# Machine Learning and Deep Learning
# Lecture-04

By: Somnath Mazumdar
Assistant Professor
sma.digi@cbs.dk

# Outline

- Principles of supervised machine learning
- K-Nearest Neighbors
- Linear Regression
- Logistic Regression

# Supervised Learning

# Fundamentals of Machine Learning

- Supervised Learning:
  - Predicting values.
  - Known targets.
  - User inputs correct answers to learn from.
  - Machine uses the information to guess new answers.
- Categories:
  - Regression: Estimate continuous values.
  - Classification: Identify a unique class.

# How supervised ML works?

- Collect data and preprocess it
- Split a dataset into training and test data
- Choose an algorithm and set its hyperparameters.
- Train a model with the training data
- Evaluate how well the model works with the test data
- Repeat steps 2-4 until you are satisfied with the results

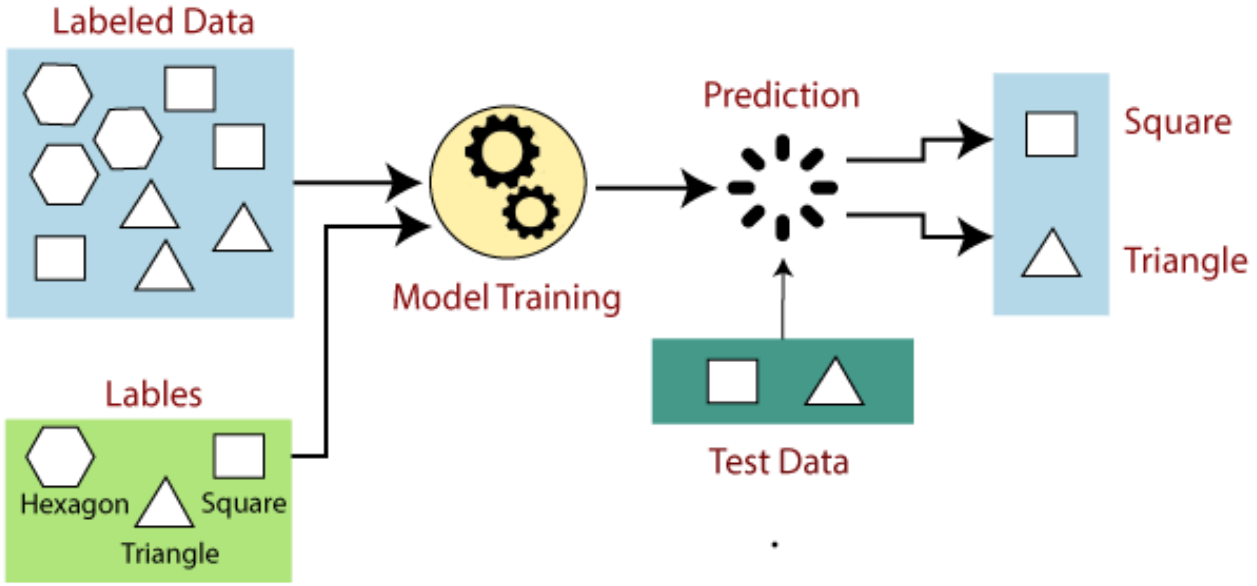
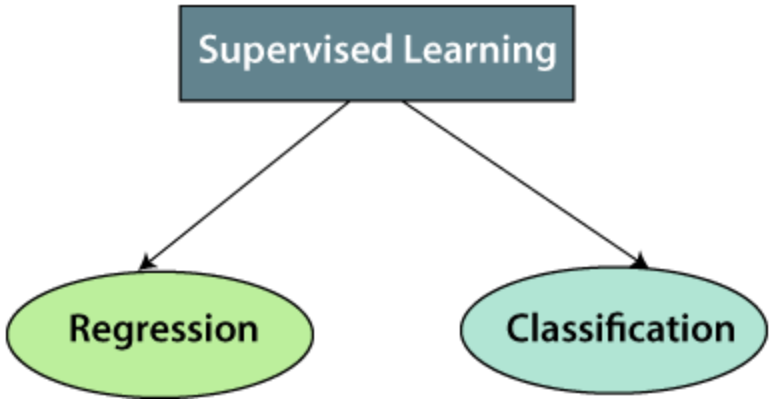



