Unit 3
Classification and Regression

Classification Vs Regression

- Classification
  - predicts categorical class labels (discrete or nominal)
  - classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data
- Regression
  - models continuous-valued functions, i.e., predicts unknown values

- Applications
  - Credit/loan approval:
  - Medical diagnosis: if a tumor is cancerous or benign
  - Fraud detection: if a transaction is fraudulent
  - Web page categorization: which category it is
  - Spam Classification
  - Weather Diagnosis /Prediction

Regression

Regression Analysis

- Statistical Technique which explains association and causation
- Correlation Analysis – association between two sets of quantitative data
- Regression analysis – variation in one variable based on the variation of one or more variables
- The variable of which variation is explained – dependent variable
- The variables which are used to explain the variation is called independent variables

Simple and Multiple Regression

- Only one independent and only one dependent variable used to explain the variations, then the model is simple regression
- If multiple independent variables used to explain one single dependent variable – Multiple regression
- Applications: Forecasting, relationship between various features
- Methodology
  - General Regression Model
    - $Y=a+b_1x_1+b_2x_2+b_3x_3+....+b_nx_n$

    Y – dependent variable

    $x_1 \ldots x_n$ – independent variable

    $b_1 \ldots b_n$- coefficients of each independent variable

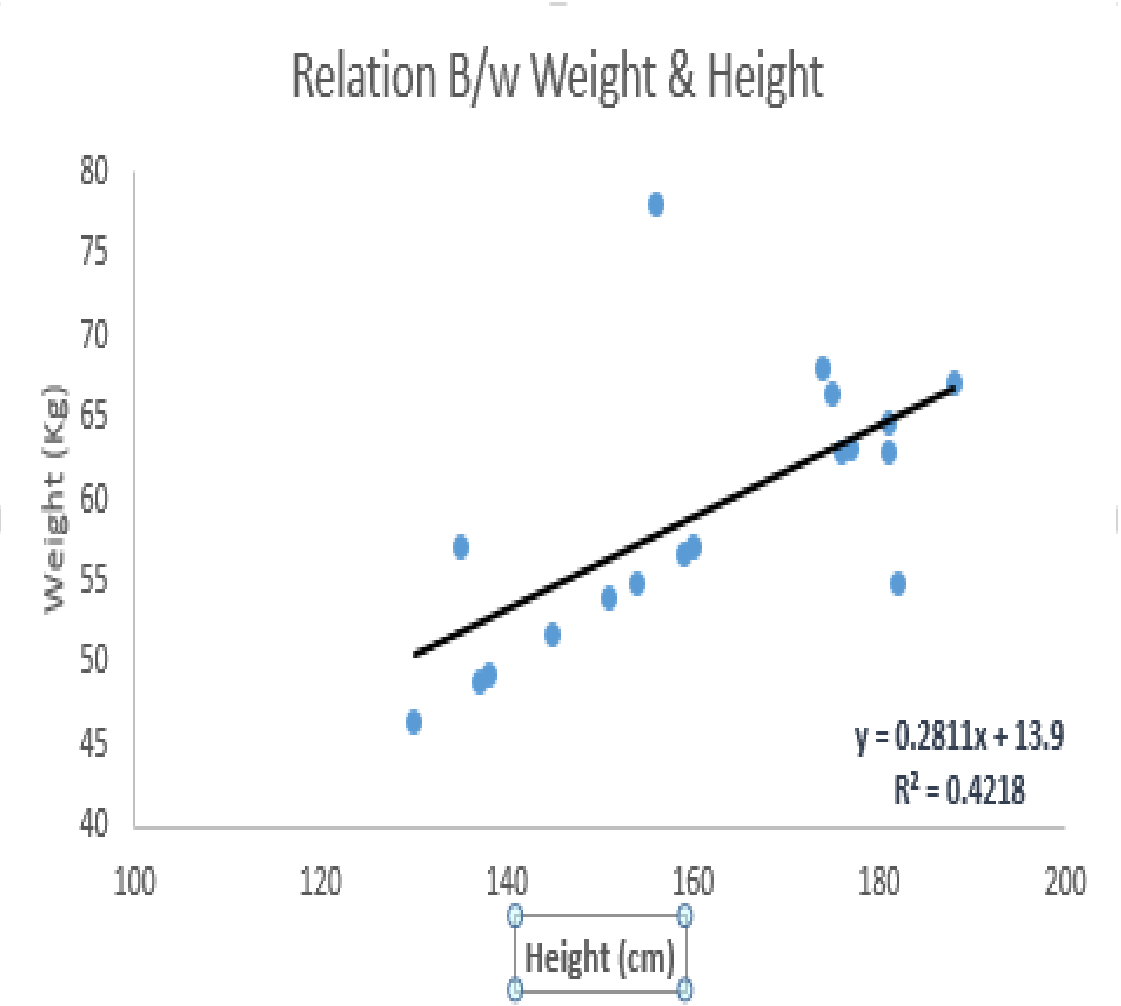
    a- intercept

- Input: Input data on Y and each of the X variable
- Output: b-coefficients of all variables

  t- test for significance of each variable

  F-test for the model as the whole

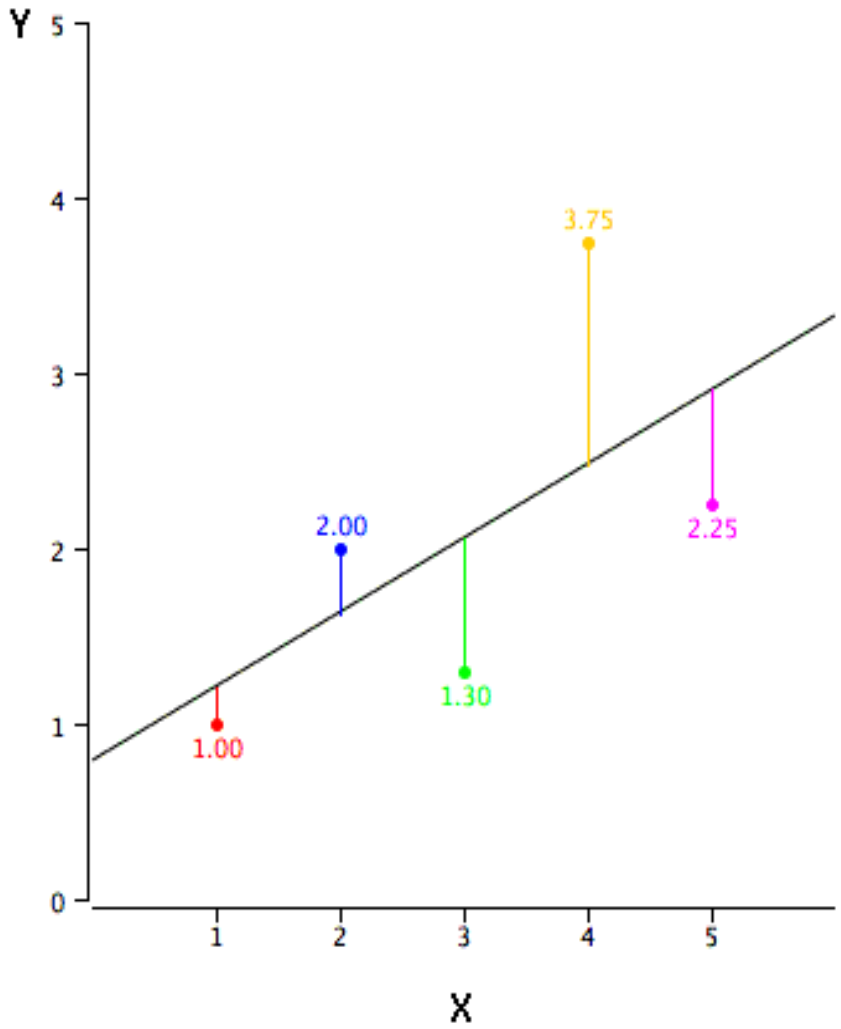
Simple Linear Regression



Line of best fit – Regression line

- Find the mean of x and mean of y.
- Any line we decide, should pass through this coordinate
- Draw all possible lines
- Notice for every point on the graph, our line is wrong by some distance
- **This amount wrong is called a residual**
- Difference between the observed value of the dependent variable (y) and the predicted value ($\hat{y}$) is called the **residual** (e).
- Each data point has one **residual**.
- **Residual** = Observed value - Predicted value.
  - $e = y - \hat{y}$

