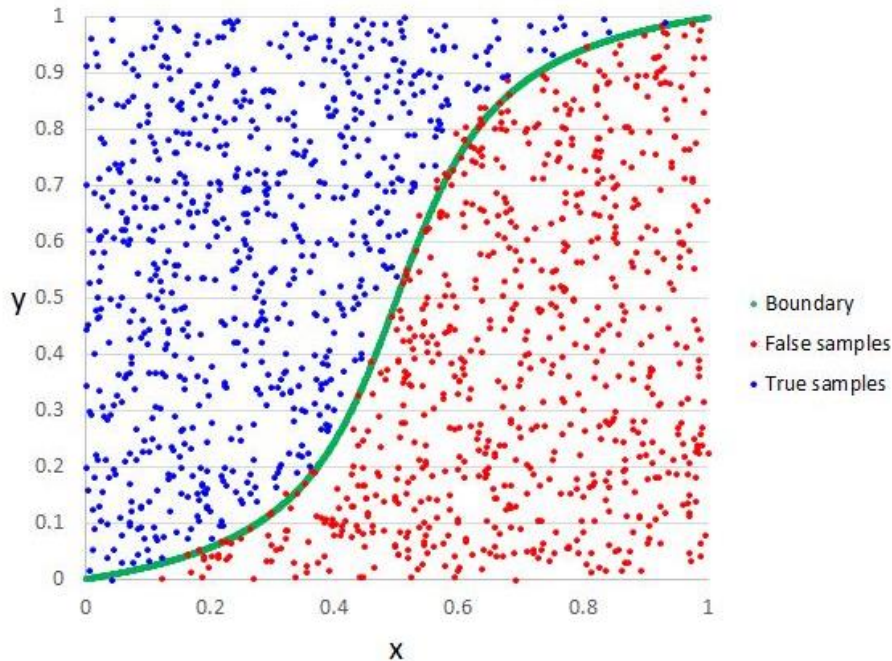
Cost Function: $L(\mathbf{w}) = \sum_i (h(\mathbf{x}_i, \mathbf{w}) - y_i)^2$

- Best fitting curve can be solved for directly by minimizing the cost function
- Relationship between independent variables and dependent variable can be viewed directly in resulting equation

- Ordinary least squares (OLS) or linear least squares is a method for estimating the unknown parameters in a linear regression model
- The goal here is to minimize the sum of the squares of the differences between the observed responses in the given dataset and those predicted by a linear function of a set of explanatory variables
- At a very basic level, the relationship between a continuous response variable (Y) and a continuous explanatory variable (X) may be represented using a line of best-fit, where Y is predicted, at least to some extent, by X

Logistic Regression

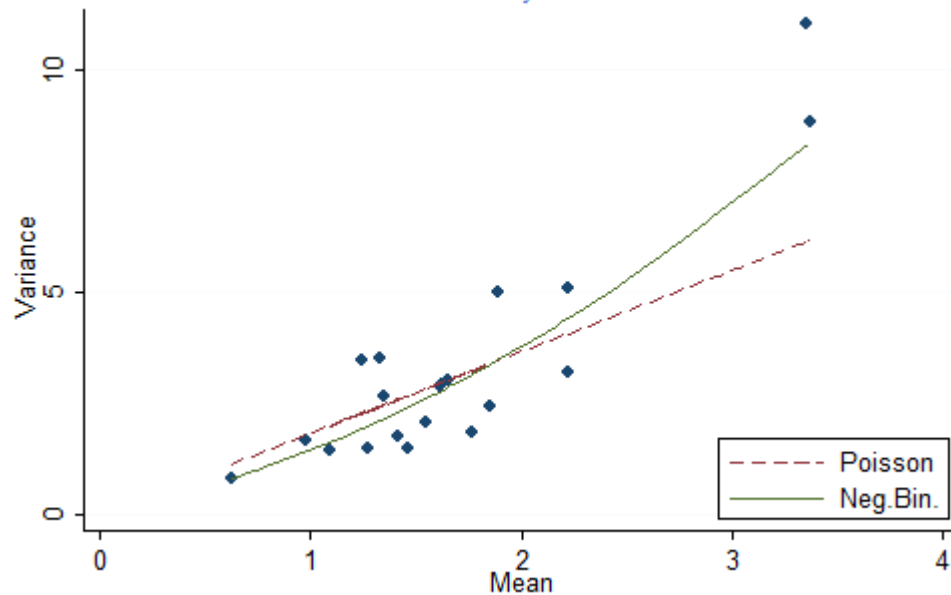
Logistic Regression



- Logistic regression, or logit regression is a regression model where the dependent variable is categorical:
 - It can be Binary dependent variables, that is, where it can take only two values, such as pass/fail, win/lose, alive/dead or healthy/sick.
 - Cases with more than two categories are referred to as multinomial logistic regression
- Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities
- The model of logistic regression, however, is based on quite different assumptions from those of linear regression:
 - Conditional distribution y/x is a Bernoulli distribution rather than a Gaussian distribution, because the dependent variable is binary
 - Predicted values are probabilities and are therefore restricted to $(0,1)$
- Disadvantages: Fundamentally a binary classifier, Hard to make incremental

Poisson Regression

Poisson Regression



$$y_i \sim \text{Poisson}(u_i)$$

$$\log u_i = y_i = \alpha + \beta_i x_i + \eta^1 z_i + f(t; \lambda) + \varepsilon_i$$

Outcome/
event

Main
exposure

Vector of
measured
covariates

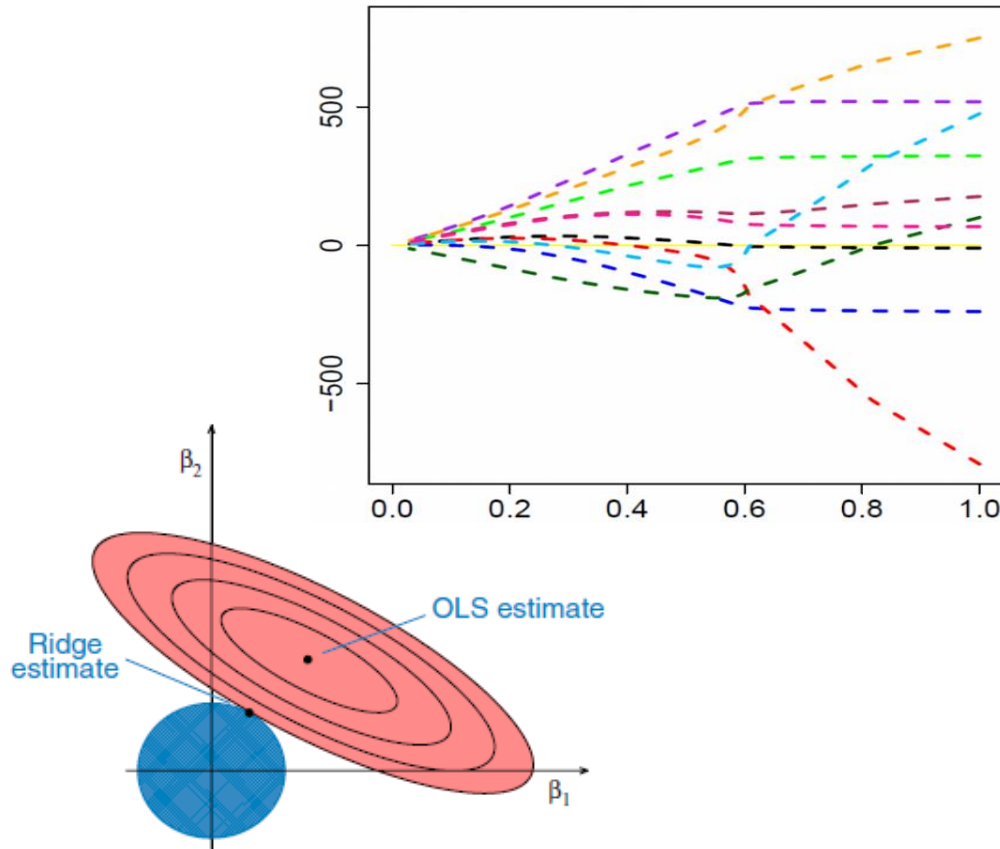
Smooth
function

Error term

- Poisson regression is also a type of GLM model where the random component is specified by the Poisson distribution of the response variable which is a count
- Poisson regression assumes the response variable Y has a Poisson distribution, and assumes the logarithm of its expected value can be modeled by a linear combination of unknown parameters
- In Poisson regression Response/outcome variable Y is a count. But we can also have Y/t , the rate (or incidence) as the response variable, where t is an interval representing time, space or some other grouping