Image src: https://www.javatpoint.com/supervised-machine-learning

# Supervised Categories

(Real-valued output)




(Discrete values, Boolean Categories)

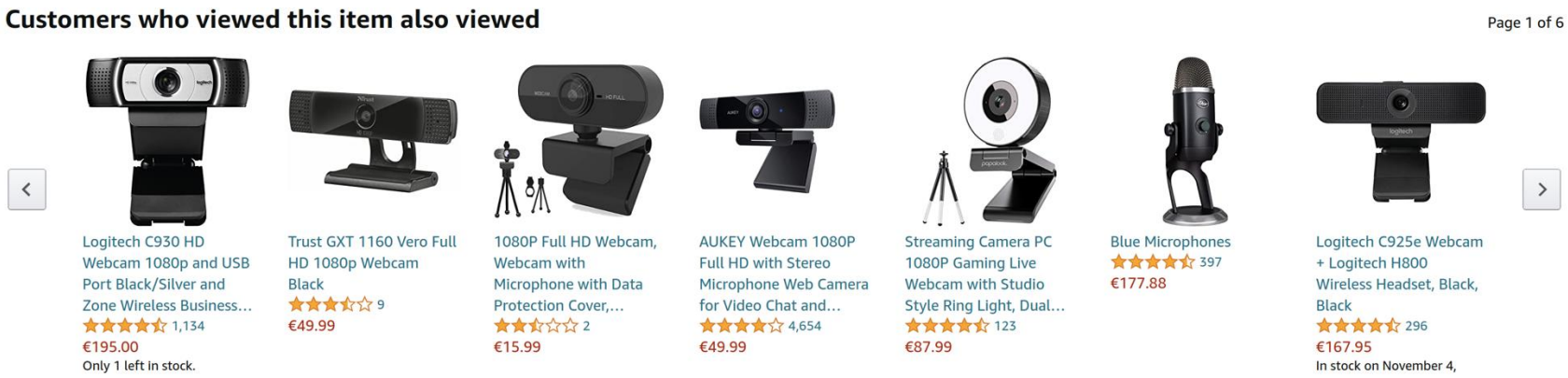
- Linear Regression
- Neural Networks*

Some NNs can be unsupervised, such as autoencoders and restricted Boltzmann.

- Decision Trees,
- K-Nearest Neighbors
- Support Vector Machine
- Logistic Regression
- Random Forests

# Three C's of ML

- Three C’s of ML:
  - Collaborative filtering
  - Clustering
  - Classification
- Collaborative filtering is a technique for recommendations
  - It’s one primary type of recommender system
  - Can use the same algorithm to recommend practically anything – Movies (Netflix)
  - Amazon uses CF to recommend a variety of products



# Three C's of ML

- Classification is a form of ‘supervised’ learning.
  - Requires training with data that has known labels.
  - Learns how to label new records based on that information.

- Applications:
  - Spam filtering: Train using a set of spam and non/spam messages --> System will eventually learn to detect unwanted e-mail.

