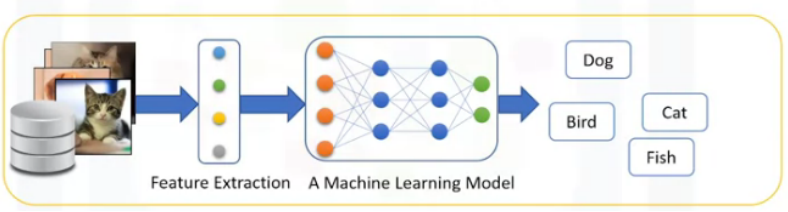
- MODULE #1: Introduction to Machine Learning
- MODULE #2: Regression
- MODULE #3: Classification
- MODULE #4: Linear Classification
- MODULE #5: Clustering

❖ **MODULE #1: Introduction to Machine Learning**

Machine Learning Algorithms

- Regression
  - Simple linear regression
  - Multiple linear regression
  - Regression trees
- Classification
  - Logistic regression
  - KNN
  - SVM
  - Multiclass prediction
  - Decision trees
- Clustering
  - K-means

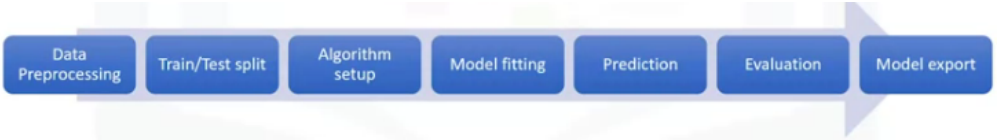
Introduction to Machine Learning

- What is machine learning?
  - Machine learning is the subfield of computer science that gives "computers the ability to learn without being explicitly programmed"
  - 

- Major machine learning techniques
  - Regression/Estimation
    - Predicting continuous values
  - Classification
    - Predicting the item/class category of a case
  - Clustering
    - Finding the structure of data; summarization
  - Associations
    - Associating frequent co-occurring items/events
  - Anomaly detection
    - Discovering abnormal and unusual cases
  - Sequence mining
    - Predicting next events; click-stream (Markov Model, HMM)
  - Dimension Reduction
    - Reducing the size of data (PCA)
  - Recommendation systems

        - Recommending items
- Difference between artificial intelligence, machine learning, and deep learning
    - AI components:
        - Computer Vision
        - Language Processing
        - Creativity
        - Etc.
    - Machine learning:
        - Classification
        - Clustering
        - Neural Network
        - Etc.
    - Revolution in ML:
        - Deep learning

Python for Machine Learning

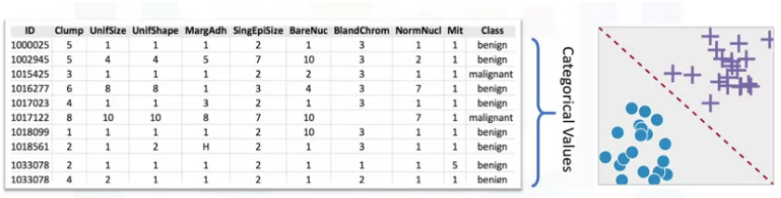
- Python libraries for machine learning
    - NumPy
    - SciPy
    - Matplotlib
    - Pandas
    - Scikit-learn
- More about scikit-learn
    - Free software machine learning library
    - Classification, Regression and Clustering algorithms
    - Works with NumPy and SciPy
    - Great documentation
    - Easy to implement
    - 

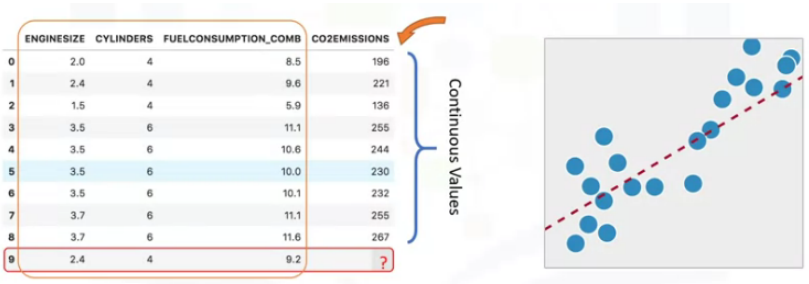
- Scikit-learn function
    - ```
      from sklearn import preprocessing
      X = preprocessing.StandardScaler().fit(X).transform(X)
      ```
      ```
      from sklearn.model_selection import train_test_split
      X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33)
      ```
      ```
      from sklearn import svm
      clf = svm.SVC(gamma=0.001, C=100.)
      ```
      ```
      clf.fit(X_train, y_train)
      ```
      ```
      clf.predict(X_test)
      ```
      ```
      from sklearn.metrics import confusion_matrix
      print(confusion_matrix(y_test, yhat, labels=[1,0]))
      ```
      ```
      import pickle
      s = pickle.dumps(clf)
      ```
      ```

Supervised vs Unsupervised

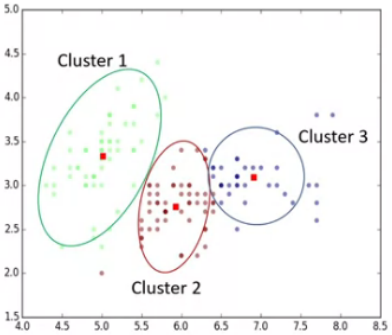
- What is supervised learning?
    - We “teach the model”, then with that knowledge it can predict unknown or future instances
- Teaching the model with labeled data
    - Columns and features
    - Value of the data, you can look at two kinds, numeric (numbers) and categorical (characters)
    -

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

- Types of supervised learning
    - Classification
    - Regression
- What is classification?
    - Classification is the process of predicting discrete class labels or categories
    - 

- What is regression?
    - Regression is the process of predicting continuous values
    - 

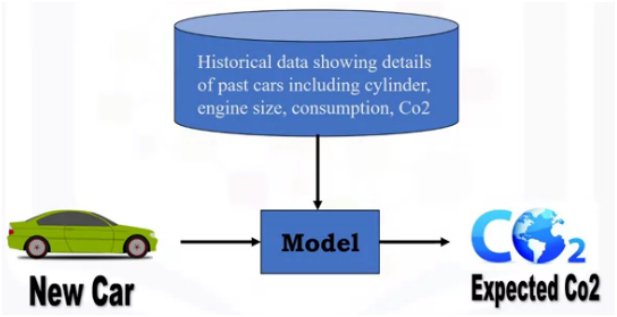
- What is unsupervised learning?
    - The model works on its own to discover information
        - The unsupervised algorithm trains on the dataset and draws conclusion from unlabeled data
    - Unsupervised learning techniques
        - Dimension reduction: reduces redundant features to make the

classification easier
        - Density estimation: very simple concept that is mostly used to explore the data to find some structure within it
        - Market basket analysis: modeling technique based upon the theory that if you buy a certain group of items, you're more likely to buy another group of items.
        - Clustering: one of the most popular unsupervised machine learning techniques used for grouping data points, or objects that are somehow similar
- What is clustering?
    - Clustering is grouping of data points or objects that are somehow similar by:
        - Discovering structure
        - Summarization
        - Anomaly detection
    - 

- Supervised vs unsupervised learning
    - Supervised learning
        - Classification: Classifies labeled data
        - Regression: Predicts trends using previous labeled data
        - Has more evaluation methods than unsupervised learning
        - Controlled environment
    - Unsupervised learning
        - Clustering: Finds patterns and groupings from unlabeled data
        - Has fewer evaluation methods than unsupervised learning
        - Less controlled environment

❖ **MODULE #2: Regression**

Introduction to Regression

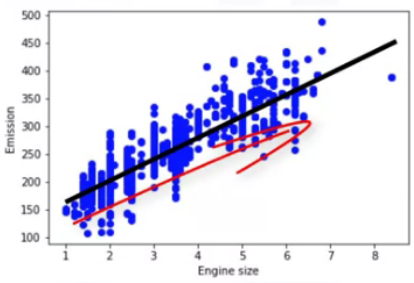
- What is regression?
    - Regression is the process of predicting a continuous value
        - Dependent variable (Y): state, target or final goal. Should be continuous, and cannot be a discrete value.
        - Independent variable (X): causes of those states
- What is a regression model?
    - 

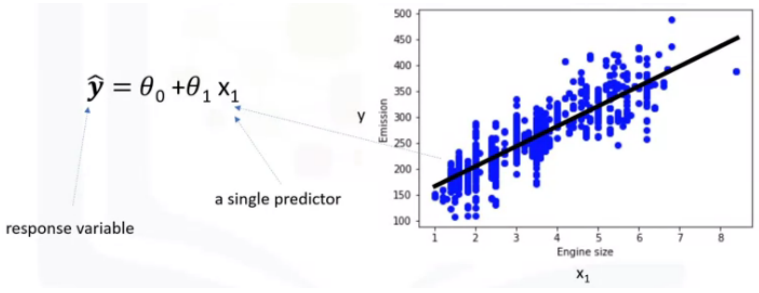
- Types of regression models
    - Simple Regression:
        - Simple Linear Regression
        - Simple Non-linear Regression
        - Ex: Predict co2emission vs EngineSize of all cars
    - Multiple Regression:
        - Multiple Linear Regression
        - Multiple Non-linear Regression
        - Ex: Predict co2emission vs EngineSize and Cylinders of all cars
- Applications of regression
    - Sales forecasting
    - Satisfaction analysis
    - Price estimation
    - Employment income
- Regression algorithms
    - Ordinal regression
    - Poisson regression
    - Fast forest quantile regression
    - Linear, Polynomial, Lasso, Stepwise, Ridge regression
    - Bayesian linear regression
    - Neural network regression
    - Decision forest regression
    - Boosted decision tree regression
    - KNN (K-nearest neighbors)

Simple Linear Regression

- Using linear regression to predict continuous variables
    - We can use linear regression to predict a continuous value such as Co2 emission by using other variables.
    - 

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

    - Simple Linear Regression there's two variables: Dependent variable (Y) and Independent variable (X)
- Linear regression topology
    - Simple Linear Regression
        - Predict co2emission vs EngineSize of all cars
            - Independent variable (X): EngineSize
            - Dependent variable (y): co2emission
    - Multiple Linear Regression:
        - Predict co2emission vs EngineSize and Cylinders of all cars
            - Independent variable (X): EngineSize, Cylinders, etc.
            - Dependent variable (y): co2emission
- How does linear regression work?
    - 

        - As the engine size increases, so do the emissions
- Linear regression model representation
    - 

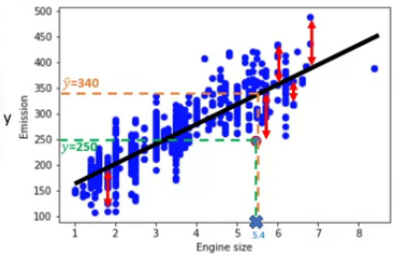
        - y hat is the dependent variable of the predicted value
        - x1 is the independent variable
        - theta0 and theta1 are the parameters of the line that we must adjust