Ridge Regression

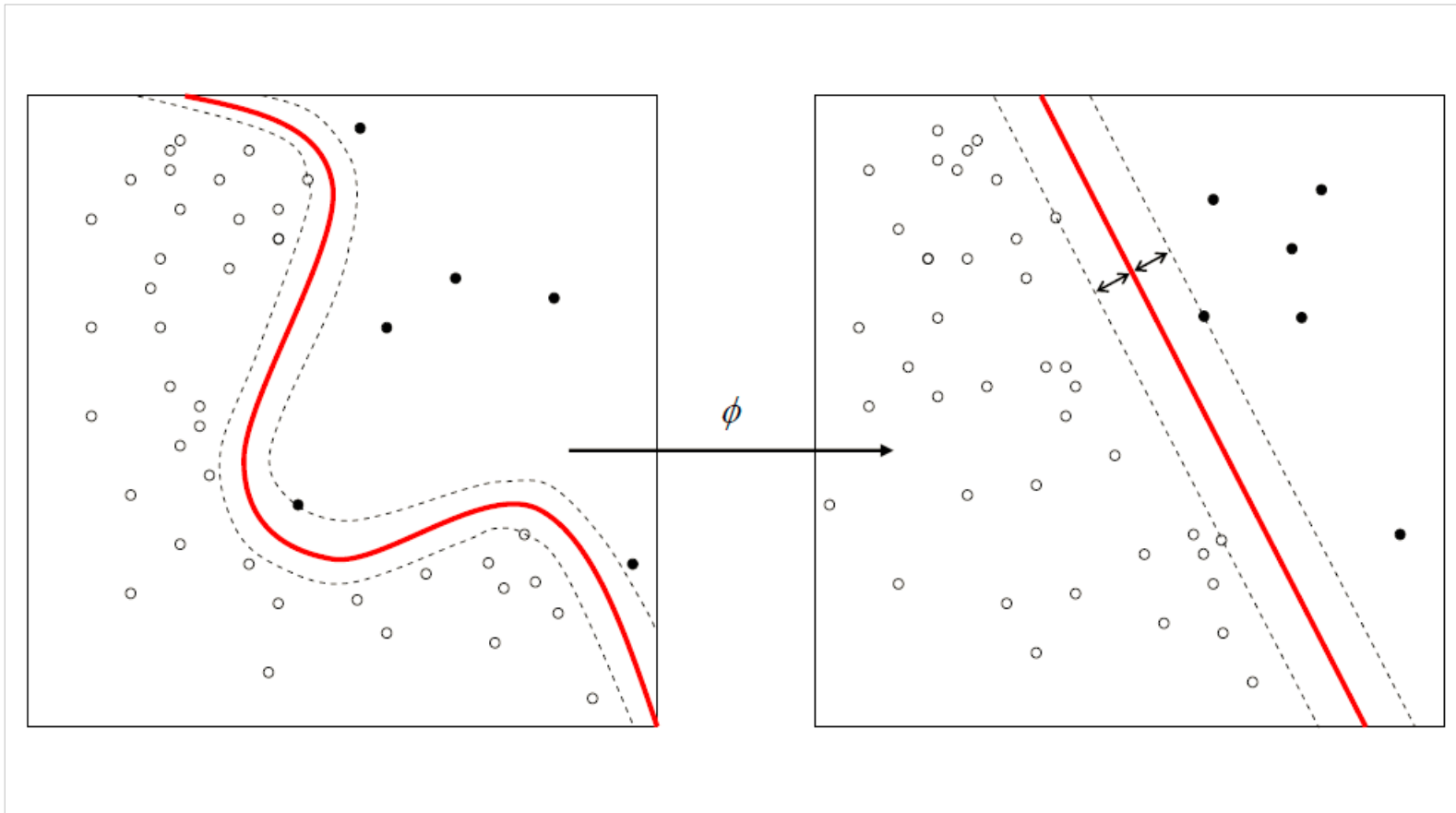
Ridge Regression



- Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity
- When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. Multicollinearity can create inaccurate estimates of the regression coefficients, inflate the standard errors of the regression coefficients, etc.
- By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors

Classification Model

Classification is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations



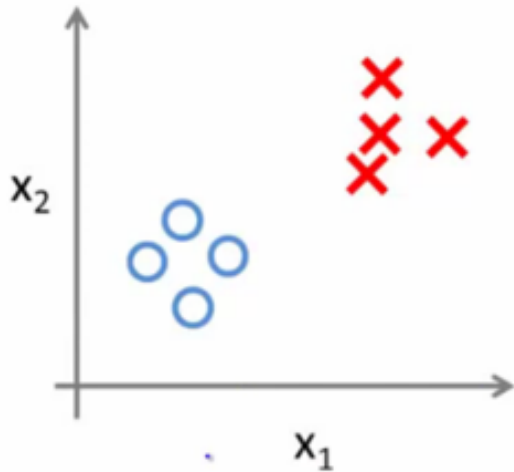
Common Classification Examples

- **SPAM filtering** – Email engines using classification to filter out unwanted emails
- **Marketing response models** – Predict if customer will respond or not to a promotion
- **Well Site Failure**- Predict if there is an impending well downtime/failure
- **Face Detection** - Using pattern recognition to identify faces
- **Sentiment Analysis** – Assigning sentiment(positive, neutral, negative) to text/voice
- **Document Classification** – Classifying documents into pre defined buckets

Classification Types

Binary Classification

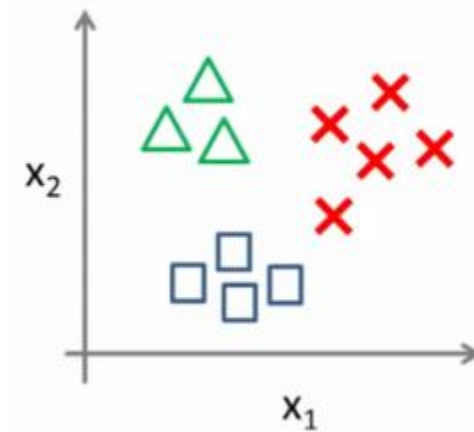
Categorizing data into two distinct classes



Ex- Prediction if it will rain or not tomorrow

Multi-Class Classification

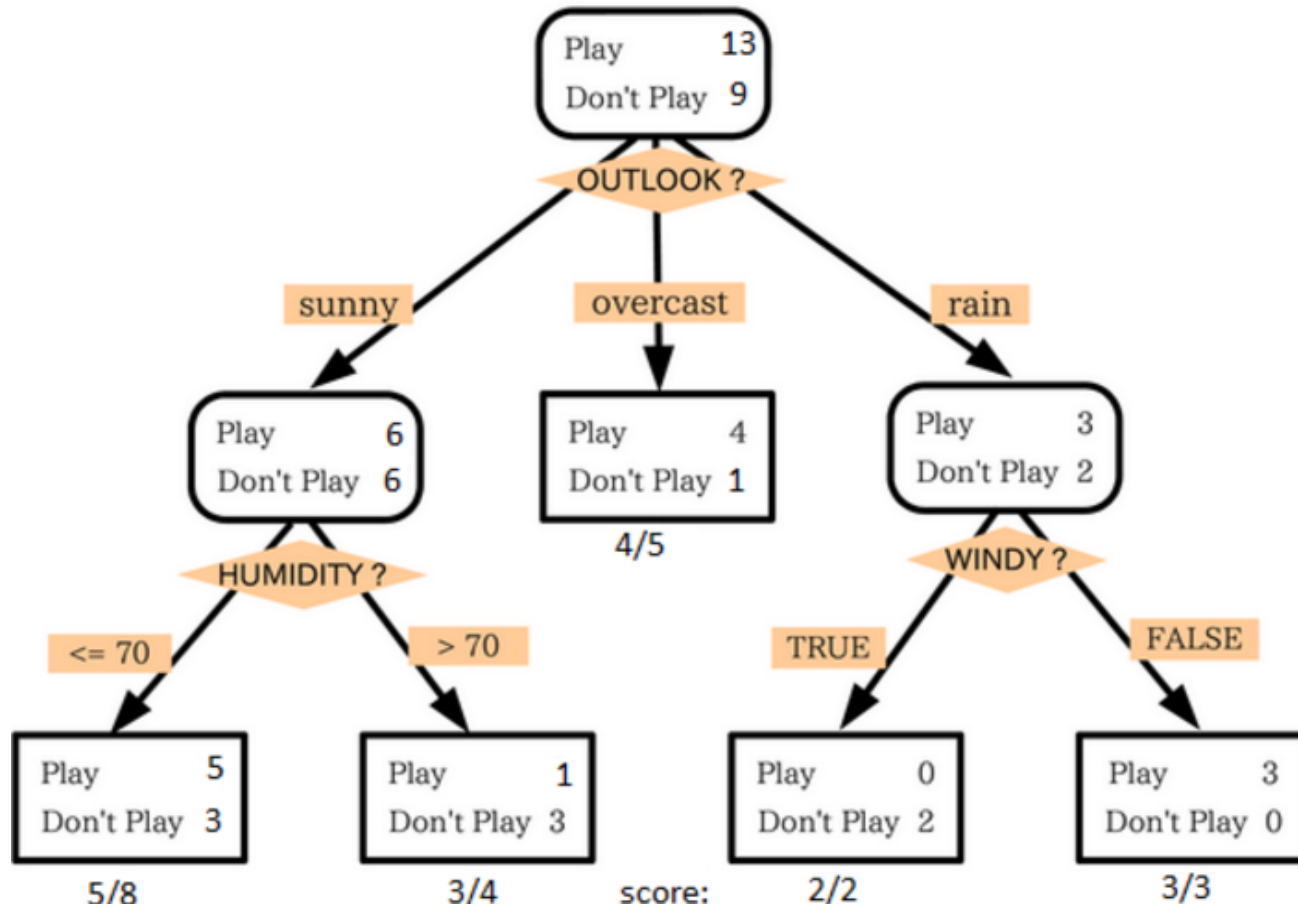
Categorizing data into more than two classes



Example – Prediction for weather as sunny, windy or rainy

Classification with Decision Trees

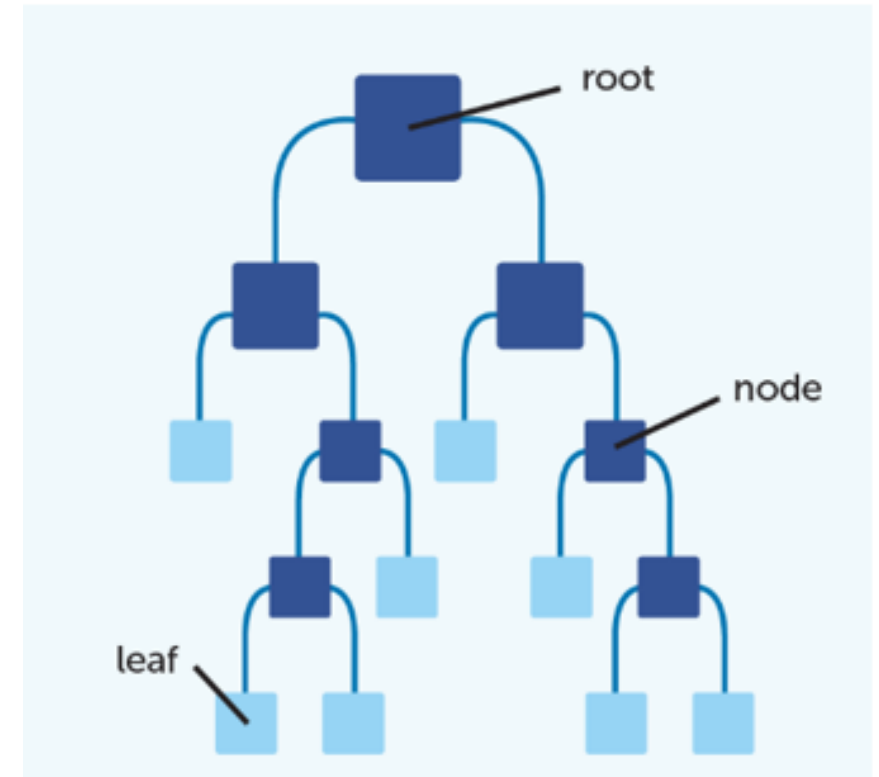
Goal is to determine likelihood of children going outside for play