  - Risk Analysis: Train using financial records of customers who do/don’t default --> System will eventually learn to identify risk customers

# K-Nearest Neighbors: KNN

# K-Nearest Neighbors: KNN

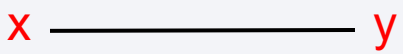
- Non-parametric, supervised learning classifier.
- Uses proximity to make classifications or predictions about grouping of an individual data point.
- Can be used for either regression or classification but used for classification.
- Evelyn Fix and Joseph Hodges are credited with the initial ideas around the KNN model in 1951.
- Application: Commonly used for simple recommendation systems, pattern recognition, data mining, financial market predictions, intrusion detection.
- Issues: As dataset grows, KNN becomes increasingly inefficient, compromising overall model performance. [Scaling problem]

https://www.ibm.com/in-en/topics/knn

# Compute KNN: Distance Metrics

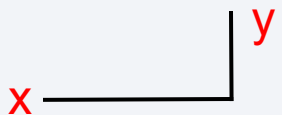
- Goal: Identify nearest neighbors of a given query point, so that we can assign a class label to that point.
- Determine your distance metrics using:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$




- Popular distance measure
- Limited to real-valued vectors.

$$\text{Manhattan Distance} = d(x,y) = \left(\sum_{i=1}^{m}|x_i - y_i|\right)$$




Measures absolute value between two points.

$$\text{Minkowski Distance} = \left(\sum_{i=1}^{n}|x_i - y_i|\right)^{1/p}$$

Generalized form of Euclidean and Manhattan distance metrics.

# Compute KNN: Defining k

- K-value in k-NN algorithm defines how many neighbors will be checked to determine the classification of a specific query point.
  - Defining k is important but not easy.
- Different k-values can lead to overfitting or underfitting.
  - Lower values of k can have high variance, but low bias.
  - Larger values of k may lead to high bias and lower variance.

https://www.ibm.com/in-en/topics/knn

# Compute KNN: Defining k

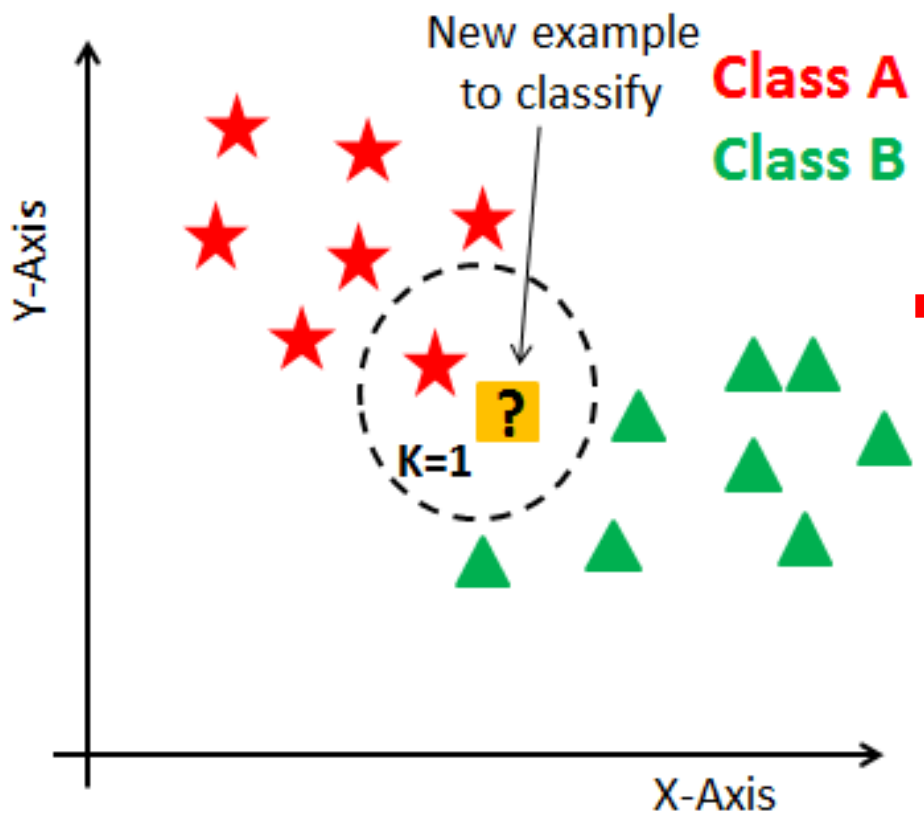
- Choice of k will largely depend on input data.
- Data with more outliers or noise will likely perform better with higher values of k.
- Must have an odd number for k to avoid ties in classification.
- Cross-validation can help choose optimal k for your dataset.