    - Theta 1 is known as the slope or gradient of the fitting line and theta 0 is known as the intercept.
    - Theta 0 and theta 1 are also called the coefficients of the linear equation.
    - You can interpret this equation as y hat being a function of x1, or y hat being dependent of x1.
    - Linear regression estimates the coefficients of the line. This means we must calculate theta 0 and theta 1 to find the best line to fit the data. This line would best estimate the emission of the unknown data points.
- How to find the best fit?
    - $x_1 = 5.4$ independent variable
      $y = 250$ actual Co2 emission of x1

      $\hat{y} = \theta_0 + \theta_1 x_1$
      $\hat{y} = 340$ the predicted emission of x1

      Error = $y - \hat{y}$
      $= 250 - 340$
      $= -90$

      $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

- Estimating the parameters
	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267

      X_1 y

      $\hat{y} = \theta_0 + \theta_1 x_1$

      $\theta_1 = \frac{\sum_{i=1}^{s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{s}(x_i - \bar{x})^2}$

      $\bar{x} = (2.0 + 2.4 + 1.5 + \ldots)/9 = 3.03$

      $\bar{y} = (196 + 221 + 136 + \ldots)/9 = 226.22$

      $\theta_1 = \frac{(2.0 - 3.03)(196 - 226.22) + (2.4 - 3.03)(221 - 226.22) + \ldots}{(2.0 - 3.03)^2 + (2.4 - 3.03)^2 + \ldots}$

      $\theta_1 = 39$

      $\theta_0 = \bar{y} - \theta_1 \bar{x}$

      $\theta_0 = 226.22 - 39 * 3.03$

      $\theta_0 = 125.74$

      $\hat{y} = 125.74 + 39x_1$
- Predictions with linear regression
	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

      $\hat{y} = \theta_0 + \theta_1 x_1$

      $Co2Emission = \theta_0 + \theta_1\ EngineSize$

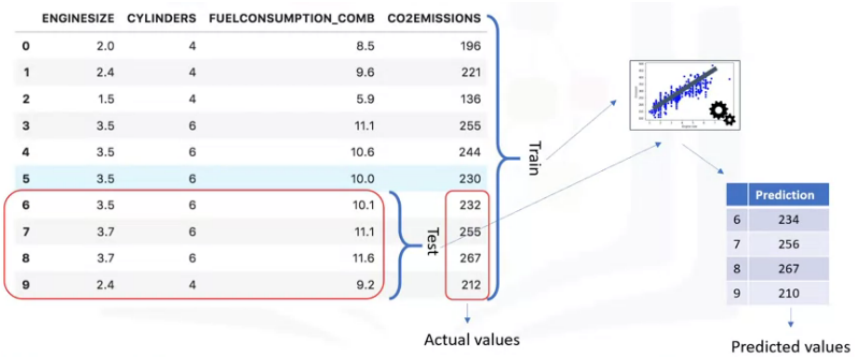
      $Co2Emission = 125 + 39\ EngineSize$

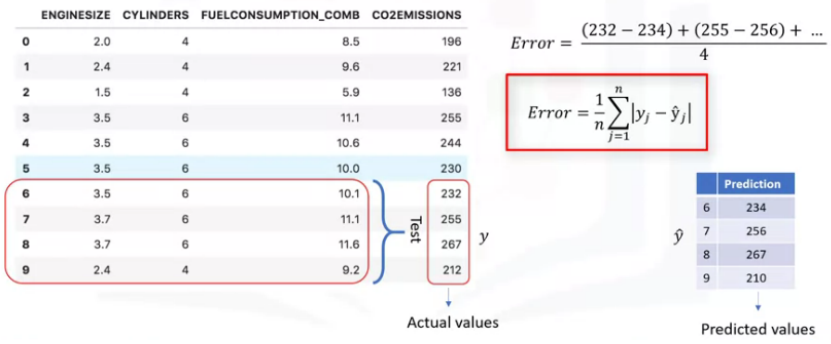
      $Co2Emission = 125 + 39 \times 2.4$

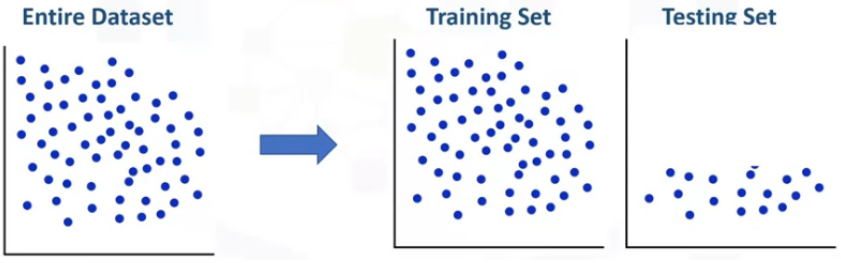
      $Co2Emission = 218.6$
- Pros of linear regression
    - Very fast
    - No parameter tuning

    - Easy to understand, and highly interpretable

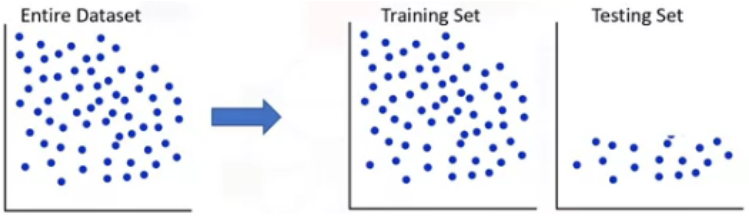
Model Evaluation in Regression Models

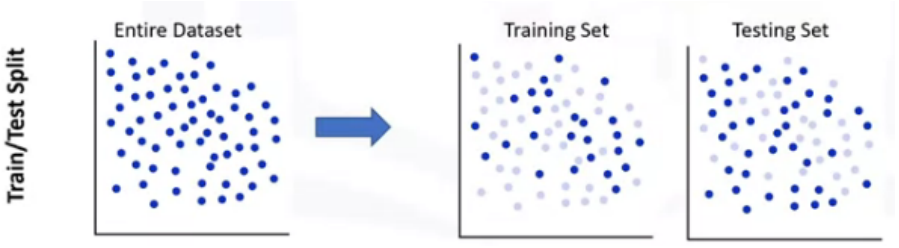
- Model evaluation approaches
    - Train and Test on the Same Dataset
    - Train/Test Split
    - Regression Evaluation Metrics
- Best approach for most accurate results?
    - One of the solutions is to select a portion of our dataset for testing.
    - 

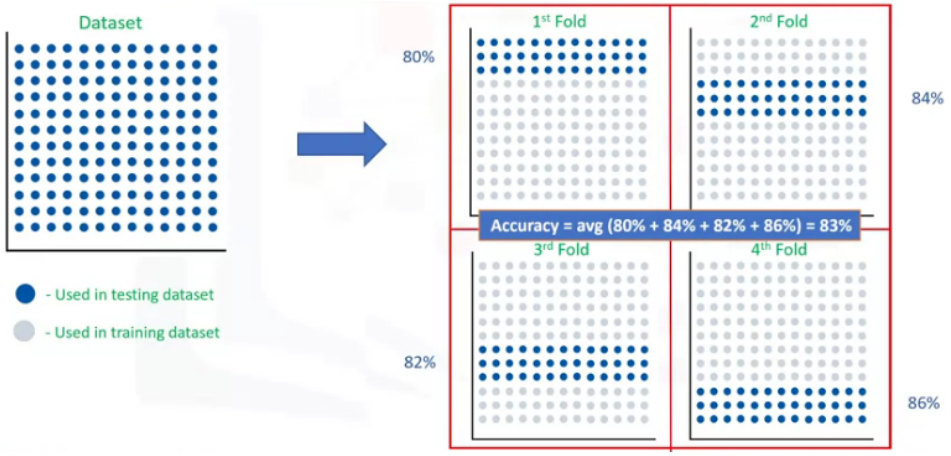
        - This indicates how accurate our model actually is
    - 

- Train and test on the same dataset
    - You train the model on the entire dataset, then you test it using a portion of the same dataset.
    - 

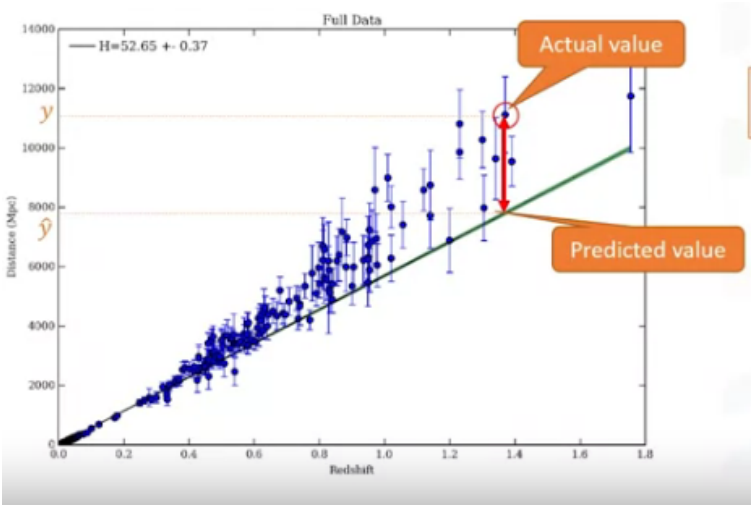
    - High “training accuracy”, Low “out-of-sample accuracy”
- What is training & out-of-sample accuracy?
    - Training Accuracy
        - High training accuracy isn’t necessarily a good thing
        - Result of over-fitting

        - Over-fit: the model is overly trained to the dataset, which may capture noise and produce a non-generalized model
    - Out-of-Sample Accuracy
        - It's important that our models have a high, out-of-sample accuracy
        - How can we improve out-of-sample accuracy?
- Train/Test split evaluation approach
    - Test on a portion of train set
        - Test-set is a portion of the train-set
        - High "training accuracy"
        - Low "out-of-sample accuracy"
        - 

    - Train/Test Split
        - Mutually exclusive
        - More accurate evaluation on out-of-sample accuracy
        - Highly dependent on which datasets the data is trained and tested
        - 

- How to use K-fold cross-validation
    - 

Evaluation Metrics in Regression Models

- Regression accuracy
    - We can compare the actual values and predicted values to calculate the accuracy of our regression model. Evaluation metrics provide a key role in the development of a model as it provides insight to areas that require improvement.
- What is an error of the model
    - Error: measure of how far the data is from the fitted regression line
    - 

    - Mean Absolute Error is the mean of the absolute value of the errors
    - Mean Squared Error is the mean of the squared error
    - Root Mean Squared Error is the root of the mean squared error
    - Relative Absolute Error takes the total absolute error and normalizes it
    - R^2 is not an error per say but is a popular metric for the accuracy of your model. It represents how close the data values are to the fitted regression line. The higher the R-squared, the better the model fits your data