- Decision trees develop a set of rules that are designed to separate the two classes effectively
- Decision trees can be used for both regression and classification problems
- Advantages -
 - Easily interpretable
 - Easier to reconfigure based on business rules
 - Fast to train and evaluate

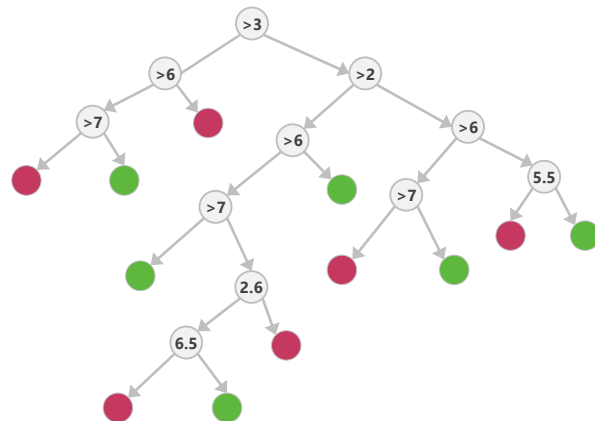
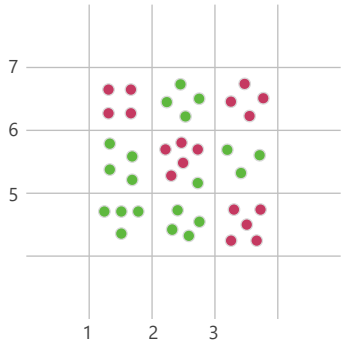
Building a Decision Tree – Key steps

1. Define the rule to split the data starting at root
2. Split data into two disjoint datasets based on the most influential variable
3. Repeat iteratively at each node keeping in mind the constraints for split – minimum population and split rule
4. Stop when leaves are almost 'pure' ~ purity is defined by Gini Index or entropy



Decision Tree Assumptions and Logic

Decision Tree



Information Gain

$$IE(f) = - \sum_{i=1}^m f_i \log_2 f_i$$

- Trees are made up of branches (decision points) and leaves. Each leaf corresponds to a classification
- Splits are selected to maximize gain on subset of features
- Split is stopped based on how “pure” the node is and other assumptions such as minimum population size → leaf
- Length of Decision trees should be an optimal cut off
 - We are classifying most records in training data
 - Not making the model complex or over fit training data
- Approach to optimal tree length is to build a large tree on training data and prune back for

Confusion Matrix

- A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known
- It identifies the model's ability to correctly identify instances of "good" and "bad"
 - TN – is the number of **correct** predictions that an instance is **negative**
 - FP – is the number of **incorrect** predictions that an instance is **positive**
 - FN – is the number of **incorrect** of predictions that an instance **negative**
 - TP – is the number of **correct** predictions that an instance is **positive**

		Actual Value (as confirmed by experiment)	
		positives	negatives
Predicted Value (predicted by the test)	positives	TP True Positive	FP False Positive
	negatives	FN False Negative	TN True Negative

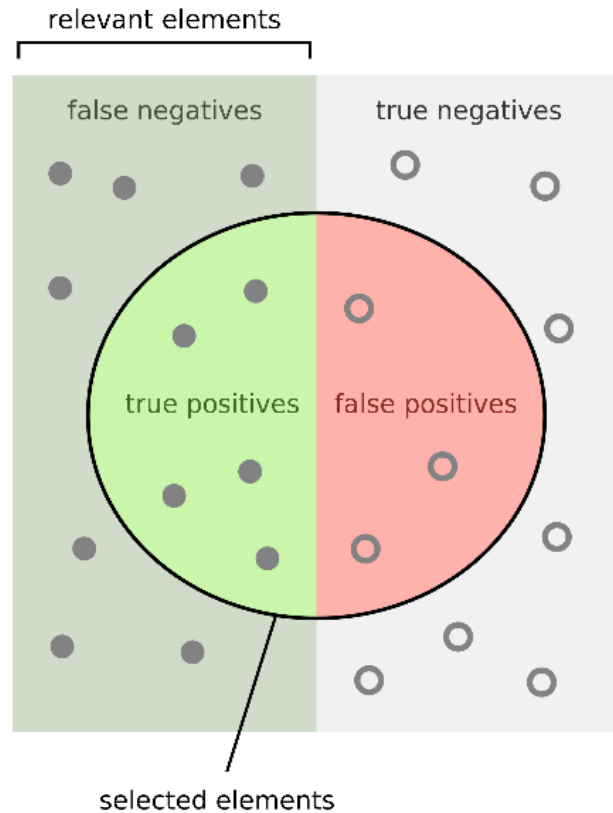
Confusion Matrix

Scenario – Student performance model which picks student scores, demographics and behavioral data to predict who will successfully graduate from school

- 1000 Students
- Actual data – 800 Pass, 200 Fail
- Predicted data – 780 Pass, 220 Fail
- TN – 120
- FP – 150
- FN – 100
- TP – 630

		Actual Value	
Predicted Value		Pass	Fail
	Pass	630	150
	Fail	100	120

Performance Metrics



How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

Recall or *true positive rate (TP)* is the proportion of positive cases that were correctly identified

Precision (P) is the proportion of the predicted positive cases that were correct

Accuracy (AC) is the proportion of the total number of predictions that were correct

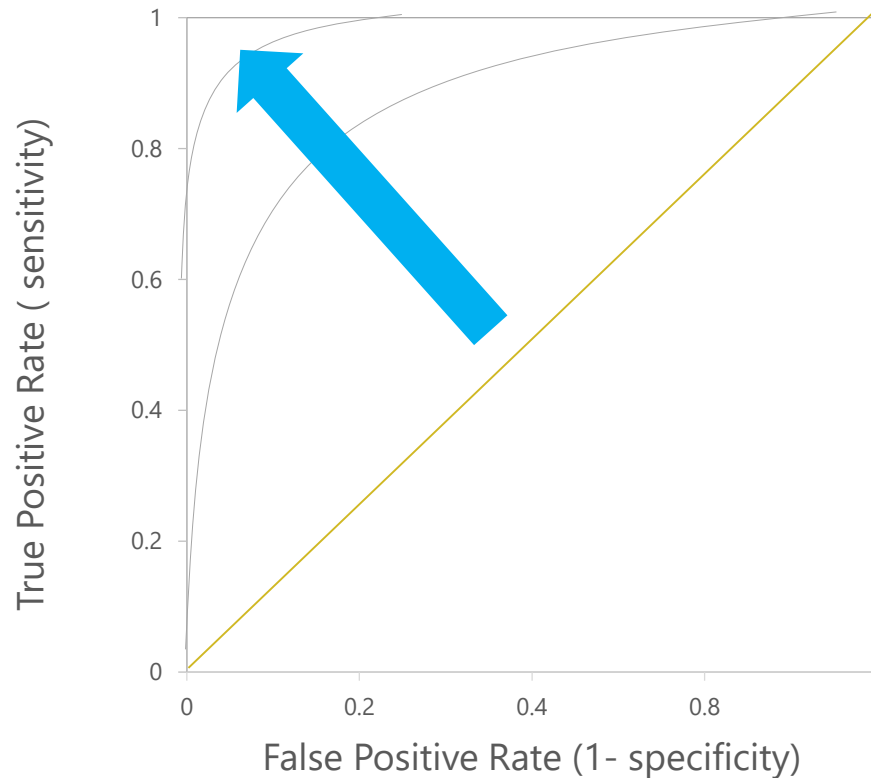
Performance Metrics – Student Graduation Model

- **Precision:** When it predicts pass, how often is it correct?
 - $TP / \text{predicted yes} = 630 / 780 = 81\%$
- **Recall:** When it's actually pass, how often does it predict pass?
 - $TP / \text{actual yes} = 630 / 730 = 86\%$
- **Accuracy:** Overall, how often is the classifier correct?
 - $(TP + TN) / \text{total} = (630 + 120) / 1000 = 75\%$
- **Misclassification Rate:** Overall, how often is it wrong?
 - $(FP + FN) / \text{total} = (150 + 100) / 1000 = 25\%$