Residuals and OLS



Line of Best Fit – Regression line

- Both the sum and the mean of the **residuals** are equal to zero.
- Find the squares and sum it, it is sum of squared residuals
- Find the least squares to fit the regression line
- **Ordinary Least squares (OLS)** is a statistical **method** used to determine a line of best fit by minimizing the sum of **squares** created by a mathematical function. A "**square**" is determined by squaring the distance between a data point and the regression line.
- **Line of Best Fit = Lowest SS residuals**

Regression diagnostics

- Two important metrics to measure Regression
  - Root mean square error -> RMSE – measures model prediction error[it is always better to have less rmse]

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

  - R-Square – squared correlation between the observed and predicted [high r-square is prefered]
  - It is the proportion of the variance of the dependent variable explained by the regression model

Coefficient of Determination → $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

Sum of Squares Total → $SST = \sum (y - \bar{y})^2$

Sum of Squares Regression → $SSR = \sum (y' - \bar{y}')^2$

Sum of Squares Error → $SSE = \sum (y - y')^2$

- Adjusted R2
  - The **adjusted R-squared** is a modified version of **R-squared** that has been **adjusted** for the number of predictors in the model.
  - The **adjusted R-squared** increases only if the new term improves the model more than would be expected by chance.
  - It decreases when a predictor improves the model by less than expected by chance.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

$$\text{Adj } R^2 = 1 - (1 - R^2)\frac{n - 1}{n - p - 1}$$

p - number of regressors
n - sample size

Interpretation of Linear Regression

STATISTIC	CRITERION
R-Squared	Higher the better *(> 0.70)*
Adj R-Squared	Higher the better
F-Statistic	Higher the better
Std. Error	Closer to zero the better
t-statistic	Should be greater 1.96 for p-value to be less than 0.05
MAPE (Mean absolute percentage error)	Lower the better
MSE (Mean squared error)	Lower the better
Min_Max Accuracy => mean(min(actual, predicted)/max(actual, predicted))	Higher the better

Logistic Regression

- Logistic Regression extends the idea of multiple linear regression, where the dependent variable Y is binary
- Input $x_1 \ldots x_n$ may be nominal or continuous or mixture of two types
- This model better suits only under these conditions
- Some Drawbacks of multiple linear regression
  - Predicted model fits outside the range of 0 and 1
  - The dependent variable is not normally distributed.
- Logistic regression overcome the above difficulties

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.
- Let us take an example of

n=165	**Predicted: NO**	**Predicted: YES**
Actual: NO	50	10
Actual: YES	5	100

n=165	**Predicted: NO**	**Predicted: YES**	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

- **Accuracy:** Overall, how often is the classifier correct?
  - (TP+TN)/total = (100+50)/165 = 0.91
- **Misclassification Rate:** Overall, how often is it wrong?
  - (FP+FN)/total = (10+5)/165 = 0.09
  - equivalent to 1 minus Accuracy
  - also known as "Error Rate"
- **True Positive Rate(Recall/Sensitivity):** When it's actually yes, how often does it predict yes?
  - TP/actual yes = 100/105 = 0.95
- **Precision:** When it predicts yes, how often is it correct?
  - TP/predicted yes = 100/110 = 0.91

Classification

Classifiers

- A Classifier is a linguistic symbol that represents a class or group of objects or subjects.
- Some of the classifiers are
  - Naïve Bayes
  - Decision Tree Classifier
  - kNN classifier
  - SVM

Classification – A Two-step Process

- Model construction: describing a set of predetermined classes
  - Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute
  - The set of tuples used for model construction is training set
  - The model is represented as classification rules, decision trees, or mathematical formulae