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

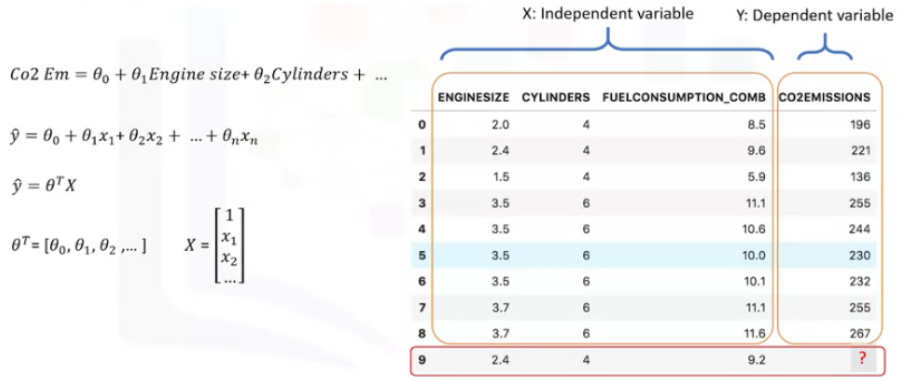
$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

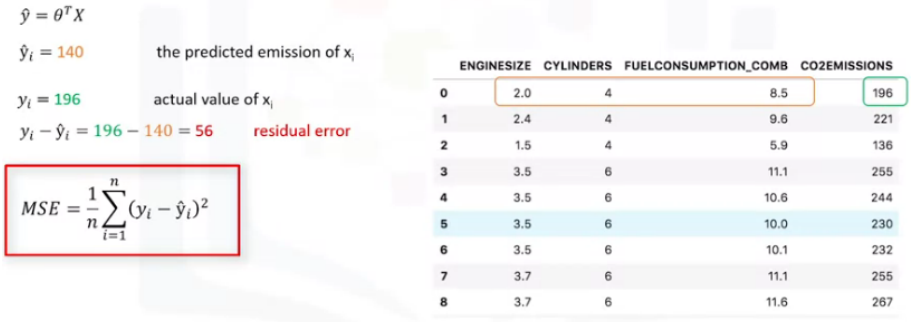
$$RAE = \frac{\sum_{j=1}^{n}|y_j - \hat{y}_j|}{\sum_{j=1}^{n}|y_j - \bar{y}|}$$

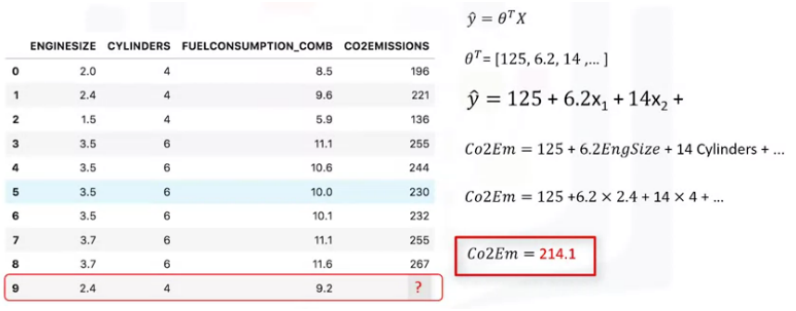
$$RSE = \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2}$$

- $R^2 = 1 - RSE$

Multiple Linear Regression

- Examples of multiple linear regression
  - Independent variables effectiveness on prediction
    - Does revision time, test anxiety, lecture attendance and gender have any effect on the exam performance of students?
  - Predicting impacts of changes
    - How much does blood pressure go up (or down) for every unit increase (or decrease) in the BMI of a patient?
- Predicting continuous values with multiple linear regression
  - 

  - 

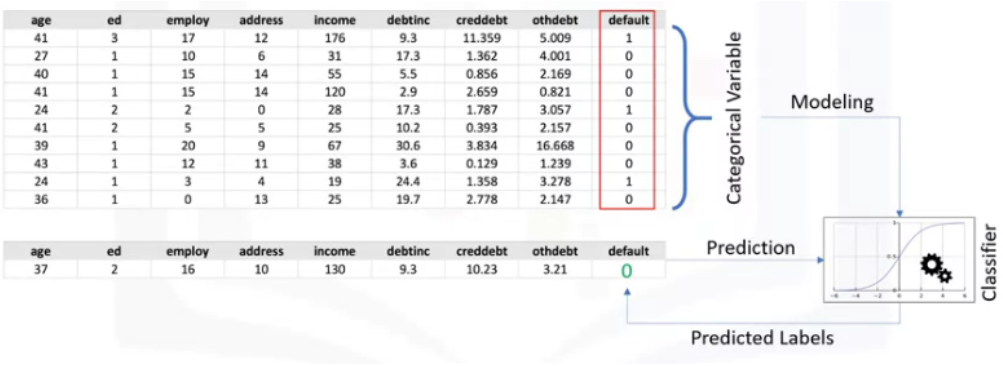
- Estimating multiple linear regression parameters
  - How to estimate Theta?
    - Ordinary Least Squares
      - Linear algebra operations
      - Takes a long time for large datasets (10K+ rows)
    - An optimization algorithm
      - Gradient Descent
      - Proper approach if you have a very large dataset
- Making predictions with multiple linear regression
  - 

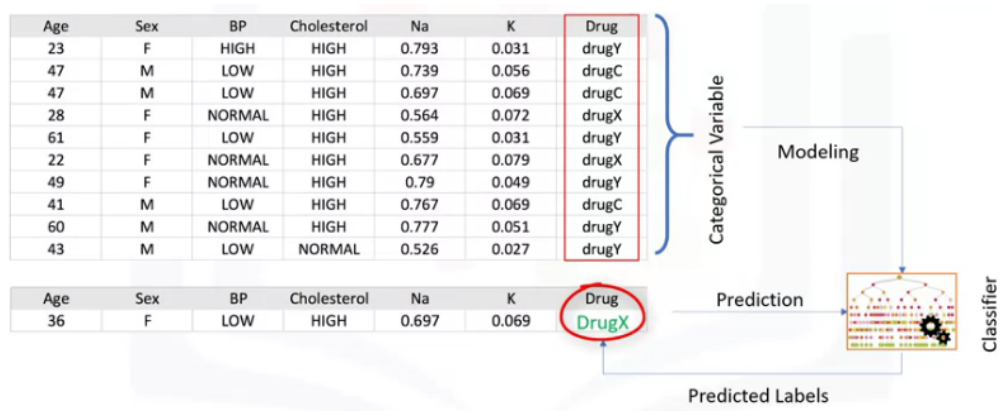
- Out of Sample Accuracy

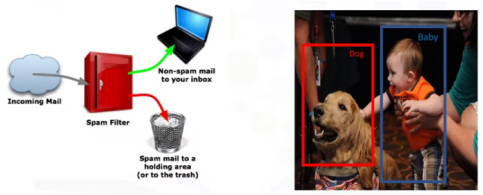
    - "Out of Sample Accuracy" is the percentage of correct predictions that the model makes on data that the model has NOT been trained on.
- When should we use Multiple Linear Regression?
    - When we would like to identify the strength of the effect that the independent variables have on a dependent variable.
    - When we would like to predict impacts of changes in independent variables on a dependent variable.
- Which sentence is TRUE about linear regression?
    - A linear relationship is necessary between the independent variables and the dependent variable.

❖ MODULE #3: Classification

Introduction to Classification

- What is classification?
  - A supervised learning approach
  - Categorizing some unknown items into a discrete set of categories of “classes”
  - The target attribute is a categorical variable
- How does classification work?
  - Classification determines the class label for an unlabeled test case
  - 

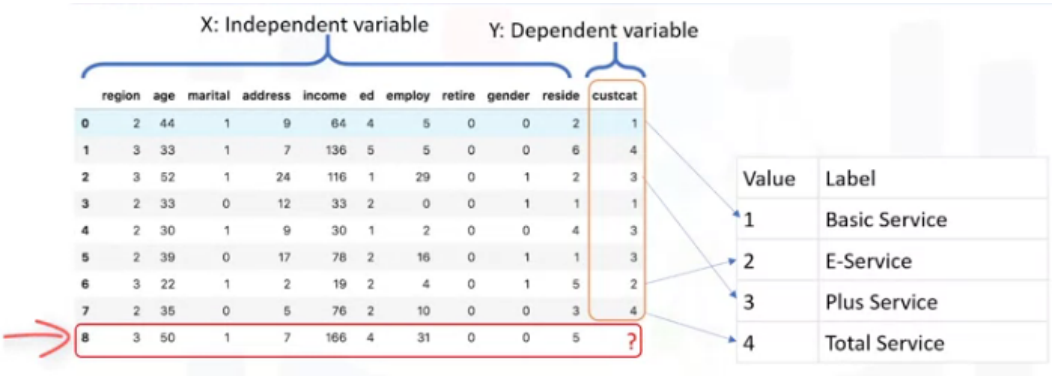
- Example of multi-class classification
  - A classifier that can predict a field with multiple discrete values, such as ”DrugA", ”DrugX" or "DrugY”.
  - 

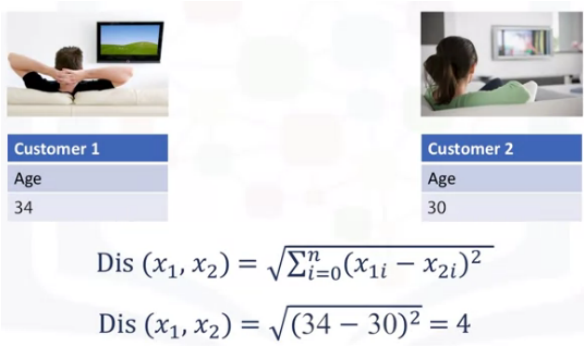
- Classification use cases
  - Which category a customer belongs to?
  - Whether a customer switches to another provider/brand?
  - Whether a customer responds to a particular advertising campaign?
- Classification applications
  - 

- Classification algorithms in machine learning

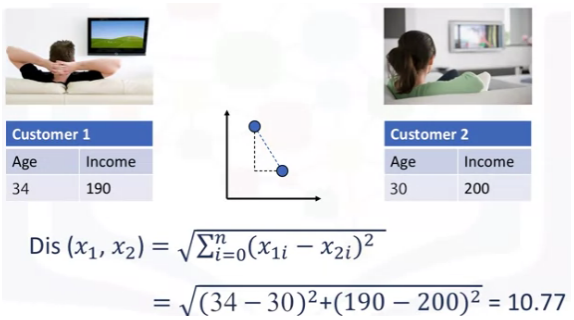
- Decision Trees (ID3, C4.5, C5.0)
- Naive Bayes
- Linear Discriminant Analysis
- K-Nearest Neighbor
- Logistic Regression
- Neural Networks
- Support Vector Machines (SVM)