		Actual Value	
Predicted Value		Pass	Fail
	Pass	630	150
	Fail	100	120

Receiver Operating Characteristic (ROC) Curve

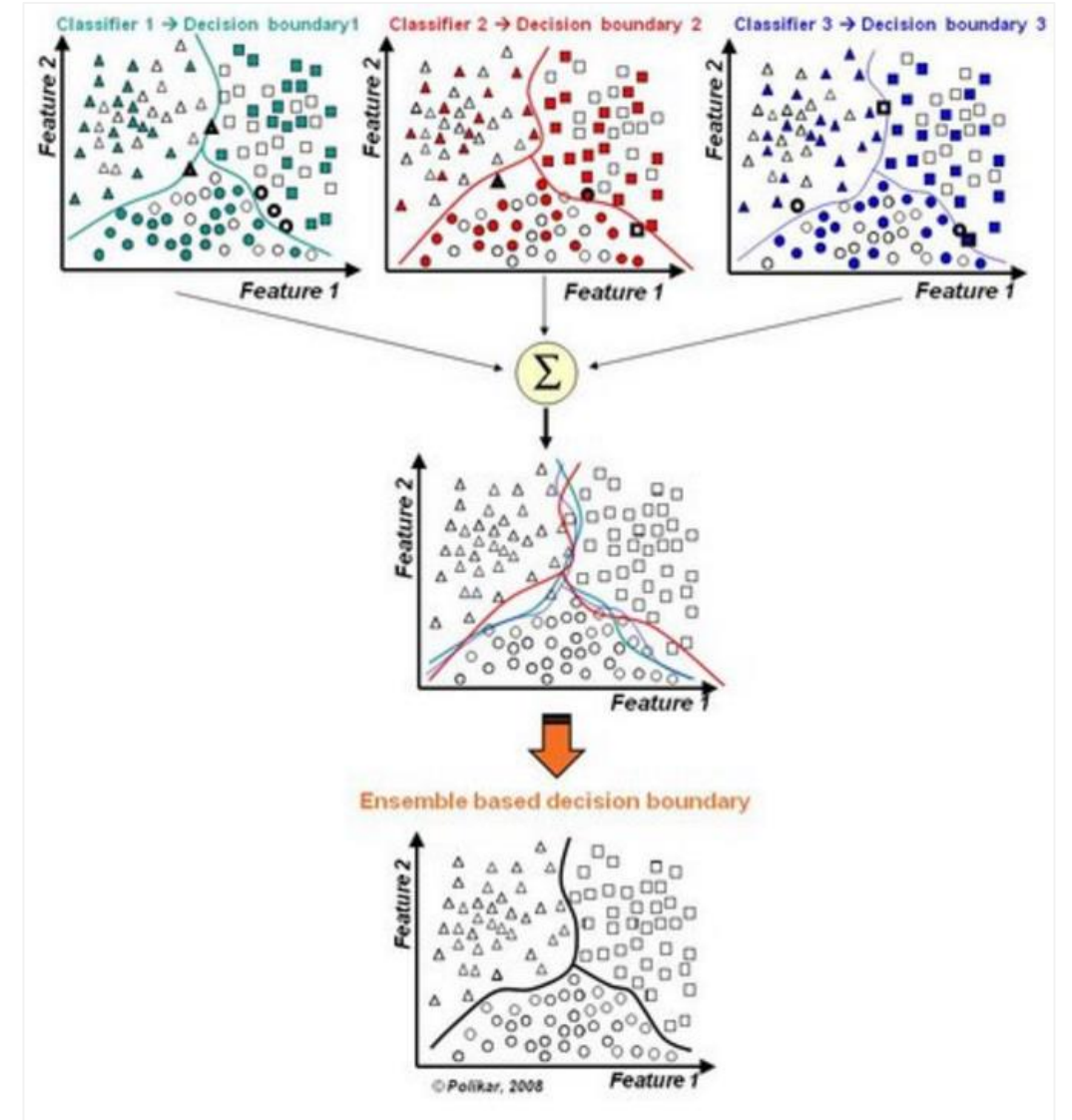
- Graphical representation of the performance of classification model
- Useful in deciding optimal model based on consideration of model success or cost/benefit analysis



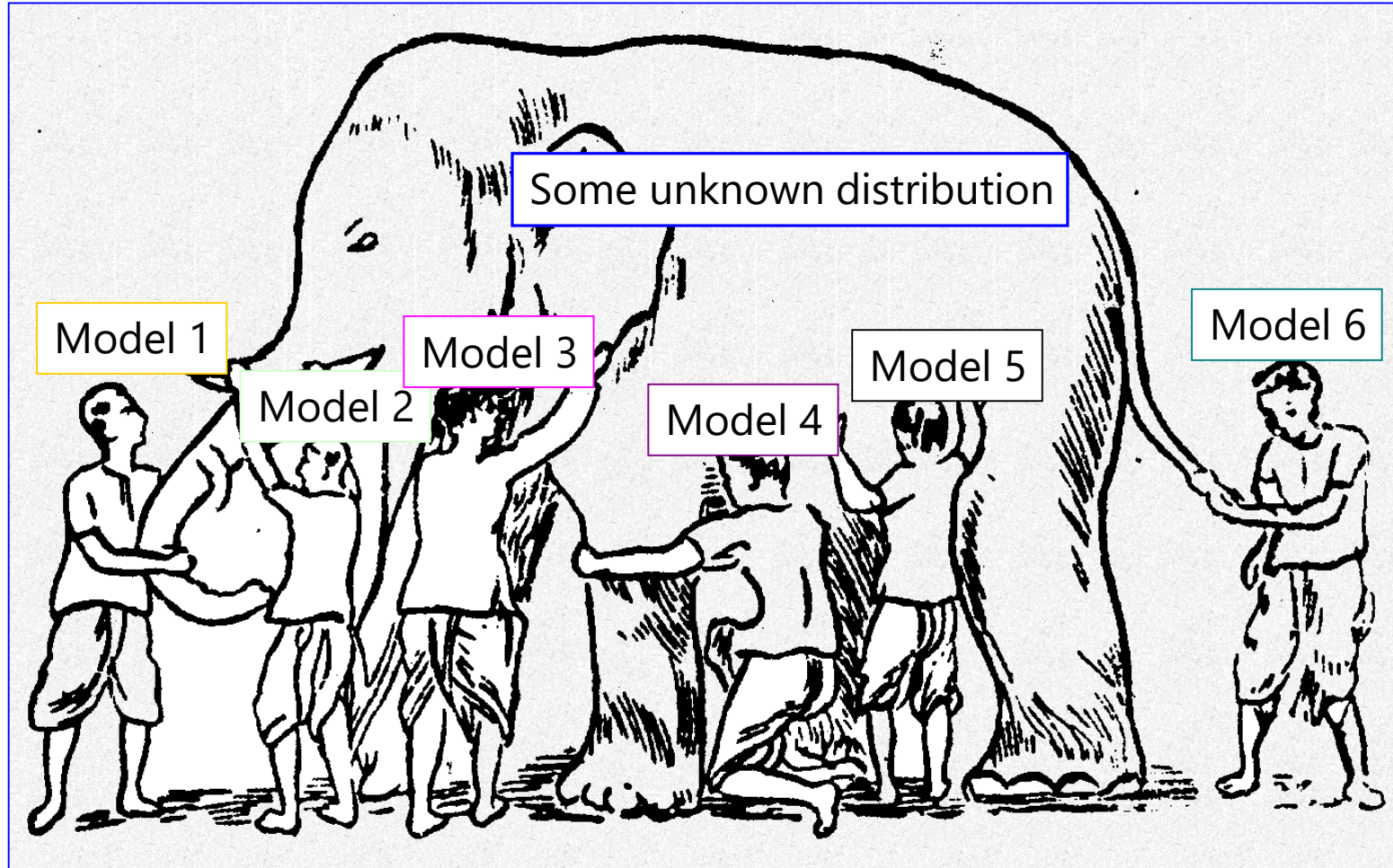
Ensemble Classifiers

- Use multiple classifiers instead of single classifier to generate more accurate and representative model
- Final output will combine predictions from multiple classifiers

- ✓ Reduces variance from single training data
- ✓ Reduces bias from single classifier



Ensemble Classifiers captures more comprehensive picture



Forming an Ensemble

Bagging

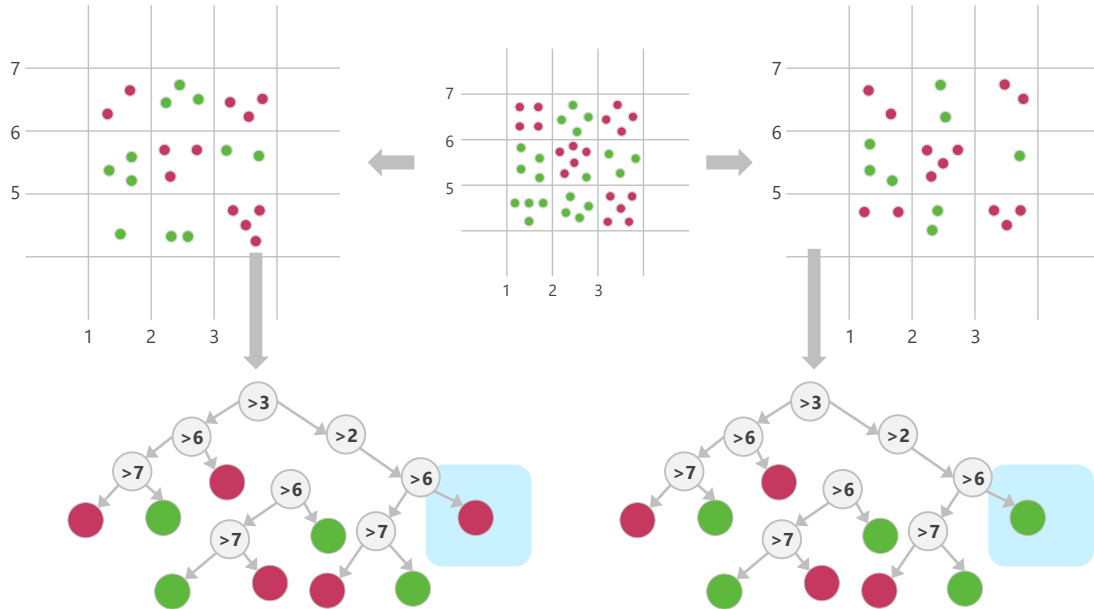
- Create multiple random sub samples of same dataset
- Build classifier on each sample
- Combine predictions to get the final outcome using majority vote
- Good with noisy data

Boosting

- Creates multiple sub samples of training data with additional weighting for instances which most recent classifier predicted incorrectly
- Similar to Bagging but ensure more representation by fine tuning across all instances which haven't been predicted accurately
- Combines predictions through a weighted vote based on model accuracy
- Good with noise free data

Decision Forest

Decision Forest

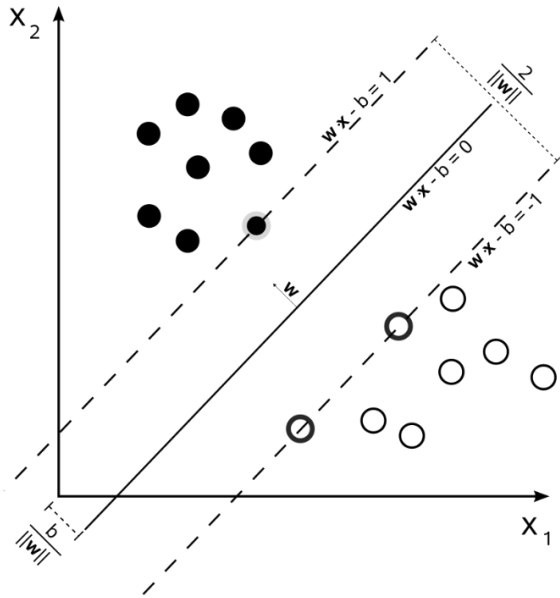


- Multiple trees are generated to avoid overfitting. Each tree gets a vote regarding the classification of the observation
- Each tree sees random selection of training set