Classification – A Two-step Process

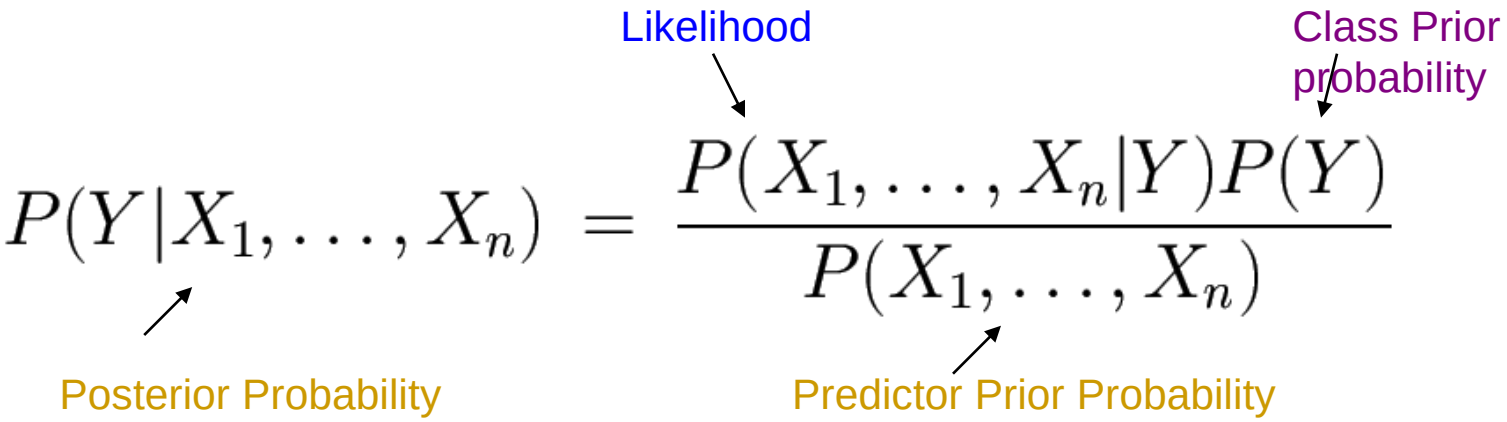
- Model usage: for classifying future or unknown objects
  - Estimate accuracy of the model
    - The known label of test sample is compared with the classified result from the model
    - Accuracy rate is the percentage of test set samples that are correctly classified by the model
    - Test set is independent of training set
    - If the accuracy is acceptable, use the model to classify new data
- Note: If *the test set* is used to select models, it is called validation (test) set

Naïve Bayes Classification

- Problem statement:
  - Given features $X_1, X_2, \ldots, X_n$
  - Predict a label Y
- Foundation: Bayes Theorem
- Know these terms:
- Prior Probability - the prior is what you believe about some quantity at particular point in time
- Posterior Probability - A posterior probability is your belief once additional information comes in.

Naïve Bayes Classification

- Bayes Rule



Also Represented as follows,

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

$$Pr(Y|X) = Pr(X_1|Y)* Pr(X_2|Y)*Pr(X_3|Y)*P(Y)$$

Assuming X_1, X_2, X_3 are independent

- Posterior probability = conditional probability of features over Predict label

  *

  individual probability of Predictor label

  ---

  individual probability of features label

Example

COLOR	TYPE	ORIGIN	STOLEN
RED	SPORTS	DOMESTIC	YES
RED	SPORTS	DOMESTIC	NO
RED	SPORTS	DOMESTIC	YES
YELLOW	SPORTS	DOMESTIC	NO
YELLOW	SPORTS	IMPORTED	YES
YELLOW	SUV	IMPORTED	NO
YELLOW	SUV	IMPORTED	YES
YELLOW	SUV	DOMESTIC	NO
RED	SUV	IMPORTED	NO
RED	SPORTS	IMPORTED	YES

P(Stolen=Yes) = 5/10
P(Stolen=No) = 5/10

	YES	NO
RED	3 /5	2/5
YELLOW	2/5	3/5

	YES	NO
SPORTS	4/5	2/5
SUV	1/5	3/5

	YES	NO
DOMESTIC	2/5	3/5
IMPORTED	3/5	2/5

- Unseen instance

X=(RED,SUV,DOMESTIC)

P(X|YES) =P(RED|YES)*P(SUV|YES)*P(DOMESTIC|YES) P(YES)

= 3/5*1/5*2/5 = 0.024 * 0.5 =**0.012**

P(X|NO)=P(RED|NO)*P(SUV|NO)*P(DOMESTIC|NO) P(NO)

=2/5*3/5*3/5 = 0.072 * 0.5 = **0.036**

0.036 > 0.012 , Hence unseen example is classified as **NO**

Example

outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes

outlook	temp.	humidity	windy	play
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Frequencies and Probabilities

Frequencies and probabilities for the weather data:

outlook			temperature			humidity			windy			play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2	3	hot	2	2	high	3	4	false	6	2	9	5
overcast	4	0	mild	4	2	normal	6	1	true	3	3		
rainy	3	2	cool	3	1								
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2/9	3/5	hot	2/9	2/5	high	3/9	4/5	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	mild	4/9	2/5	normal	6/9	1/5	true	3/9	3/5		
rainy	3/9	2/5	cool	3/9	1/5								

Classifying an Unseen Observation

Now assume that we have to classify the following new instance:

outlook	temp.	humidity	windy	play
sunny	cool	high	true	?