K-Nearest Neighbors

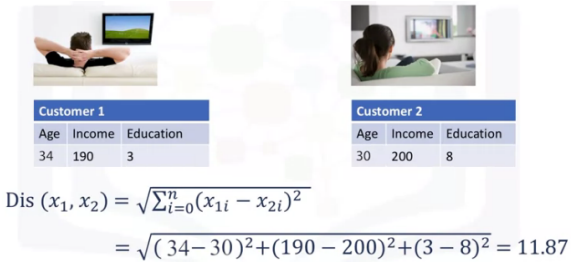
- Intro to KNN
  - Given the dataset with predefined labels, we need to build a model to be used to predict the class of a new or unknown case.
  - 

- What is K-Nearest Neighbor (or KNN)?
  - A method for classifying cases based on their similarity to other cases
  - Cases that are near each other are said to be “neighbors”
  - Based on similar cases with same class labels are near each other
- The K-Nearest Neighbors algorithm
  - 1. Pick a value for K
  - 2. Calculate the distance of unknown case from all cases
  - 3. Select the K-observations in the training data that are “nearest” to the unknown data point
  - 4. Predict the response of the unknown data point using the most popular response value from the K-nearest neighbors
- Calculating the similarity/distance in a 1-dimensional space
  - 

- Calculating the similarity/distance in a 2-dimensional space



- Calculating the similarity/distance in a multi-dimensional space

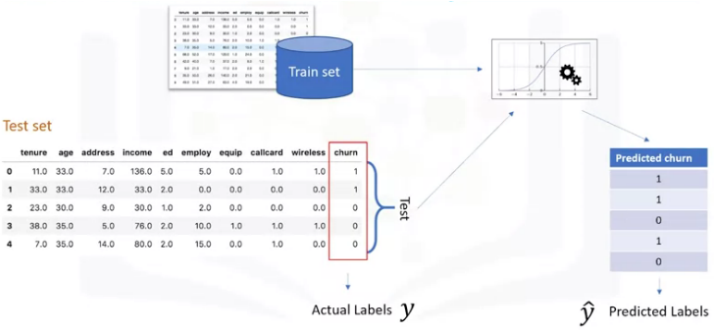


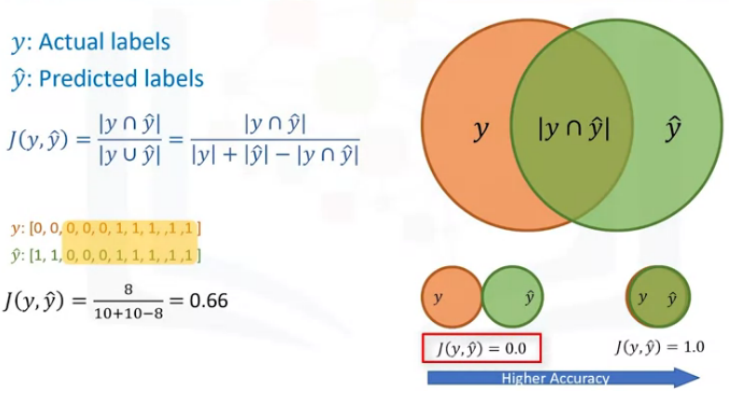
- What is the best value of K for KNN?
  - The general solution is to reserve a part of your data for testing the accuracy of the model. Once you've done so, choose K=1, and then use the training part for modeling and calculate the accuracy of prediction using all samples in your test set. Repeat this process increasing the K and see which K is best for your model. For example, in our case, K equals four will give us the best accuracy.

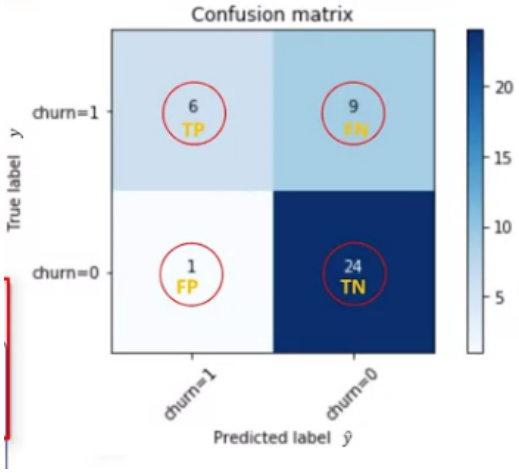
- Computing continuous targets using KNN
  - KNN can also be used for regression

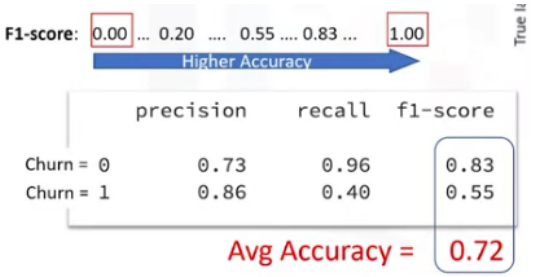
Evaluation Metrics in Classification

- Classification accuracy

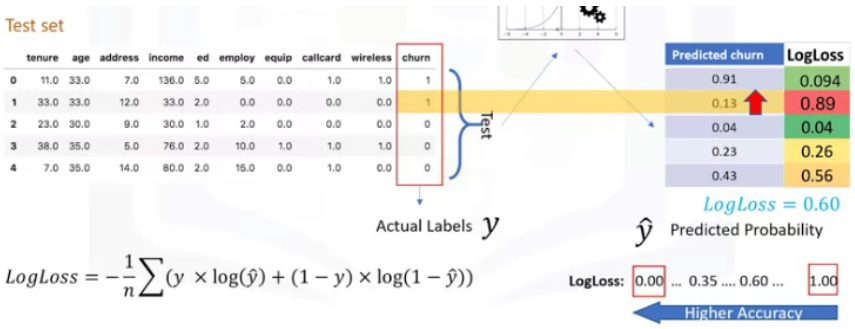


- Jaccard index
  - 

- F1-score
  - 

  - TP = True Positive. FP = False Positive
  - Precision = TP / (TP+FP)
  - Recall = TP / (TP+FN)
  - F1-score = 2x(prc x rec) / (prc + rec)
    - 

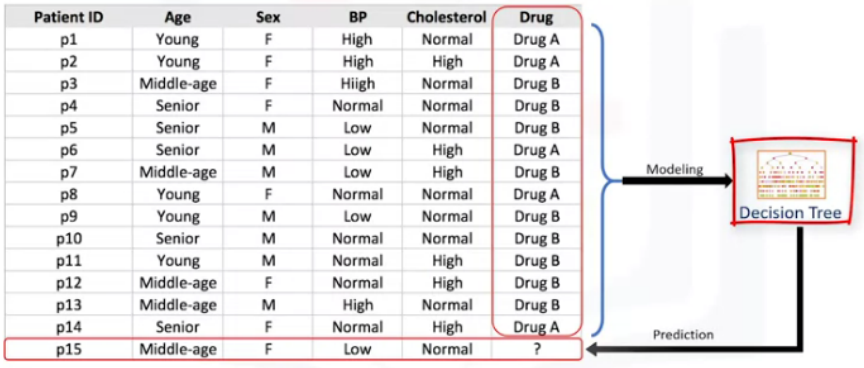
- Log loss
  - Performance of a classifier where the predicted output is a probability value between 0 and 1.



- The smaller the log loss, the better.

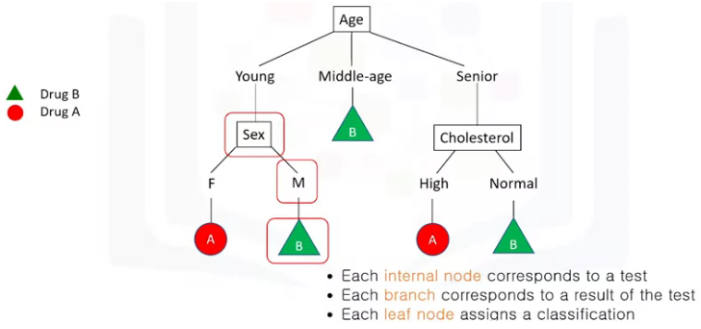
Introduction to Decision Trees

- What is a decision tree?
  - The basic intuition behind a decision tree is to map out all possible decision paths in the form of a tree.
  - A Decision Tree is a type of clustering approach that can predict the class of a group, for example, DrugA or DrugB.
- How to build a decision tree?



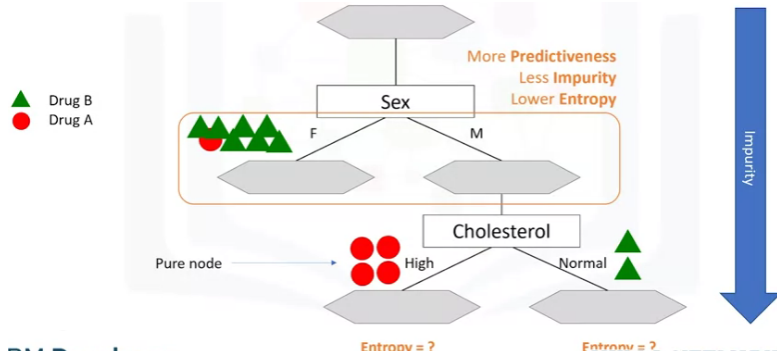
Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hiigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A
p15	Middle-age	F	Low	Normal	?

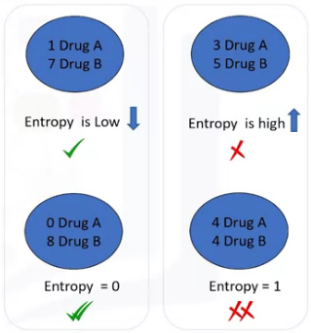
- Building a decision tree with the training set

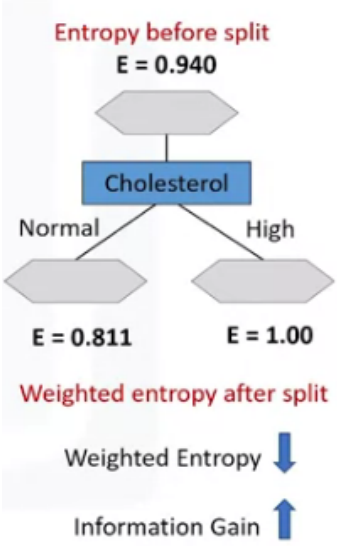


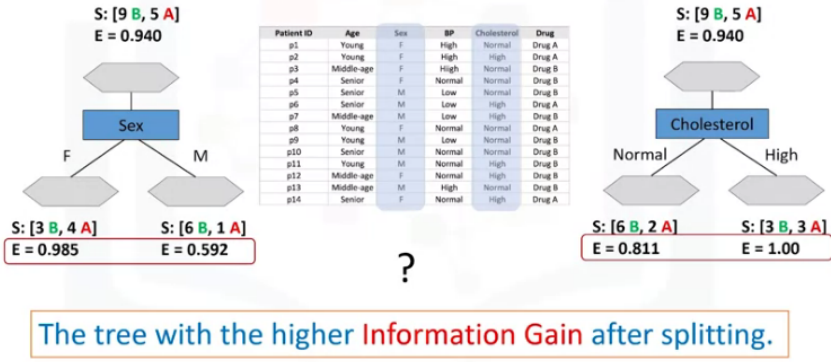
- Decision tree learning algorithm
  - 1. Choose an attribute from your dataset
  - 2. Calculate the significance of attribute in splitting of data
  - 3. Split data based on the value of the best attribute
  - 4. Go to step 1