- One of the most useful aspects of decision forests is that they force you to consider as many possible outcomes of a decision as you can think of
- A decision forest can help you weigh the likely consequences of one decision against another
- A drawback of using decision forests is that the outcomes of decisions, subsequent decisions and payoffs may be based primarily on expectations. When actual decisions are made, the payoffs and resulting decisions may not be the same as those you've planned for. It may be impossible to plan for all contingencies that can arise as a result of a decision.
- This can lead to an unrealistic decision tree that could guide you toward a bad decision

Support Vector Machine

Support Vector Machine



$$\arg \min_{(w, b)} \frac{1}{2} ||w||^2$$

subject to
(for any $i = 1, \dots, n$)

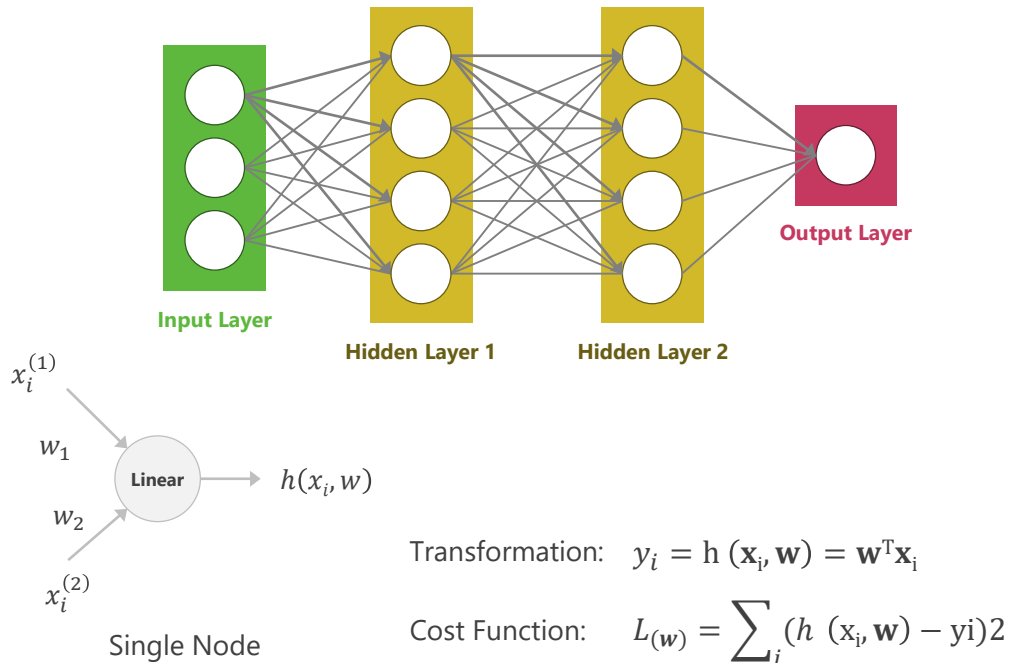
$$y_i(w \cdot x_i - b) \geq 1$$

- Creates a hyperplane that maximizes separation between known data points and axis origin
- Hyperplane is used to categorize new observations. Observations that not classified with existing data are labelled as anomalies

- Support vector machines (SVMs) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis
- Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier
- Unique features of SVM include:
 - Not affected by local minima
 - Do not suffer from the curse of dimensionality
- Disadvantages:
 - Picking/finding the right kernel can be a challenge
 - Results/output are incomprehensible
 - No standardized way for dealing with multi-class problems; fundamentally a binary classifier

Neural Networks

Neural Networks

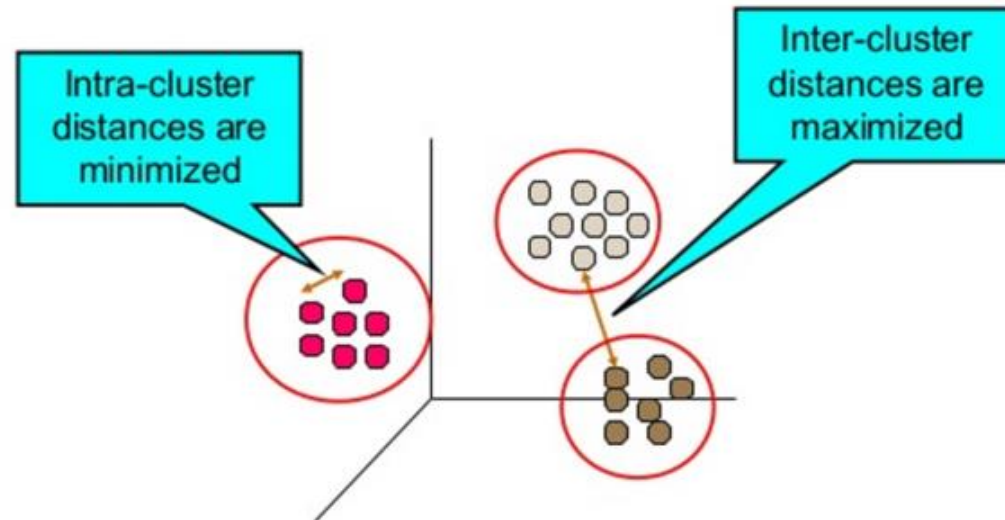


- Neural networks are a combinations of neurons. Each neuron applies a transformation function to input variables
- Backpropagation algorithm used to determine transformation function over entire network

- A neural network can be thought of as a network of “neurons” organized in layers:
 - The predictors (or inputs) form the bottom layer, and
 - The forecasts (or outputs) form the top layer
 - There may be intermediate layers containing “hidden neurons”
- The very simplest networks contain no hidden layers and are equivalent to linear regression
- The coefficients attached to these predictors are called “weights”. The forecasts are obtained by a linear combination of the inputs. The weights are selected in the neural network framework using a “learning algorithm” that minimizes a “cost function” such as MSE
- Disadvantages:
 - Picking the correct topology is difficult
 - Training takes a long time/requires a lot of data

Clustering

- Clustering is the task of **grouping** a set of **objects** in such a way that objects in the same **group** or cluster are more **similar** (in some sense or another) to each other than to those in other groups
- Form of unsupervised classification with **no** labels
- Useful for understanding the distribution of data and feature transformation into more *intuitive* and *usable* information



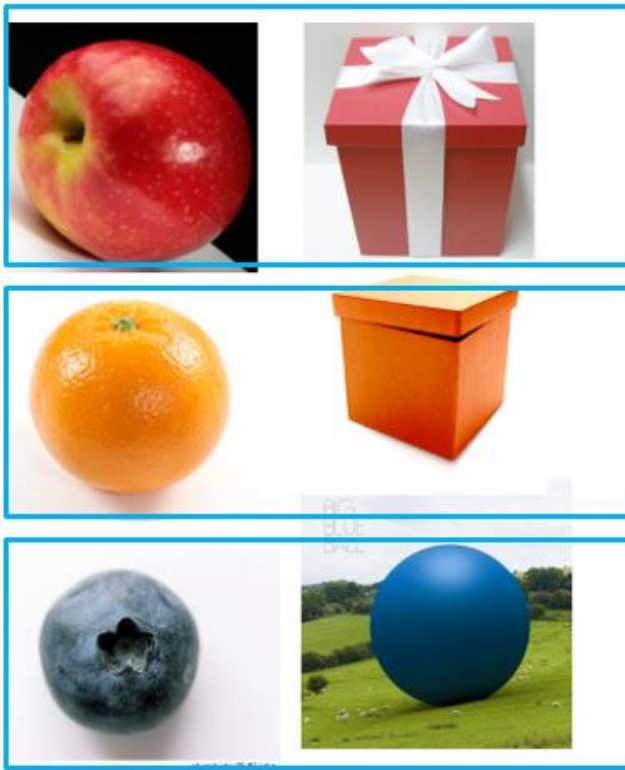
Cluster Analysis: No One Right Way!



How many clusters do you expect?

Forming Clusters – Using one metric to calculate distance

Cluster by Colour



Cluster by Shape



Forming Clusters – Using more than one metric

Cluster by Shape



- More complex relationships can be uncovered by adding more variables
- Care must be taken to properly weight variables so that one does not dominate the algorithm results.

Using Clustering To Find Anomalies

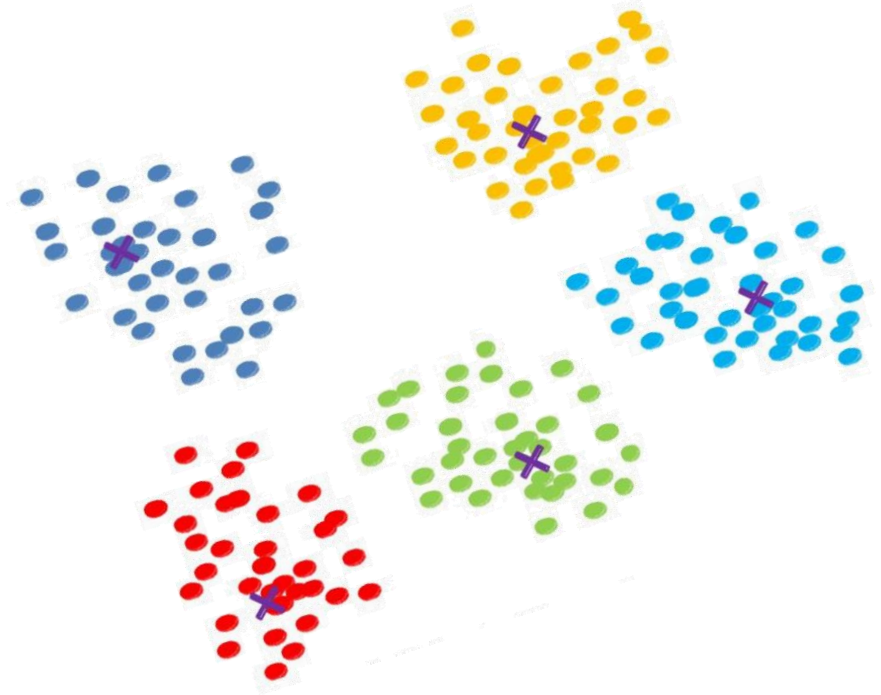
Search for Outliers



- Look for observations that are either consistently placed far from the centroid or are included in relatively small or even their own clusters.
- Look to the distance metric for guidance.