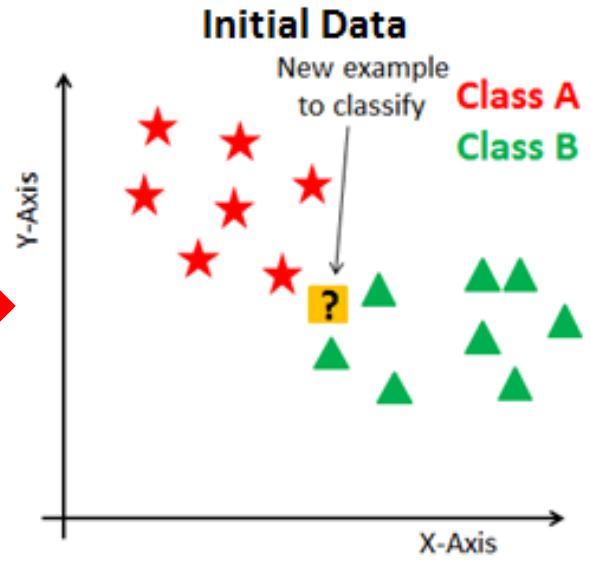
# K-nearest neighbors

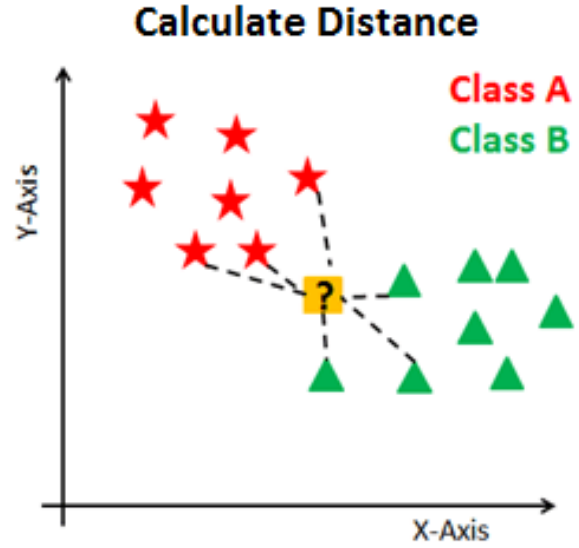
- Given a training dataset $X$ and a new instance $x_{new}$
- Find $k$ points in $X$ that are closest to $x_{new}$
- Using the selected distance measure:
- Predict label for $x_{new}$
- Classification: majority vote among the $k$ nearest neighbors.
- Regression: mean of the $k$ nearest neighbors.