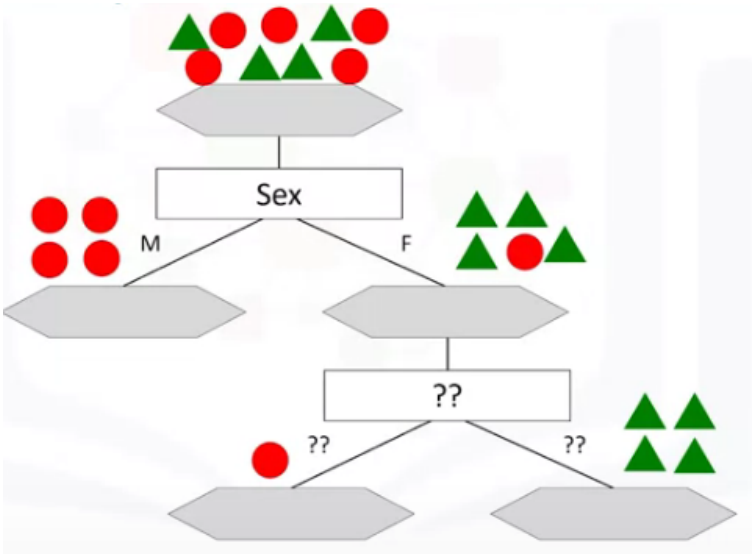
Building Decision Trees

- Which attribute is the best?
  - 

- Entropy
  - Measure of randomness or uncertainty
  - The entropy in a node is the amount of information disorder calculated in each nod
  - The lower the Entropy, the less uniform the distribution, the purer the node
  - 

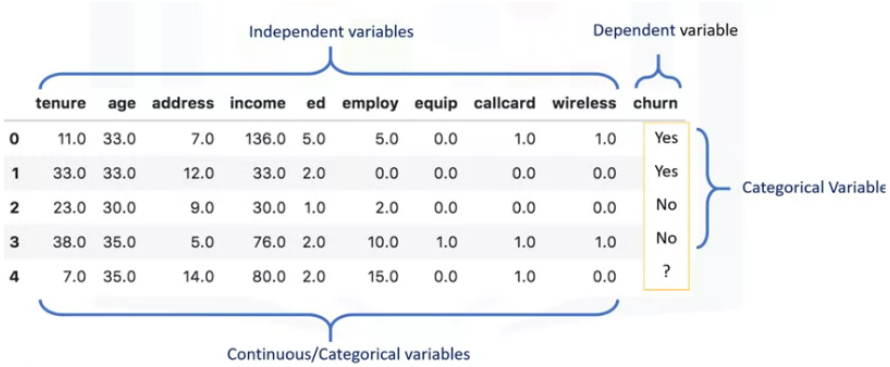
- What is information gain?
  - Information gain is the information that can increase the level of certainty after splitting
  - Information Gain = (Entropy before split) – (weighted entropy after split)
  - 

- Which attribute is the best?
  - 

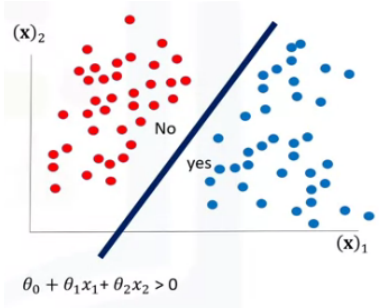
    The tree with the higher Information Gain after splitting.
- Correct way to build a decision tree
  - 

❖ **MODULE #4: Linear Classification**

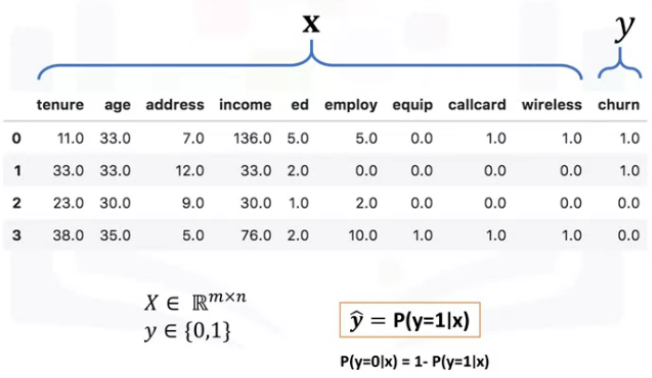
Intro to Logistic Regression

- What is logistic regression?
  - Logistic regression is a classification algorithm for categorical variables
  - In logistic regression, we use one or more independent variables such as tenure, age, and income to predict an outcome, such as churn, which we call the dependent variable representing whether or not customers will stop using the service.
  - 

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	Yes
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	Yes
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	No
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	No
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	?

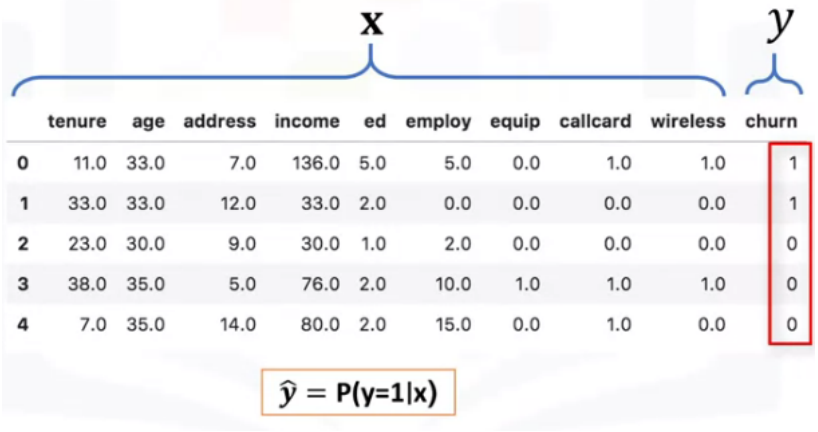
- Logistic regression applications
  - Predicting the probability of a person having a heart attack
  - Predicting the mortality in injured patients
  - Predicting a customer's propensity to purchase a product or halt a subscription
  - Predicting the probability of failure of a given process or product
  - Predicting the likelihood of a homeowner defaulting on a mortgage
- When is logistic regression suitable?
  - If your data is binary
    - 0/1, YES/NO, True/False
  - If you need probabilistic results
  - When you need a linear decision boundary
  - If you need to understand the impact of a feature
  - 

- Building a model for customer churn

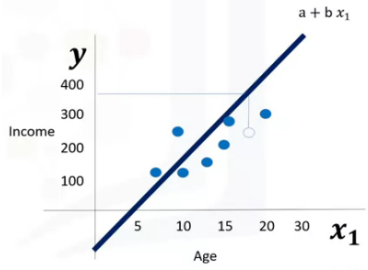


Logistic regression vs Linear regression

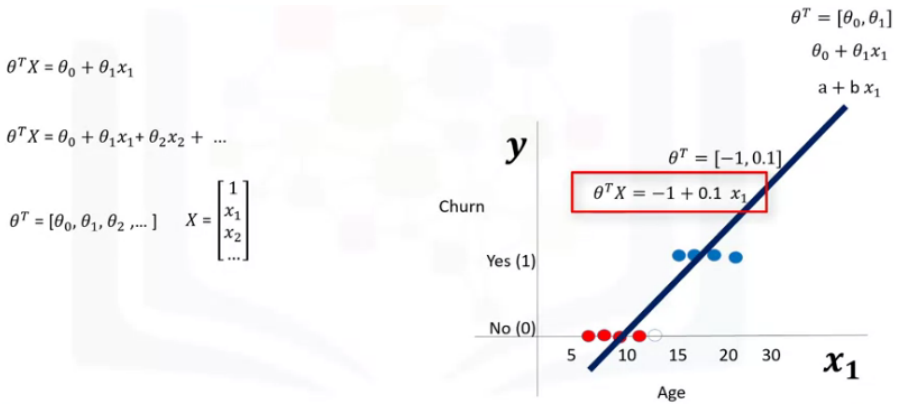
- Model of customer churn data



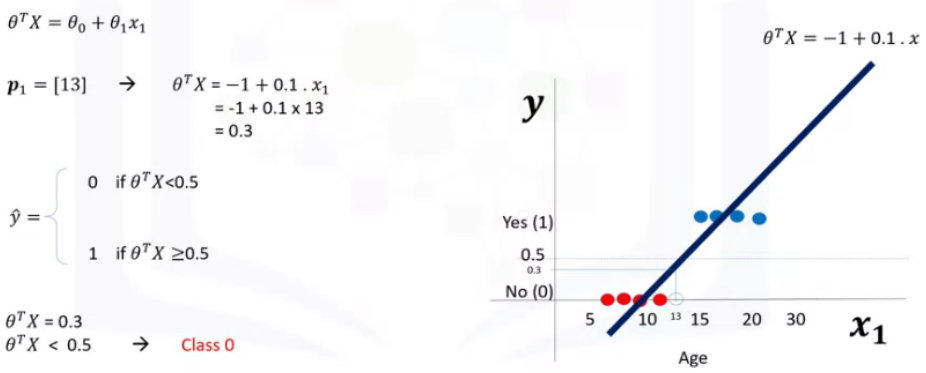
- Predicting customer income



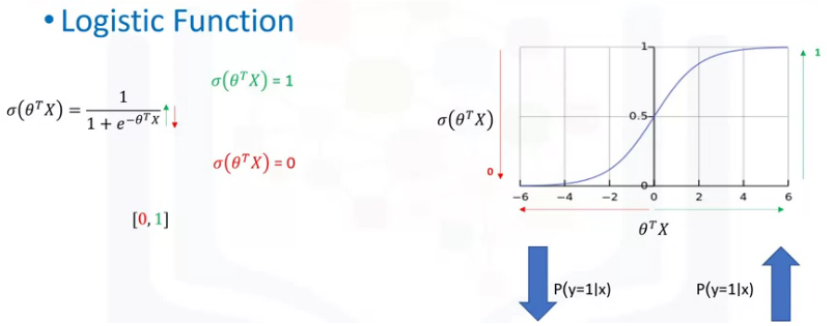
- Predicting churn using linear regression

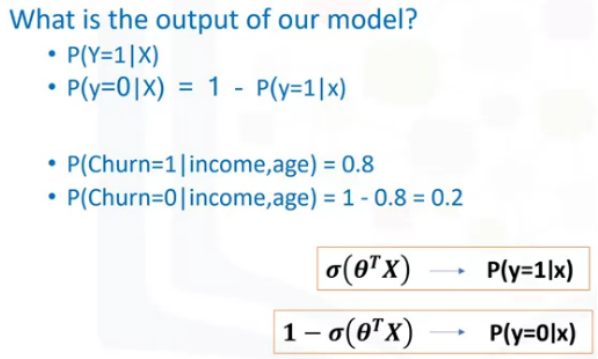


- Linear regression in classification problems?