k-Means Clustering

- K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem
- Theoretically k-means clustering aims to partition 'n' observations into 'k' clusters
- Each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.



k-Means Clustering

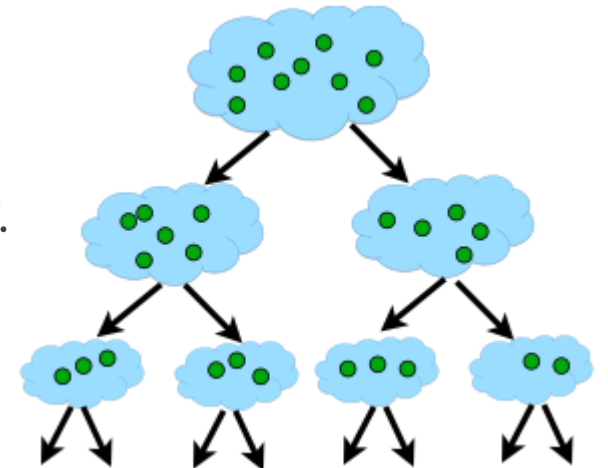
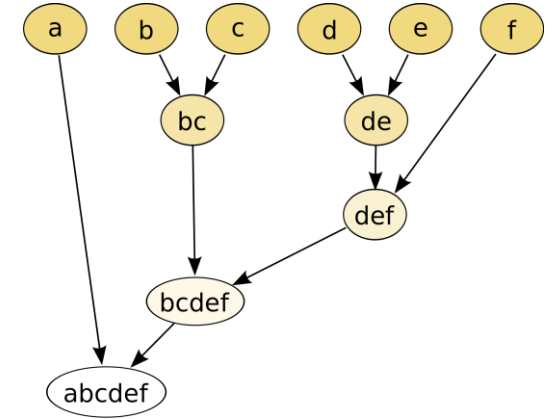
- Generate starting points for k initial random centroid
- Partition the data by associating each data point with the closest centroid
- Recalculate the mean point, or centroid for each set
- Repeat the process until convergence (stability) is achieved



Hierarchical Clustering

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

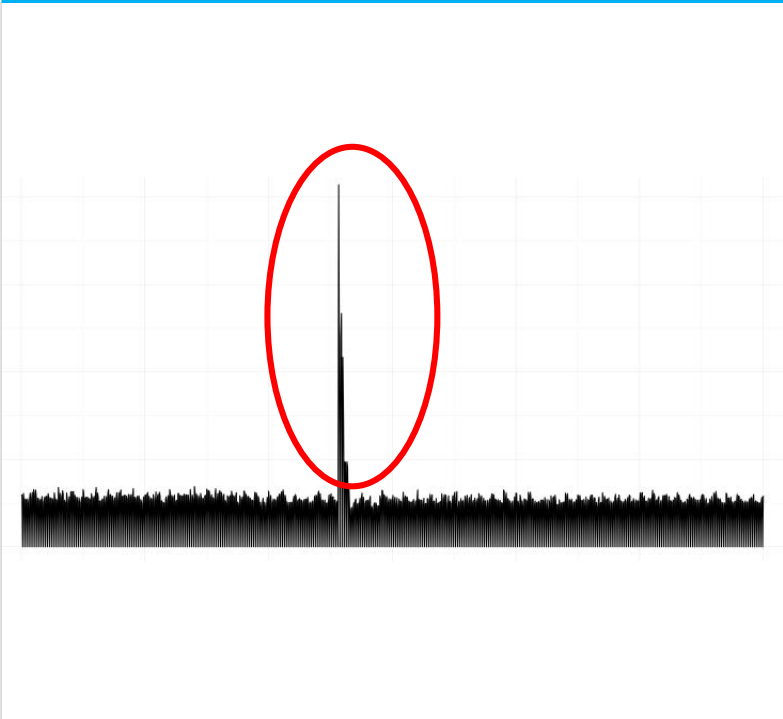
- **Agglomerative:** This is a "bottom up" approach
 - Initially, each point is a cluster
 - Repeatedly combines the two "nearest" clusters into one
- **Divisive:** This is a "top down" approach:
 - All observations start in one cluster
 - Splits are performed recursively as one moves down the hierarchy.



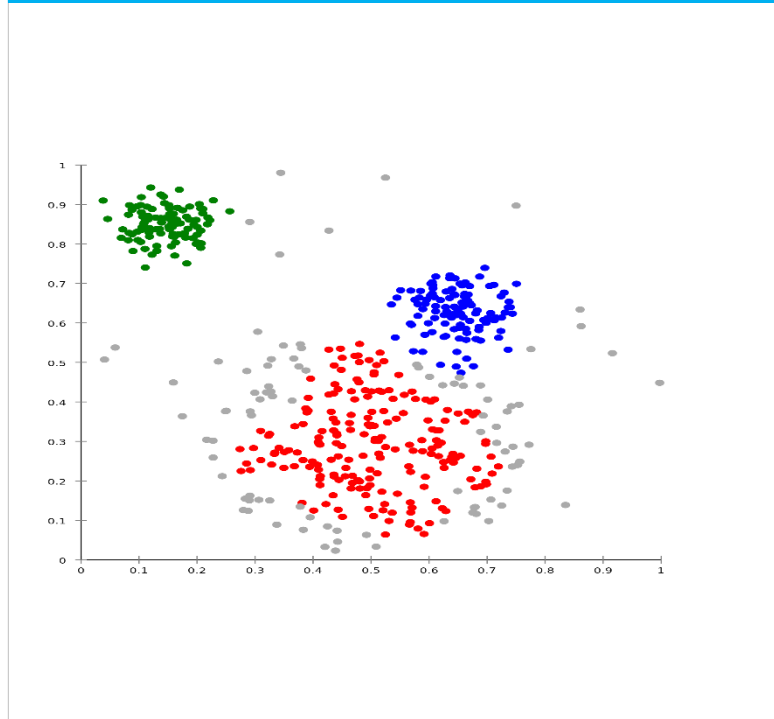
Anomaly Detection

- Anomaly detection helps in identifying events or observations which are outside the bounds of expected pattern. It allows prior experience and given data to identify scope of “normal” and “failure”
- Extremely useful for detecting “unknown” failure patterns which haven’t occurred before or happened with varying intensity
- In Oil & Gas scenario where downtime will lead to big losses building an automated, self-adaptive anomaly detection system is crucial

Anomaly Detection



Modeling Approach



Modelling Approach

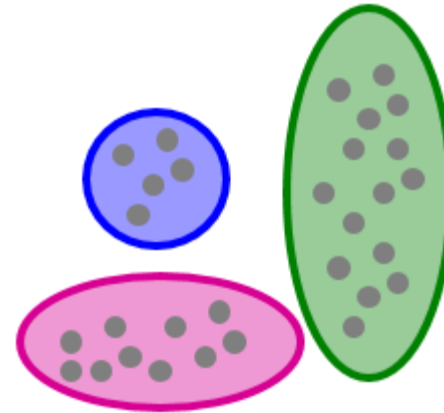
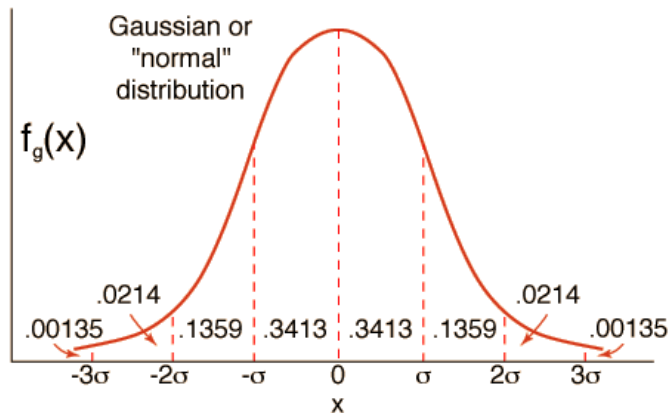