sklearn.neighbors.KNeighborsClassifier

# K-nearest neighbors



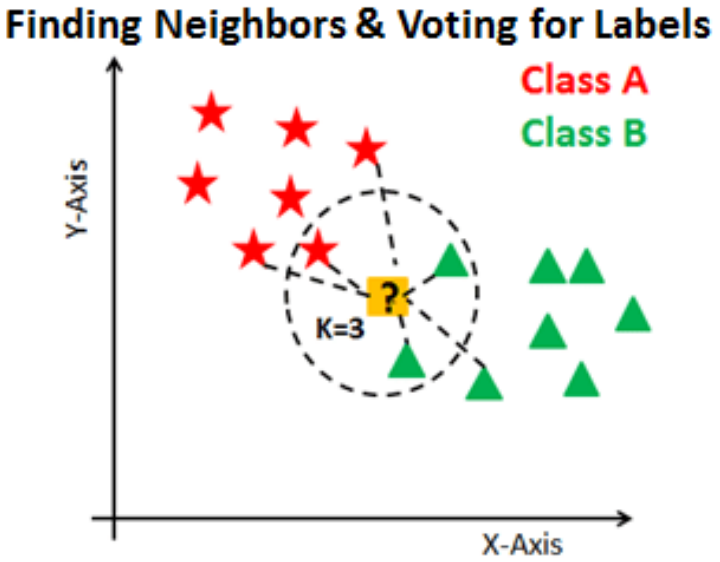




Image from Datacamp

# Compute KNN: Code

```
from sklearn.neighbors import KNeighborsClassifier
model_name = 'K-Nearest Neighbor Classifier'
knnClassifier = KNeighborsClassifier(n_neighbors = 5, metric = 'minkowski', p=2)
knn_model = Pipeline(steps=[('preprocessor', preprocessorForFeatures), ('classifier' ,
knnClassifier)])
knn_model.fit(X_train, y_train)
y_pred = knn_model.predict(X_test)
```

# Applications of k-NN

1. Data preprocessing: Datasets frequently have missing values, but KNN can estimate for those values [Missing data imputation].
2. Recommendation Engines: Using clickstream data from websites, KNN can provide automatic recommendations to users on additional content.
3. Finance: Using KNN on credit data can help banks assess risk of a loan.
4. Healthcare: Predicting risk of heart attacks and prostate cancer.
5. Pattern Recognition: KNN has also assisted in identifying patterns in text and in digit classification.

https://www.ibm.com/in-en/topics/knn

# Advantages and Disadvantages

- Advantages:
  - Easy to implement
  - Adapts easily
  - Few hyperparameters

- Disadvantages:
  - Does not scale well
  - Curse of dimensionality
  - Prone to overfitting

# Remember!!

- Features need to be scaled:
    - A feature with a big scale can dominate all the distances.
    - A feature with a small scale would get neglected.

- “Curse of dimensionality”:
    - Problems with high-dimensional spaces (e.g. more than 5 features).
        - Volume of space grows exponentially with dimension.
        - Need more points to ‘fill’ a high-dimensional volume.

# Linear Regression

# Linear Regression

- Linear regression analysis is used to predict the value of a variable based on the value of another variable.
  - Dependent variable: You want to predict.
  - Independent variable: You are using to predict the other variable's value.
- Estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable.

# Linear Regression

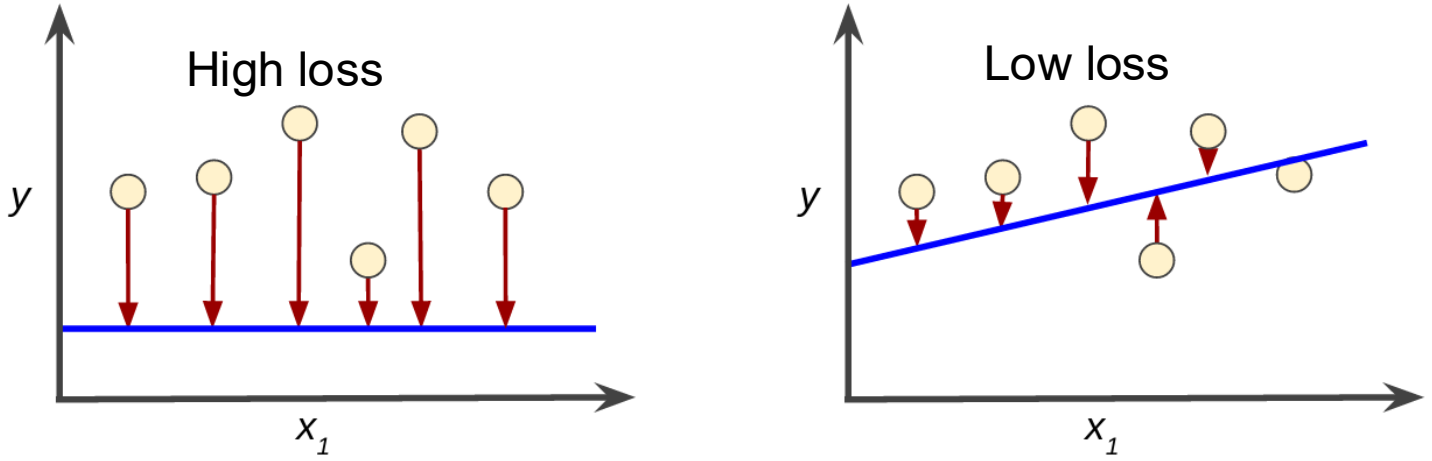
- You can take large amounts of raw data and transform it into actionable information.
  - Better insights by uncovering patterns and relationships.