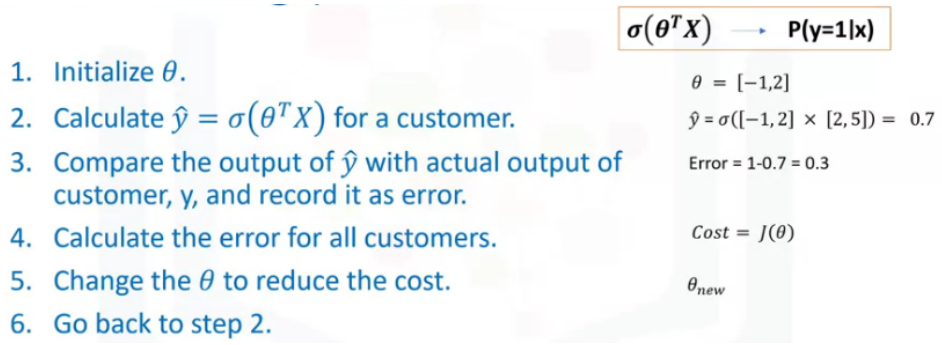
- 

- The problem with using linear regression
  - 

- Sigmoid function in logistic regression
  - 

- Clarification of the customer churn model
  - 

    - Now, how can we achieve this?
- The training process



- Key differences
  - Purpose
    - Linear regression is used for predicting a continuous outcome variable based on one or more predictor variables. It assumes a linear relationship between the independent variables and the dependent variable.
    - Logistic regression is used for predicting the probability of an event occurring or not. It is commonly used for binary classification problems, where the outcome is either 0 or 1.
  - Output Type
    - Linear regression predicts a continuous output.
    - Logistic regression predicts the probability of a binary outcome.
  - Equation Form
    - Linear regression uses a linear equation.
    - Logistic regression uses the logistic function to model the probability.
  - Interpretation
    - In linear regression, the coefficients represent the change in the dependent variable for a one-unit change in the independent variable.
    - In logistic regression, the coefficients represent the change in the log-odds of the dependent variable for a one-unit change in the independent variable.
  - In summary, choose linear regression when you are dealing with a continuous outcome, and choose logistic regression when you are dealing with a binary outcome and want to model probabilities.

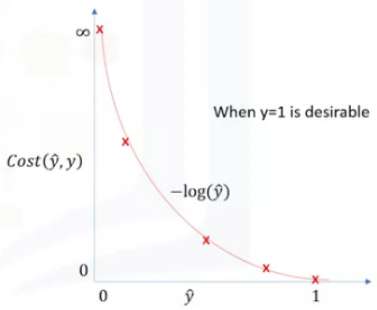
Logistic Regression Training

- General cost function

- $\sigma(\theta^T X) \rightarrow$ P(y=1|x)
- Change the weight -> Reduce the cost
- Cost function $Cost(\hat{y}, y) = \frac{1}{2}(\sigma(\theta^T X) - y)^2$

$$J(\theta) = \frac{1}{m}\sum_{i-1}^{m} Cost(\hat{y}, y)$$

- Plotting the cost function of the model
  - Model $\hat{y}$
  - Actual Value y=1 or 0
  - If Y=1, and $\hat{y}$=1 → cost = 0
  - If Y=1, and $\hat{y}$=0 → cost = large



- Logistic regression cost function
  - So, we will replace cost function with:

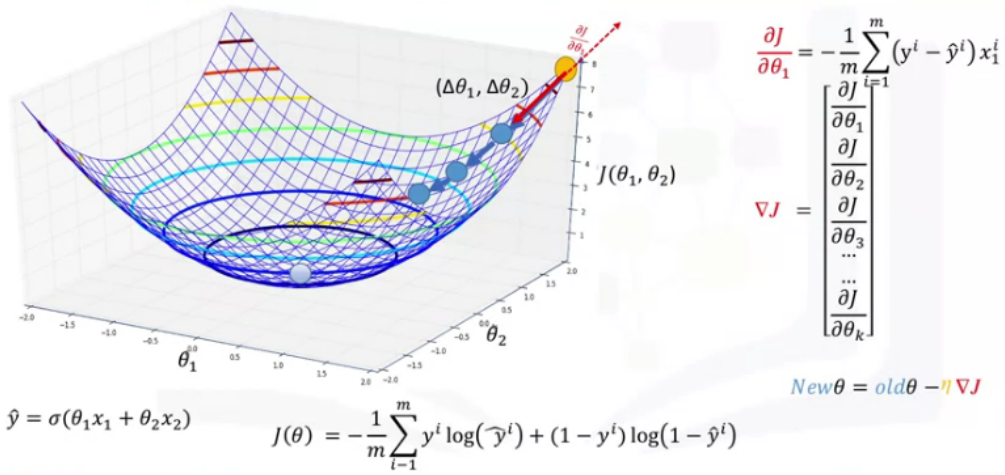
$$Cost(\hat{y}, y) = \frac{1}{2}(\sigma(\theta^T X) - y)^2$$

$$Cost(\hat{y}, y) = \begin{cases} -\log(\hat{y}) & if\ y = 1 \\ -\log(1 - \hat{y}) & if\ y = 0 \end{cases}$$

$$J(\theta) = \frac{1}{m}\sum_{i-1}^{m} Cost(\hat{y}, y)$$

$$J(\theta) = -\frac{1}{m}\sum_{i-1}^{m} y^i \log(\hat{y}^i) + (1 - y^i)\log(1 - \hat{y}^i)$$

- Minimizing the cost function of the model
  - How to find the best parameters of our model?
    - Minimize the cost function
  - How to minimize the cost function?
    - Using Gradient Descent
  - What is gradient descent?
    - A technique to use the derivative of a cost function to change the parameter values, in order to minimize the cost
- Using gradient descent to minimize the cost
  - Error curve

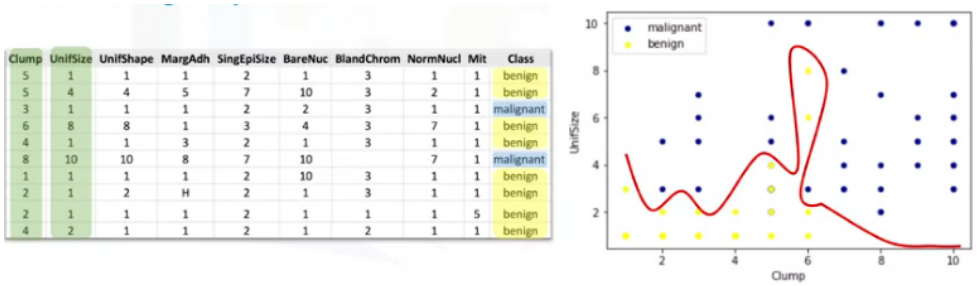


- Training algorithm recap
  - 1. Initialize the parameters randomly
  - 2. Feed the cost function with training set, and calculate the error
  - 3. Calculate the gradient of cost function
  - 4. Update weights with new values
  - 5. Go to step 2 until cost is small enough
  - 6. Predict the new customer X

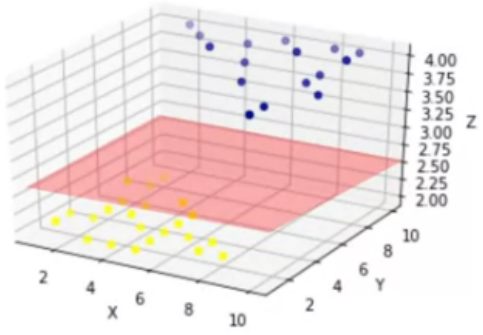
Support Vector Machine (SVM)

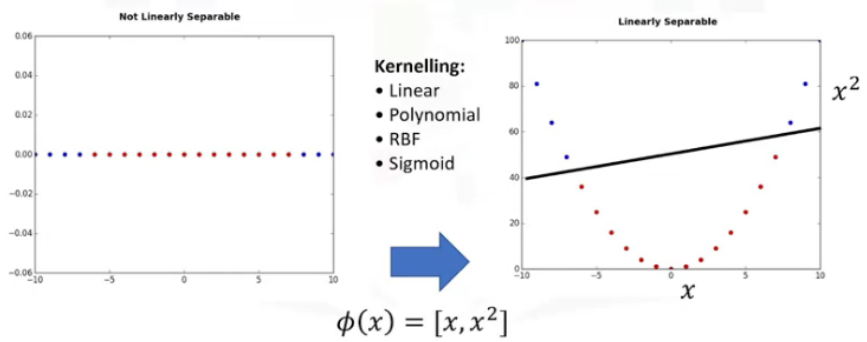
- Classification with SVM

- What is SVM?
  - SVM is a supervised algorithm that classifies cases by finding a separator
    - 1. Mapping data to a high-dimensional feature space
    - 2. Finding a separator

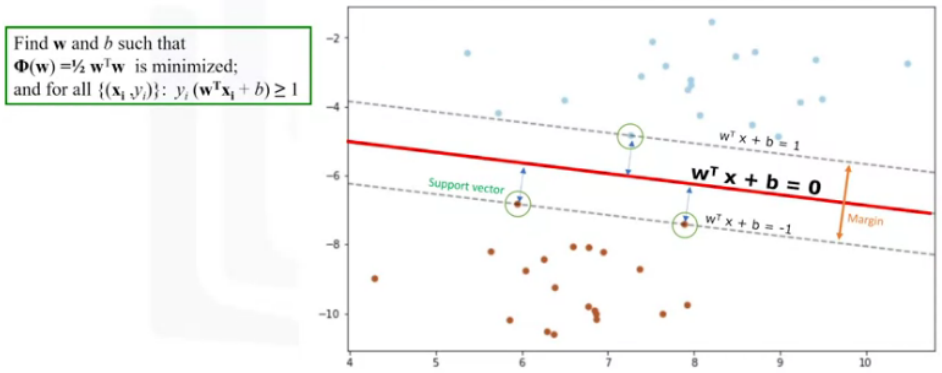
- 

Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
5	1	1	1	2	1	3	1	1	benign
5	4	4	5	7	10	3	2	1	benign
3	1	1	1	2	2	3	1	1	malignant
6	8	8	1	3	4	3	7	1	benign
4	1	1	3	2	1	3	1	1	benign
8	10	10	8	7	10		7	1	malignant
1	1	1	1	2	10	3	1	1	benign
2	1	2	H	2	1	3	1	1	benign
2	1	1	1	2	1	1	1	5	benign
4	2	1	1	2	1	2	1	1	benign

- 

    - How do we transfer data in such a way that a separator could be drawn as a hyperplane?
    - How can we find the best or optimized hyperplane separator after transformation?
- Data transformation
    - 