Bradley-Fayyad-Reina

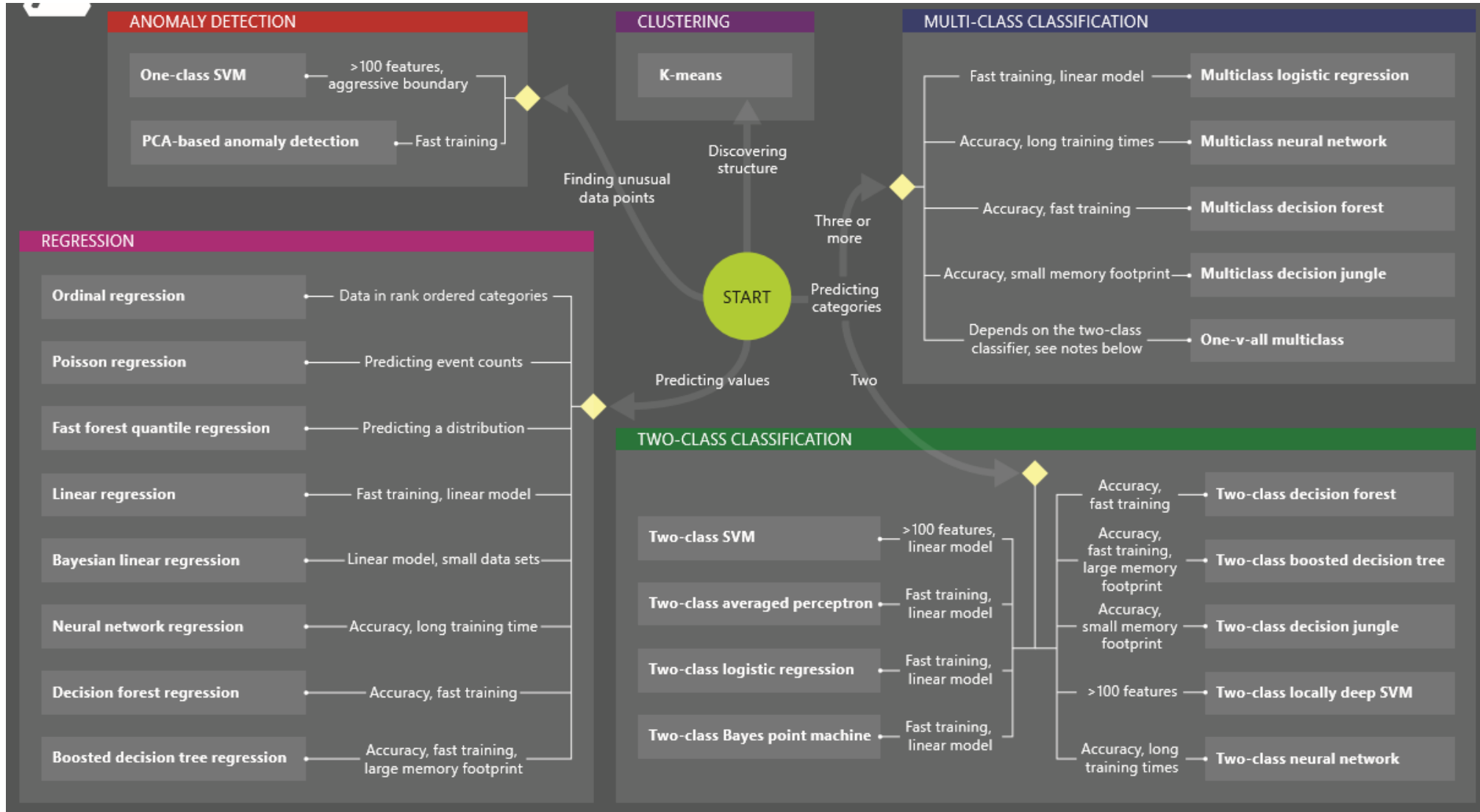
BFR (Bradley-Fayyad-Reina) is a variant of k -means clustering algorithm and is designed to handle very large (disk-resident) data sets.

BFR assumes -

- Clusters are normally distributed around a centroid in a Euclidean space.
- Clusters are axis-aligned ellipses



AML Cheat Sheet



Exploratory Data Analysis – Lab Session

- Introduction to AML
- Picking one scenario to start with
- Classification Model
- Regression Model
- Clustering(Separate data)



Azure HackFest Training

Overall Machine Learning Terminology

- **Examples:** Items or instances used for learning or evaluation.
- **Features:** Set of attributes represented as a vector associated with an example.
- **Labels:** Values or categories assigned to examples. In classification the labels are categories; in regression the labels are real numbers.
- **Target:** The correct label for a training example. This is extra data that is needed for supervised learning.
- **Output:** Prediction label from input set of features using a model of the machine learning algorithm.
- **Training sample:** Examples used to train a machine learning algorithm.
- **Validation sample:** Examples used to tune parameters of a learning algorithm.
- **Model:** Information that the machine learning algorithm stores after training. The model is used when predicting the output labels of new, unseen examples.
- **Test sample:** Examples used to evaluate the performance of a learning algorithm. The test sample is separate from the training and validation data and is not made available in the learning stage.
- **Loss function:** A function that measures the difference/loss between a predicted label and a true label. We will design the learning algorithms so that they minimize the error (cumulative loss across all training examples).
- **Hypothesis set:** A set of functions mapping features (feature vectors) to the set of labels. The learning algorithm chooses one function among those in the hypothesis set to return after training. Usually we pick a class of functions (e.g., linear functions) parameterized by a set of free parameters (e.g., coefficients of the linear function) and pinpoint the final hypothesis by identifying the parameters that minimize the error.
- **Model selection:** Process for selecting the free parameters of the algorithm (actually of the function in the hypothesis set).

42

Neal Analytics Oil & Gas Predictive Maintenance

Neal Analytics is actively working with 25 upstream oil and gas companies to reduce production costs with cloud-based predictive maintenance models.

About Neal Analytics



About

- Founded in 2011
- Offices in Seattle and Mumbai, India
- OSIsoft & Microsoft Gold Partner

Capabilities

- 25 data scientists on staff
- Predictive Analytics, data visualization and data engineering
- Oil and gas industry experience

Technologies

- Microsoft Azure, Azure ML, SQL Server and PowerBI
- R, SAS, SPSS, and Weka

Approach

- Demonstrate capabilities with Proof-of-Concept (POC)
- Demonstrate value with Proof-of-Profit (POP) methodology

Service Offerings



Cost Management

Total cost of production financial models for upstream oil and gas

Automation Experience

Joint product development with leading oil and gas automation firm

Predictive Maintenance

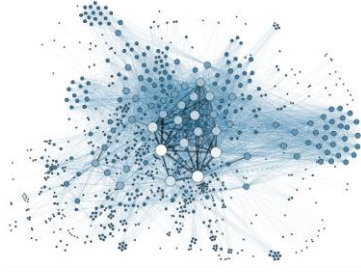
Optimize maintenance schedules by extending SCADA systems with predictive analytics

Optimization

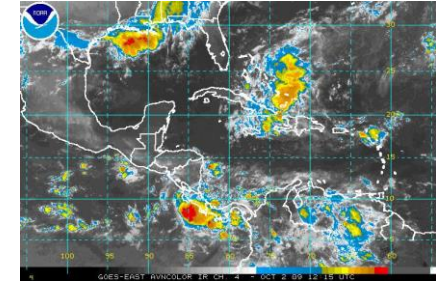
Truck and pumper optimization experience

Data Science Approach

Classification



Regression



Build end-to-end model in AML

Load/Cleanse Data

- Transfer data to Azure based data storage (Azure SQL, Blob Storage)
- Create associations between data tables
- Import data into Azure ML using Reader Module

Transform Data

- Data cleaning by doing audit checks on data looking at min, mean and max.
- Complete or remove incomplete records
- Variable treatment by handling missing, outliers or special values
- Create new derived measures in line with business sense

Exploratory Analysis

- Generate exploratory visualizations in R to identify relationships between variables
- Determine correlation and significance of predictive variables
- Graph temporal relationships on varying time scales (hours, days)

Build & Validate Model

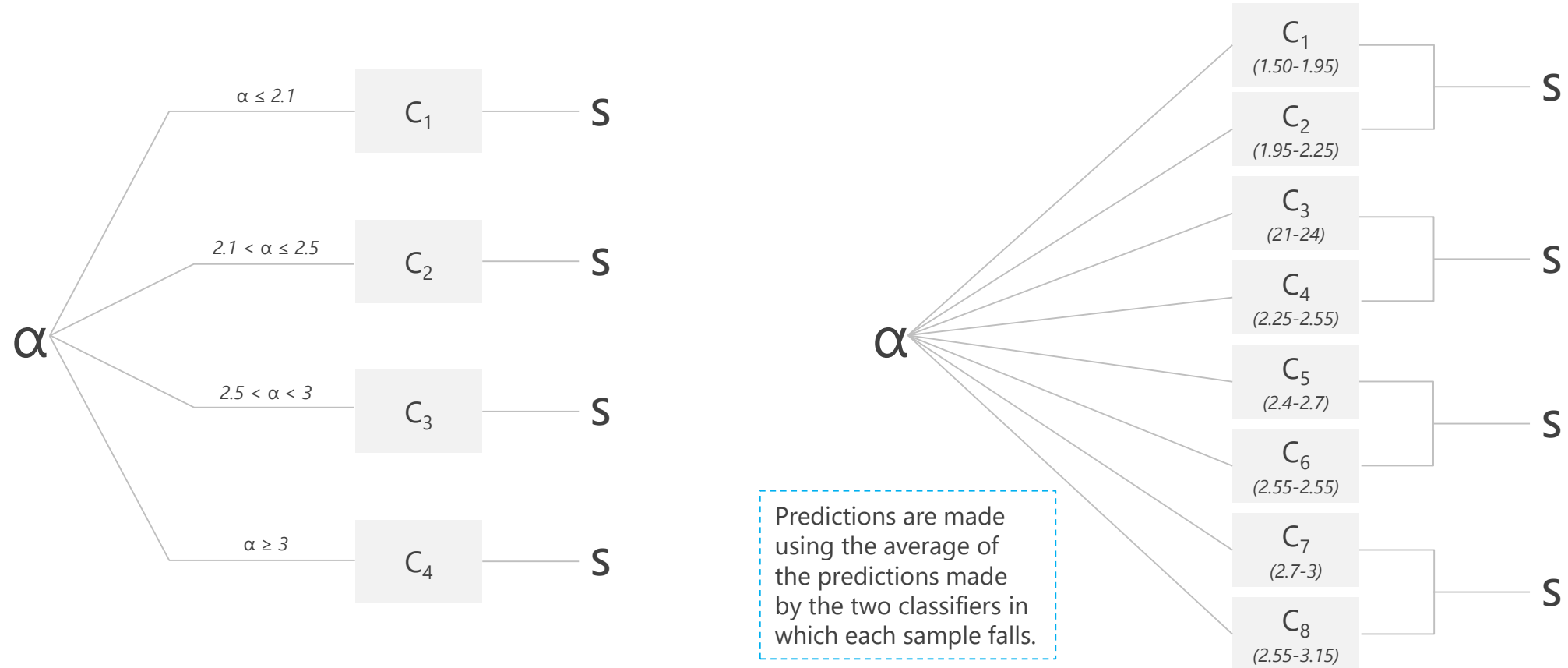
- Split data into training and scoring data sets
- Do iterative process of feature selection
- Train Azure ML based predictive model using subset of data
- Validate model using either evaluate model or cross validate model to test stability of predictive features in Azure ML

Publish Model

- Save trained model for scoring
- Input new data set to trained model
- Analyze model performance
- Benchmark performance of several algorithm variations
- Publish model as API service