- For example, performing an analysis of sales and purchase data can help you uncover specific purchasing patterns on particular days or at certain times.

- Disadvantages:
  - Performs poorly when there are non-linear relationships.
  - Sensitive to outliers.

# Linear Regression

- A common method for regression problems.
- Linear regression models are parametric models.
- Fits a straight line or surface that minimizes the discrepancies between predicted and actual output values.



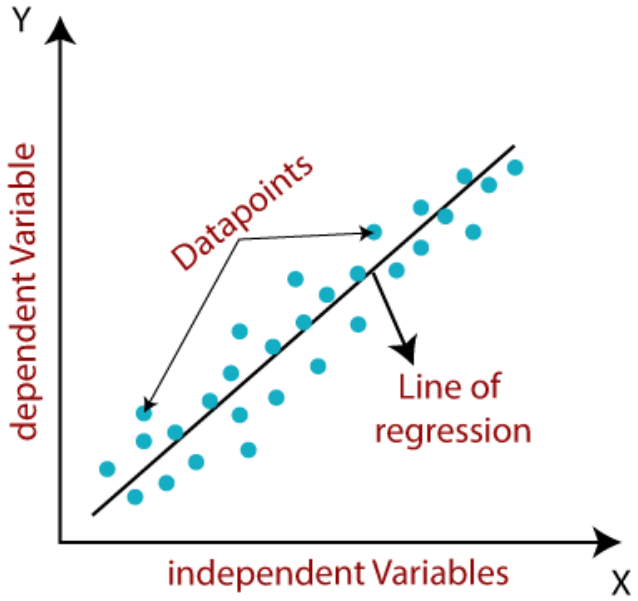

- Simple linear regression calculators use a “least squares” method to discover the best-fit line for a set of paired data.



https://developers.google.com/machine-learning/crash-course/descending-into-ml/training-and-loss

# Linear Regression

- In statistics, a simple linear regression is represented as:
- $y = mx + b$




- ***y*** *is* the value we're trying to predict
- ***m*** is the slope of the line
- ***x*** is the value of our input feature
- ***b*** is the y-intercept

- Linear-regression models are relatively simple.
- Linear regression is a long-established statistical procedure.
- Linear-regression models are well understood and can be trained very quickly.

# Linear Regression

- LR: $y' = b + w_1 x_1$ ; $y' = b + w_1x_1 + w_2x_2 + w_3x_3$ (Three features)
  - ***y'*** is predicted label (a desired output)
  - ***b*** is bias (y-intercept)
  - $w_1$ is weight of feature 1 (slope);
  - ***$x_1$*** is feature (input)

sklearn.linear_model.LinearRegression()

- ML algorithm for linear regression attempts to find values for b and $w_i$ that minimizes the loss (e.g. mean squared error).