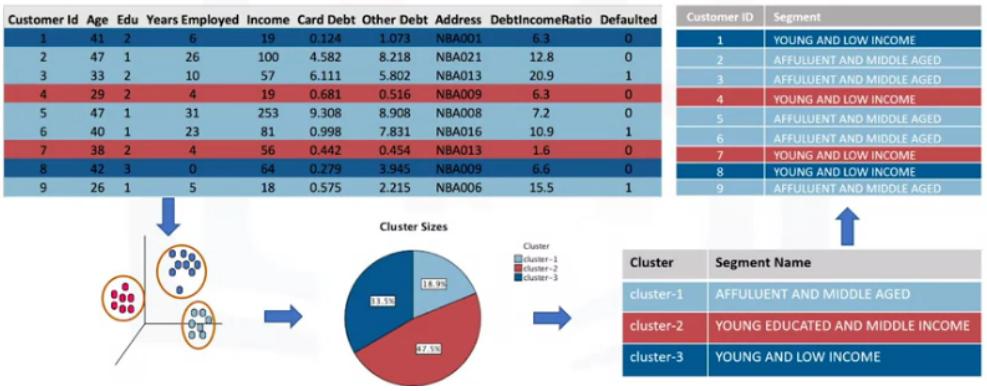
    - The mathematical function used for the transformation is known as the kernel function such as linear, polynomial, RBF and sigmoid
    - What is the meaning of "Kernelling" in SVM?
        - Mapping data into a higher dimensional space, in such a way that can change a linearly inseparable dataset into a linearly separable dataset.
- Using SVM to find the hyperplane
    - The goal is to choose a hyperplane with as big a margin as possible.
    - Examples closest to the hyperplane are support vectors. We tried to find the hyperplane in such a way that it has the maximum distance to support vectors.
        - The hyperplane and boundary decision lines have their own equations.
    - The hyperplane is learned from training data using an optimization procedure that maximizes the margin.

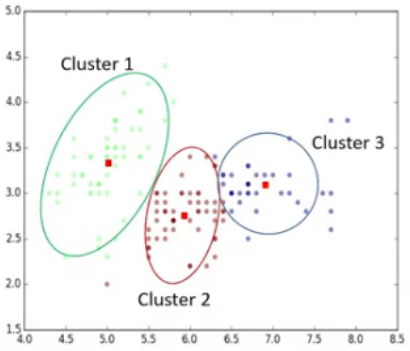
- 

- Pros and cons of SVM
  - Advantages
    - Accurate in high-dimensional spaces
    - Memory efficient
  - Disadvantages
    - Prone to over-fitting
    - No probability estimation
    - Small datasets (if your dataset is big it won't work)
- SVM applications
  - Image recognition
  - Text category assignment
  - Detecting spam
  - Sentiment analysis
  - Gene Expression Classification
  - Regression, outlier detection and clustering

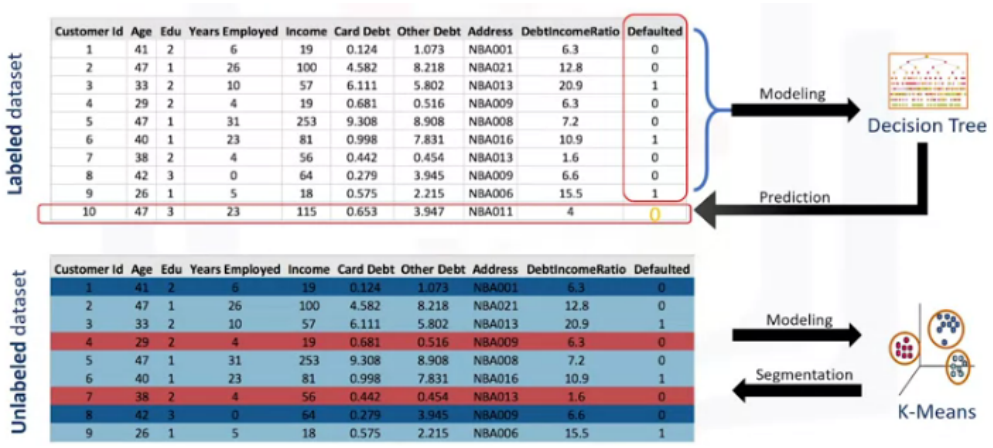
❖ MODULE #5: Clustering

Intro to Clustering

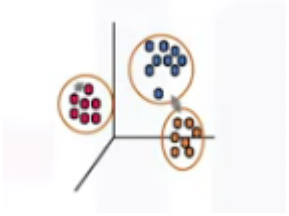
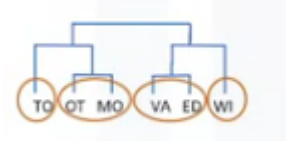
- Clustering for segmentation
    - Clustering can group data only unsupervised, based on the similarity of customers to each other.
    - It will partition your customers into mutually exclusive groups.
    - For example, into three clusters. The customers in each cluster are similar to each other demographically.
    - 

- What is clustering?
    - A group of objects that are similar to other objects in the cluster, and dissimilar to data points in other clusters.
    - 

- Clustering vs. Classification
    - Classification
        - Classification algorithms predict categorical classed labels. This means assigning instances to predefined classes such as defaulted or not defaulted.
        - For example, if an analyst wants to analyze customer data in order to know which customers might default on their payments, she uses a labeled dataset as training data and uses classification approaches such as a decision tree, Support Vector Machines or SVM, or logistic regression, to predict the default value for a new or unknown customer.

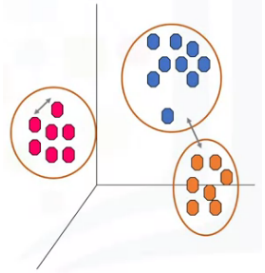
        - Generally speaking, classification is supervised learning where each training data instance belongs to a particular class.
    - Clustering
        - In clustering however, the data is unlabeled and the process is unsupervised.
        - For example, we can use a clustering algorithm such as k-means to group similar customers as mentioned, and assign them to a cluster, based on whether they share similar attributes, such as; age, education, and so on.
    - 

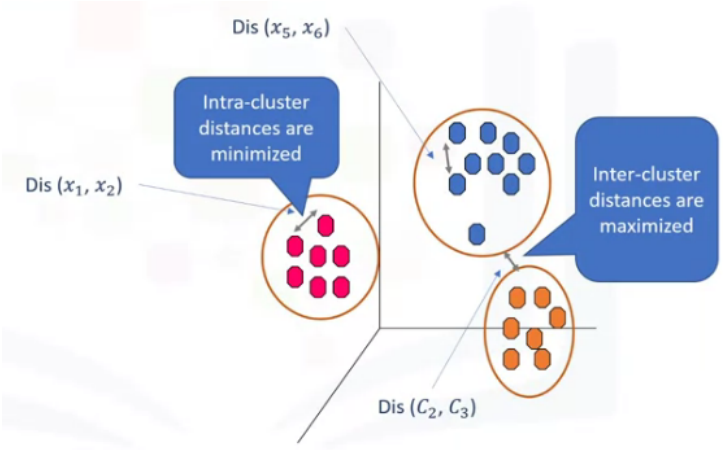
- Clustering applications
    - Retail/Marketing
        - Identifying buying patterns of customers
        - Recommending new books or movies to new customers
    - Banking
        - Fraud detection in credit card use
        - Identifying clusters of customers (e.g., loyal)
    - Insurance
        - Fraud detection in claims analysis
        - Insurance risk of customers
    - Publication
        - Auto-categorizing news based on their content
        - Recommending similar news articles
    - Medicine
        - Characterizing patient behavior
    - Biology
        - Clustering genetic markers to identify family ties
- Why clustering?
    - Exploratory data analysis
    - Summary generation
    - Outlier detection
    - Finding duplicates
    - Pre-processing step
- Clustering algorithms

    - Partitioned-based clustering
        - Relatively efficient
        - E.g. k-Means, k-Median, Fuzzy c-Means
        - 
    - Hierarchical Clustering
        - Produces trees of clusters
        - E.g. Agglomerative, Divisive
        - 

    - Density-based Clustering
        - Produces arbitrary shaped clusters
        - E.g. DBSCAN
        - 

Intro to k-Means

- What is k-Means clustering?
    - K-Means can group data only unsupervised based on the similarity of customers to each other.
- k-Means algorithms
    - Partitioning Clustering
    - K-means divides the data into non-overlapping subsets (clusters) without any cluster-internal structure
    - Examples within a cluster are very similar
    - Examples across different clusters are very different
    - 
- Determine the similarity or dissimilarity



- What is the objective of k-means?
  - To form clusters in such a way that similar samples go into a cluster, and dissimilar samples fall into different clusters.
  - To minimize the “intra cluster” distances and maximize the “inter-cluster” distances.
  - To divide the data into non-overlapping clusters without any cluster-internal structure

- 1-dimensional similarity/distance

Customer 1
Age
54

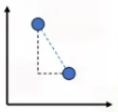
Customer 2
Age
50

$$\text{Dis}(x_1, x_2) = \sqrt{\sum_{i=0}^{n}(x_{1i} - x_{2i})^2}$$

$$\text{Dis}(x_1, x_2) = \sqrt{(54 - 50)^2} = 4$$

- 2-dimensional similarity/distance

Customer 1	
Age	Income
54	190



Customer 2	
Age	Income
50	200

$$\text{Dis}(x_1, x_2) = \sqrt{\sum_{i=0}^{n}(x_{1i} - x_{2i})^2}$$

$$= \sqrt{(54 - 50)^2 + (190 - 200)^2} = 10.77$$

- Multi-dimensional similarity/distance

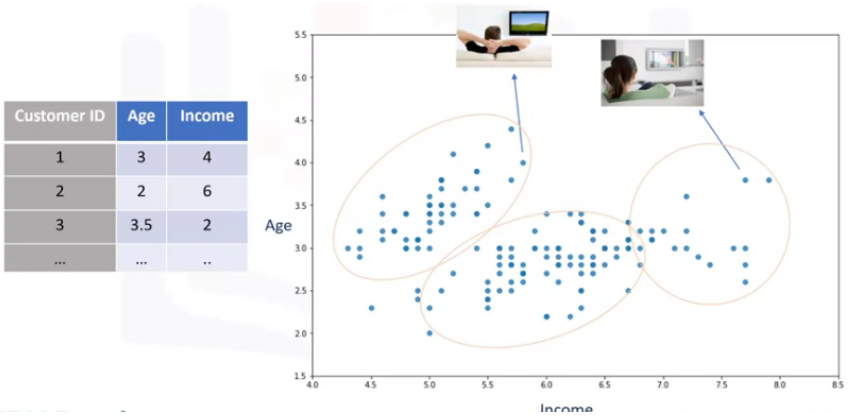
Customer 1		
Age	Income	education
54	190	3

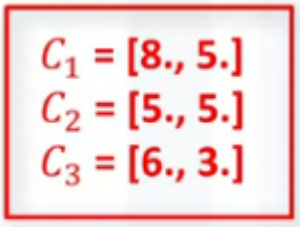
Customer 2		
Age	Income	education
50	200	8