- *Key idea:* compute a probability for each class based on the probability distribution in the training data.
- First take into account the probability of each attribute. Treat all attributes **equally important**, i.e., multiply the probabilities:

$$P \text{ yes } 2/9 * 3/9 * 3/9 * 3/9 = 0.0082$$

$$P \text{ no } 3/5 * 1/5 * 4/5 * 3/5 = 0.0577$$

- Now take into account the **overall probability** of a given class.
- Multiply it with the probabilities of the attributes:
- P yes $=0$ $.0082 * 9/14 = 0.0053$
- P no $= 0.0577 * 5/14 = 0.0206$
- Now choose the class so that it **maximizes** this probability. This means that the **new instance will be classified as No**.

NB - Advantages

- It is easy and fast to predict class of test data set. It also perform well in multi class prediction
- When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you need less training data.
- It perform well in case of categorical input variables compared to numerical variable(s). For numerical variable, normal distribution is assumed (bell curve, which is a strong assumption).

NB - Disadvantages

- If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as "Zero Frequency". To solve this, we can use the smoothing technique. One of the simplest smoothing techniques is called Laplace estimation.
- On the other side naive Bayes is also known as a bad estimator, so the probability outputs from predict_proba are not to be taken too seriously.
- Another limitation of Naive Bayes is the assumption of independent predictors. In real life, it is almost impossible that we get a set of predictors which are completely independent.

Problem of Zero frequency

- If a particular attribute value does not occur in the training set in conjunction with every class value , then things go badly
- **Laplace Estimation/ smoothing:** Add 1 to all the numerator and compensate by adding the factor times to the denominator.

outlook			temperature		
	yes	no		yes	no
sunny	2	3	hot	2	2
overcast	4	0	mild	4	2
rainy	3	2	cool	3	1
	yes	no		yes	no
sunny	2/9	3/5	hot	2/9	2/5
overcast	4/9	0/5	mild	4/9	2/5
rainy	3/9	2/5	cool	3/9	1/5

	yes	no
sunny	2/9	4/8
overcast	4/9	1/8
rainy	3/9	3/8

Model Measurement

- Accuracy – (TP+FN)/(TP+TN+FP+FN)
- FPR - Precision - TP/TP+FP
- TPR or Recall - TP/TP+TN
- AUC - ROC curve
  - AUC (**Area Under The Curve**)
  - ROC (**Receiver Operating Characteristics**)
  - The ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis.
  - ROC value should be higher than 50%. Higher the ROC better the model. When it less than or equal to 50%, model is worse.