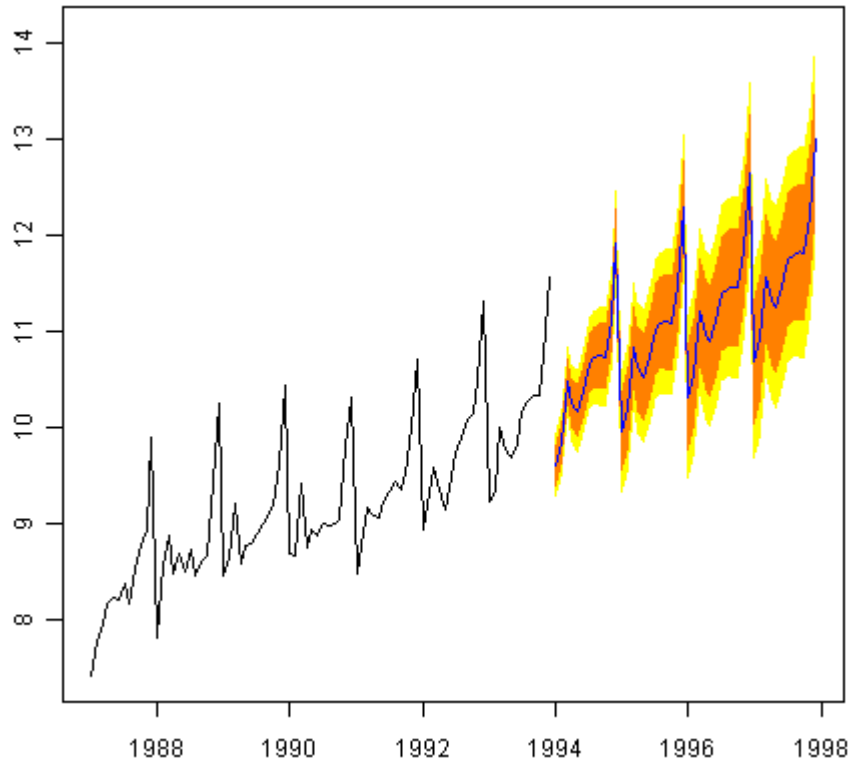
Advanced Topic: Combining Models – ensemble.

An alternative of model combinations is to select one of the models to make the classification, and let the choice of the model be driven by an input parameter or by an a priori knowledge.



Time Series Analysis - ARIMA

ARIMA

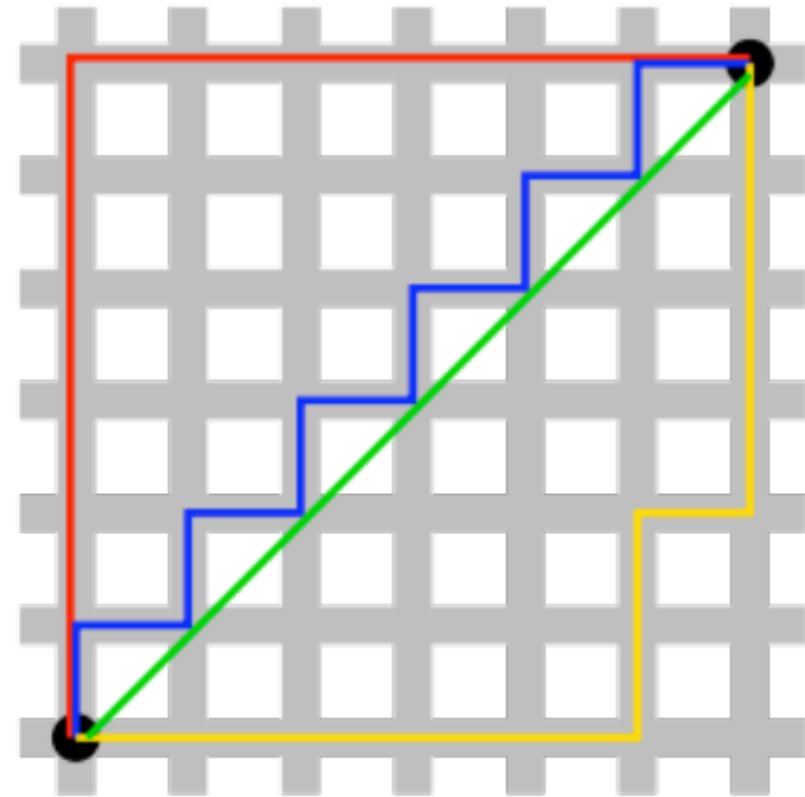


- In particular in time series analysis, an autoregressive integrated moving average (ARIMA) model is a generalization of an autoregressive moving average (ARMA) model
- Both of these models are fitted to time series data either to better understand the data or to predict future points in the series
- What does ARIMA mean:
 - The AR part of ARIMA indicates that the evolving variable of interest is regressed on its own lagged (i.e., prior) values
 - The MA part indicates that the regression error is actually a linear combination of error terms whose values occurred contemporaneously and at various times in the past
 - The I (for "integrated") indicates that the data values have been replaced with the difference between their values and the previous values

k-Means Clustering Algorithm Details

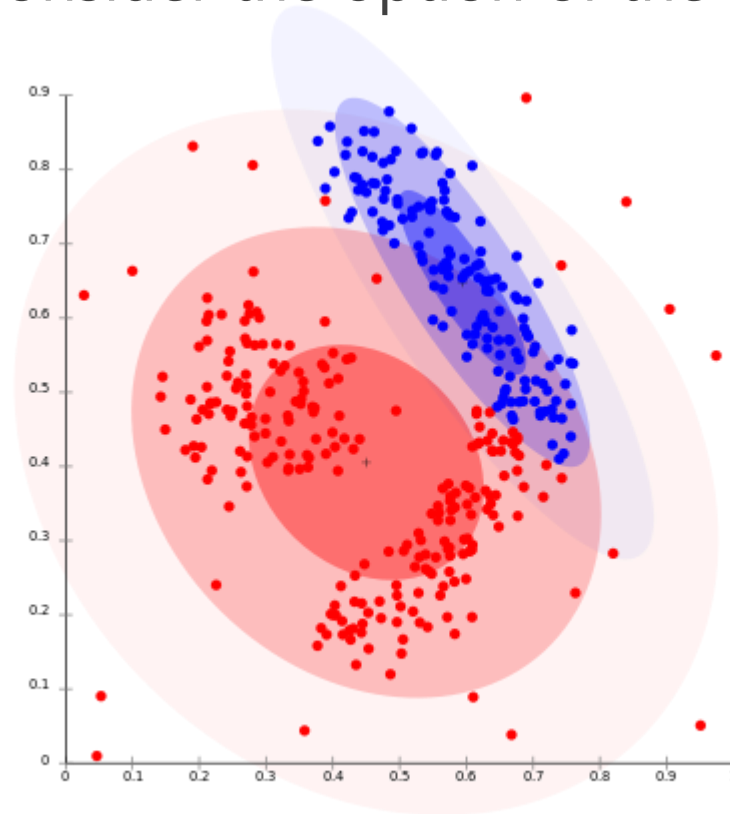
- Determine the similarity between two clusters based on a defined metric

Names	Formula
Euclidean distance	$\ a - b\ _2 = \sqrt{\sum_i (a_i - b_i)^2}$
squared Euclidean distance	$\ a - b\ _2^2 = \sum_i (a_i - b_i)^2$
Manhattan distance	$\ a - b\ _1 = \sum_i a_i - b_i $
maximum distance	$\ a - b\ _\infty = \max_i a_i - b_i $
Mahalanobis distance	$\sqrt{(a - b)^\top S^{-1} (a - b)}$ where S is the covariance matrix
cosine similarity	$\frac{a \cdot b}{\ a\ \ b\ }$



Optimizing K-Means Performance

- Find the right balance between two competing levers: number of clusters (K) and variance of clusters seen in different model runs
- The number of clusters is inversely related to the variance. The larger the number of clusters, the more stable the result! Consider the option of the trivial case $K = n$, and variance = 0.
- Criteria for performance:
 - BIC – Bayesian Information Criteria
 - AIC – Akaike Information Criteria
 - Davis- Bouldin Index



A Quick Note On Normalization

- In the simplest cases, normalization of ratings means adjusting values measured on different scales to a notionally common scale, often prior to averaging
- The purpose of normalization is to allow the comparison of corresponding normalized values for different datasets in a way that eliminates the effects of certain gross influences
- Some algorithms like decision trees can still perform well data of various types and scales, while others struggle (neural networks).



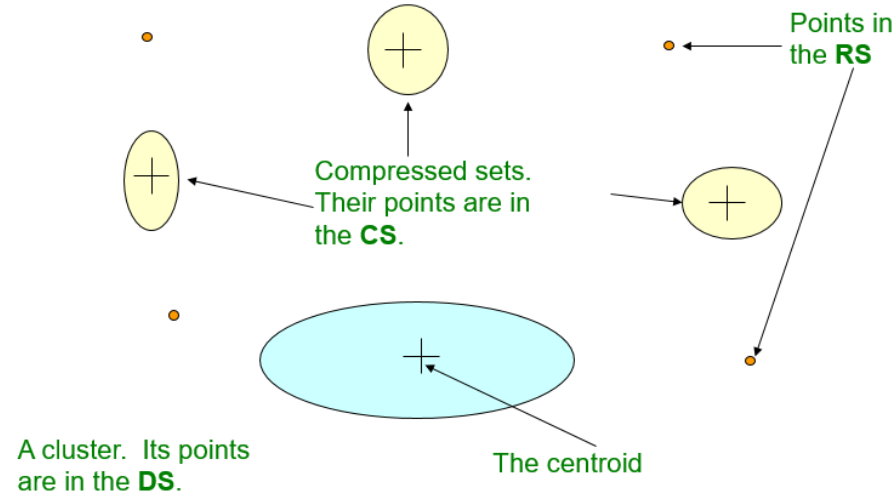
Var	$x' = \frac{x - \bar{x}}{\sigma_x}$
Range [0,1]	$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$
Log	$x' = \ln(x - \min(x) + 1)$
Softmax	$\hat{x} = \frac{x - \bar{x}}{\sigma_x} \quad x' = \frac{1}{1 + e^{-\hat{x}}}$

BFR- 3 Classes of Points

Points are read one main-memory-full at a time.

Discard Set(DS)

Points close enough to a centroid to be summarized



Retained Set(RS)

Isolated points waiting to be assigned to a compression set

Compression Set (CS)

Groups of points that are close together but not close to any existing centroid
These points are Isolated points waiting to be assigned to a compression set

Euclidean Distance illustration