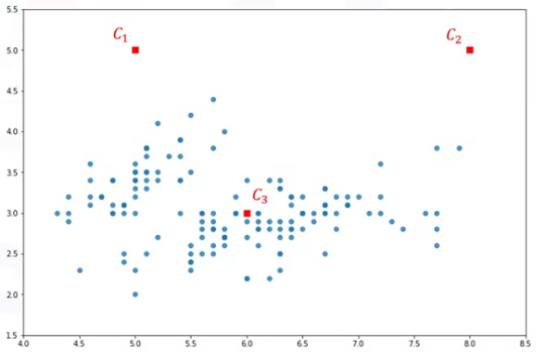
$$\text{Dis}(x_1, x_2) = \sqrt{\sum_{i=0}^{n}(x_{1i} - x_{2i})^2}$$

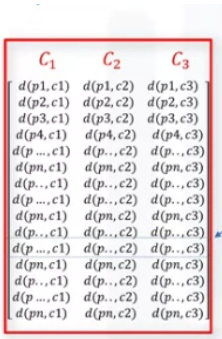
$$= \sqrt{(54 - 50)^2 + (190 - 200)^2 + (3 - 8)^2} = 11.87$$

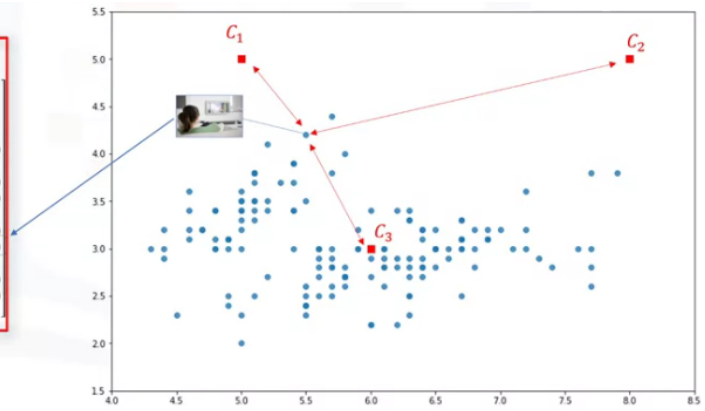
- How does k-Means clustering work?

- 

- k-Means clustering – initialize k
  - 1. Initialize k=3 centroids randomly
    - There are two approaches to choose these centroids.
    - 1. We can randomly choose three observations out of the dataset and use these observations as the initial means.
    - or
    - 2. We can create three random points as centroids of the clusters which is our choice that is shown in the plot with red color.
    - 

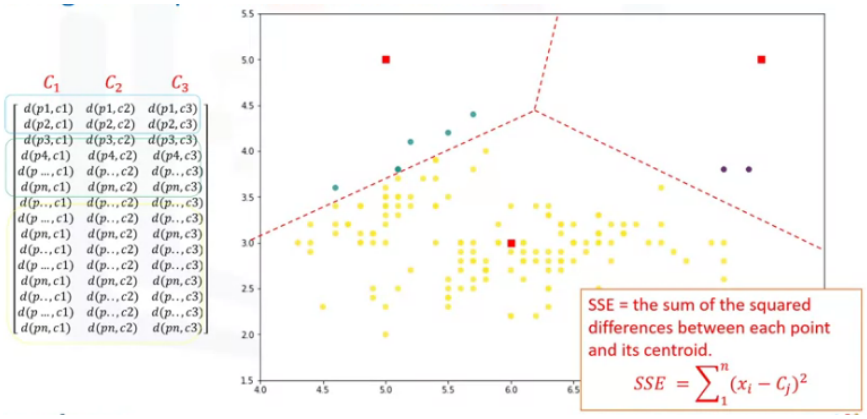
    

  - 2. Distance calculation
    - You will form a matrix where each row represents the distance of a customer from each centroid. It is called the Distance Matrix.
    - 

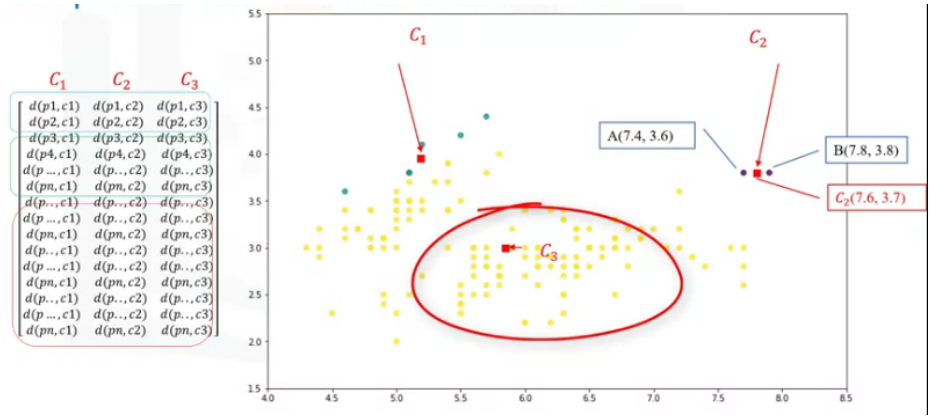
    

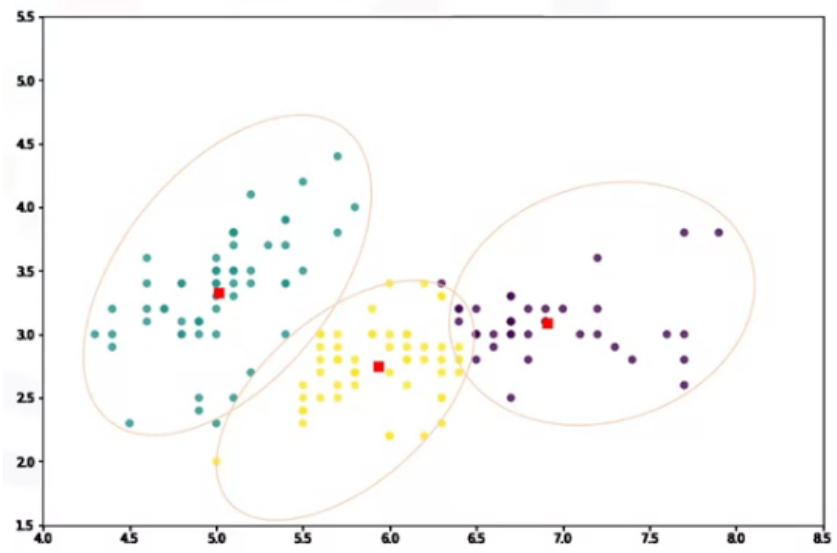
  - 3. Assign each point to the closest centroid

- So, in this step, we have to find the closest centroid to each data point. We can use the distance matrix to find the nearest centroid to data points.
- Finding the closest centroids for each data point, we assign each data point to that cluster. In other words, all the customers will fall to a cluster based on their distance from centroids.
- In other words, all the customers will fall to a cluster based on their distance from centroids.
- We can easily say that it does not result in good clusters because the centroids were chosen randomly from the first. Indeed, the model would have a high error.
- Here, error is the total distance of each point from its centroid. It can be shown as a within-cluster sum of squares error. Intuitively, we try to reduce this error. It means we should shape clusters in such a way that the total distance of all members of a cluster from its centroid be minimized.

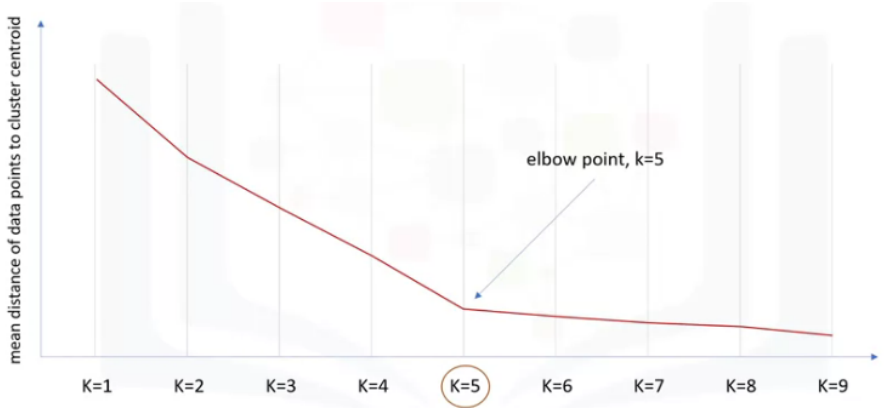


- 4. Compute the new centroids for each cluster
  - In the next step, each cluster center will be updated to be the mean for datapoints in its cluster. Indeed, each centroid moves according to their cluster members.
  - In other words the centroid of each of the three clusters becomes the new mean.
  - For example, if point A coordination is 7.4 and 3.6, and B point features are 7.8 and 3.8, the new centroid of this cluster with two points would be the average of them, which is 7.6 and 3.7.
  - Now, we have new centroids. The points are reclustered and the centroids move again. This continues until the centroids no longer move. Please note that whenever a centroid moves, each point's distance to the centroid needs to be measured again.

- 

- 5. Repeat until there are no more changes
  - Yes, K-Means is an iterative algorithm and we have to repeat steps two to four until the algorithm converges.
  - In each iteration, it will move the centroids, calculate the distances from new centroids and assign data points to the nearest centroid. It results in the clusters with minimum error or the most dense clusters.
  - However, as it is a heuristic algorithm, there is no guarantee that it will converge to the global optimum and the result may depend on the initial clusters. It means, this algorithm is guaranteed to converge to a result, but the result may be a local optimum i.e. not necessarily the best possible outcome.
  - To solve this problem, it is common to run the whole process multiple times with different starting conditions. This means with randomized starting centroids, it may give a better outcome. As the algorithm is usually very fast, it wouldn't be any problem to run it multiple times.
  - 

More on k-Means

- k-Means clustering algorithm
    - 1. Randomly placing k centroids, one for each cluster.
    - 2. Calculate the distance of each point from each centroid.
    - 3. Assign each data point (object) to its closest centroid, creating a cluster.
    - 4. Recalculate the position of the k centroids.
    - 5. Repeat the steps 2-4, until the centroids no longer move.
- k-Means accuracy
    - External approach
        - Compare the clusters with the ground truth, if it is available
    - Internal approach
        - Average the distance between data points within a cluster
- Choosing K
    - The correct choice of K is often ambiguous because it's very dependent on the shape and scale of the distribution of points in a dataset.
    - This metric can be mean, distance between data points and their cluster's centroid, which indicate how dense our clusters are or, to what extent we minimize the error of clustering. Then, looking at the change of this metric, we can find the best value for K.
    - But the problem is that with increasing the number of clusters, the distance of centroids to data points will always reduce. This means increasing K will always decrease the error. So, the value of the metric as a function of K is plotted and the elbow point is determined where the rate of decrease sharply shifts. It is the right K for clustering.
    - This method is called the ELBOW METHOD.
    - 

- k-Means recap
    - Med and Large sized databases (Relatively efficient)
    - Produces sphere-like clusters
    - Needs number of clusters (k)