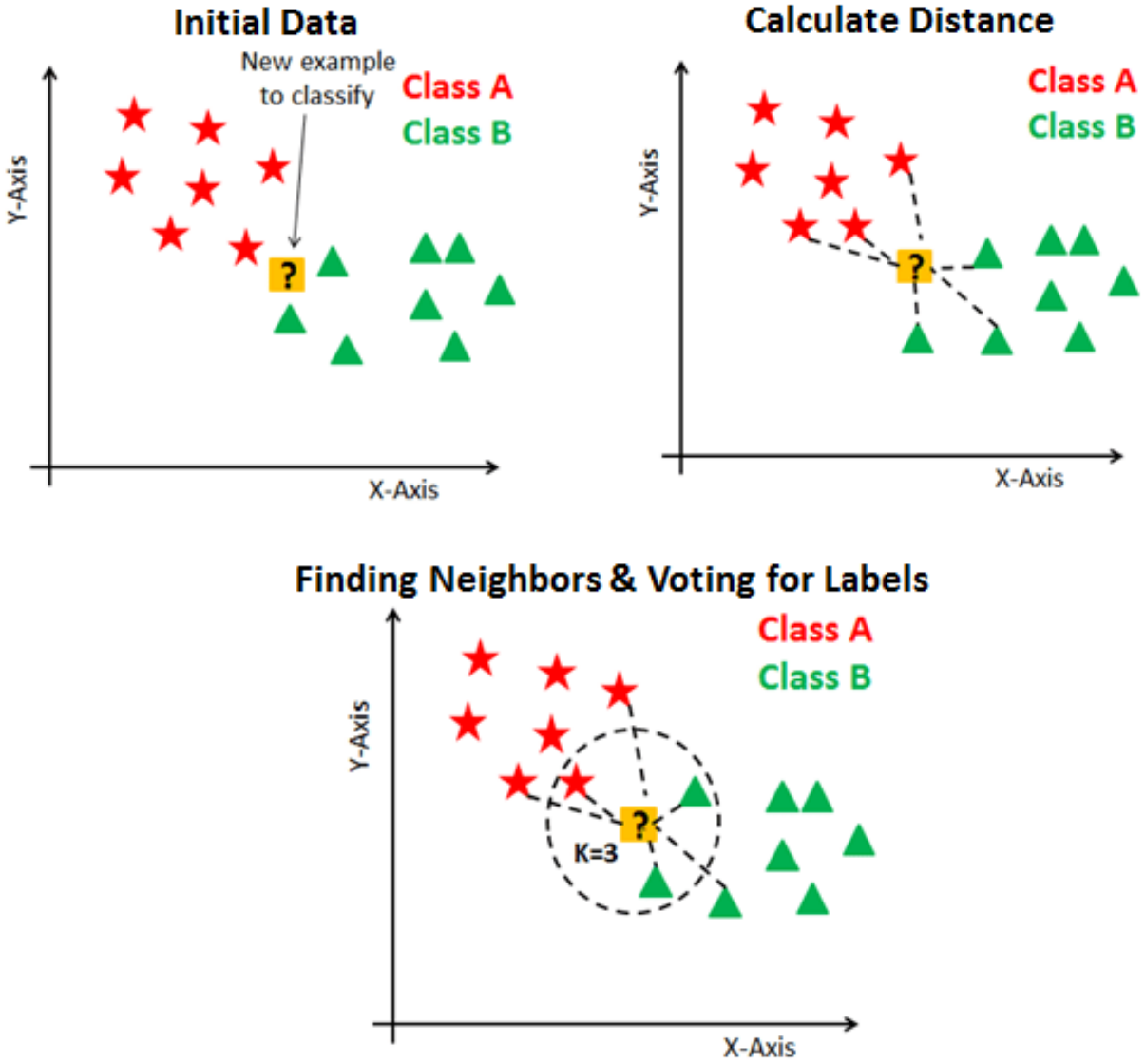
K Nearest Neighbour Classification (kNN)

- K Nearest Neighbour(KNN) is a very simple, easy to understand, versatile and one of the topmost machine learning algorithms.
- KNN used in the variety of applications such as finance, healthcare, political science, handwriting detection, image recognition and video recognition.
- KNN is a non-parametric and lazy learning algorithm.
- Non-parametric means there is no assumption for underlying data distribution.
- Lazy algorithm means it does not need any training data points for model generation.

- All training data used in the testing phase.
- This makes training faster and testing phase slower and costlier.
- Costly testing phase means time and memory.
- In KNN, K is the number of nearest neighbours. The number of neighbours is the core deciding factor.
- K is generally an odd number if the number of classes is 2.
- Working of Algorithm
  - Calculate distance (Euclidean, Manhattan, Chebychev, Hamming…)

$$d(x, x') = \sqrt{(x_1 - x'_1)^2 + (x_2 - x'_2)^2 + \ldots + (x_n - x'_n)^2}$$

  - Find closest neighbours
  - Vote for labels


Initial Data
New example
to classify
Class A
Class B
Y-Axis
X-Axis
?
Calculate Distance
Class A
Class B
Y-Axis
X-Axis
?
Finding Neighbors & Voting for Labels
Class A
Class B
Y-Axis
X-Axis
?
K=3

Example

Name	Acid Durability	Strength	class
Type-1	7	7	Bad
Type-2	7	4	Bad
Type-3	3	4	Good
Type-4	1	4	Good
Type-5	3	7	??

- Find the distance measure
- Euclidean distance $d(p,q)=sqrt((q1-p1)^2 +(q2-p2)^2+\ldots)$

Name	Acid Durability	Strength	class	Distance	Rank
Type-1	7	7	Bad	$Sqrt(7-3)^2+(7-7)^2=4$	3
Type-2	7	4	Bad	5	4
Type-3	3	4	Good	3	1
Type-4	1	4	Good	3.6	2
Type-5	3	7	??		

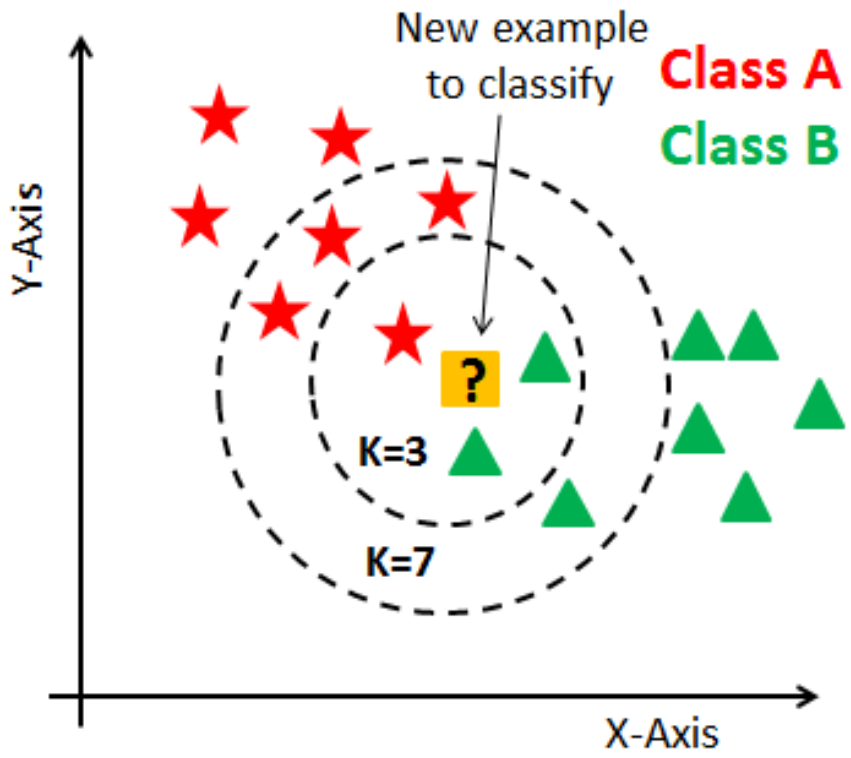
- If k=1 =>The new test data is close to Type-3 and hence class label is good.
- If k=3=>The data that is close to test data is Type-3, Type-4 and Type-1 where 2 are good and 1 is bad. The probability of good is high, hence the new test data belongs to class label good.

kNN for Large Datasets???

- **Curse of Dimensionality**
  - KNN performs better with a lower number of features
  - Increase in dimension also leads to the problem of overfitting.
  - To avoid the curse of dimensionality, PCA can be used for dimensionality reduction and then apply kNN.

Choose optimal k

- **How do you decide the number of neighbours in KNN?**
  - No optimal number of neighbours suits all kind of data sets. Each dataset has it's own requirements.
  - In the case of a small number of neighbours, the noise will have a higher influence on the result, and a large number of neighbours make it computationally expensive.


New example
to classify
Class A
Class B
Y-Axis
?
K=3
K=7
X-Axis

Evolutionary Computing and Genetic Algorithms

- Evolution theory Principles
  - Evolution takes place at the level of chromosomes
  - Nature tends to make more copies of chromosomes which produce a more "fit" organism
  - Diversity must be maintained in the population
- Operation is called crossover

Genetic Algorithm

- A genetic algorithm is a search heuristic that mimics the process of evolution
- There are five phases
  - Initial population
  - Fitness function
  - Selection
  - Crossover
  - Mutation
- The primary advantage comes from the crossover operation

- Step1: Initialize: create a population of N elements ,each with randomly generated DNA
- Step2: Selection: Evaluate the fitness of each element of the population and build a mating pool
- Step 3: Reproduction: Repeat N times:
  - Pick two parents with probability according to fitness function
  - Crossover: create a child by combining the DNA of the two parents
  - Mutation: mutate the child's DNA based on the given probability
  - Add the new child to the new population
- Replace the old population with new population and return to step 2.

- Applications
  - Text mining
  - Web mining
  - Email classification
  - Natural language processing
  - Pattern recognition