# Linear Regression: Key Assumptions

1. Data: Dependent and independent variables should be quantitative. (Convert Categorical variables to binary).
   - For each variable: Consider number of valid cases, mean and SD.
2. For each model: Consider regression coefficients, correlation matrix, variance-covariance matrix etc
3. Plots: Consider scatterplots, partial plots, histograms.
4. Other assumptions:
   - For each value of the independent variable, the distribution of the dependent variable must be normal.
   - The relationship between the dependent variable and each independent variable should be linear and all observations should be independent.



https://www.ibm.com/topics/linear-regression

# Linear Regression: Data Assumptions

- Variables should be measured at a continuous level (time, sales, weight).
- Use a scatterplot to find out quickly if there is a linear relationship between those two variables.
- The observations should be independent of each other.
- Your data should have no significant outliers.
- Check that the variances along the best-fit linear-regression line remain similar all through that line.
- The residuals (errors) of the best-fit regression line follow normal distribution.



https://www.ibm.com/topics/linear-regression

# Linear Regression

- Linear regression models identify relationship between a continuous dependent variable and one or more independent variables.
- Linear regression produces a probability.
- Simple linear regression supports one independent variable and one dependent variable
- Multiple linear regression supports more independent variables.
- For each type of linear regression, it seeks to plot a line of best fit through a set of data points.

# Logistic Regression

# Logistic Regression

- Logistic regression is a data analysis technique that uses mathematics to find the relationships between two data factors.
  1. Then uses relationship to predict the value of one of those factors based on the other.
  2. Prediction usually has a finite number of outcomes, like yes or no.
- Steps:
  1. Identify the question: Any data analysis begins with a business question. For logistic regression, you should frame the question to get particular outcomes:
     - Do rainy days impact our monthly sales? (yes or no)
     - What type of credit card activity is the customer performing? (authorized, fraudulent, or potentially fraudulent)
  2. Collect historical data
  3. Train the regression analysis model
  4. Make predictions for unknown values
- logistic regression is a classification algorithm. It cannot predict actual values for continuous data. It can answer questions like "Will the price of rice increase by 50% in 10 years?"

# Logistic Regression

- Logistic regression estimates the probability of an event occurring.
  - Example: voted or didn’t vote, based on a given dataset of independent variables.
  - Outcome is a probability, dependent variable is bounded between 0 and 1
- Also known as logit model.
- Often used for classification and predictive analytics.
- A logit transformation is applied on the P(Success)/ P(Failure). [log odds, or the natural logarithm of odds].
- Logistic regression does require large sample size.

# Logistic Regression Equation

$$\text{Logit}(pi) = 1/(1+ \exp(-pi))$$

$$\ln(pi/(1-pi)) = \text{Beta}_0 + \text{Beta}_1*X_1 + \ldots + B_k*K_k$$

- logit(pi) is the dependent variable and x is the independent variable.
- Beta parameter is estimated via maximum likelihood estimation (MLE).
- Process:
  1. MLE tests different values of beta through multiple iterations to optimize for the best fit of log odds.
  2. Iterations produce the log likelihood function, and logistic regression seeks to maximize this function to find the best parameter estimate.
  3. Once the optimal coefficient is found, the conditional probabilities for each observation can be calculated, logged, and summed together to yield a predicted probability.

- For binary classification, a probability < 0.5 will predict 0 while a probability > 0 will predict 1.
- Evaluate "How well the model predicts the dependent variable?"[goodness of fit].

https://www.ibm.com/topics/logistic-regression

# Logistic Regression

- Logistic regression needs an adequate sample to represent values across all the response categories.
- Without a larger and representative sample, the model may not have sufficient statistical power to detect a significant effect.
- Three types of logistic regression models, which are defined based on categorical response.
  - Binary logistic regression
  - Multinomial logistic regression
  - Ordinal logistic regression

# Logistic Regression Types

- Binary logistic regression: Dependent variable has only two possible outcomes (e.g. 0 or 1).
  - Ex: Predicting if an e-mail is spam or not spam.
  - Most common classifiers for binary classification.

- Multinomial logistic regression: Dependent variable has three or more possible outcomes but no specified order.
  - Example: Movie studios want to predict what genre of film a moviegoer is likely to see, to market films more effectively.

- Ordinal logistic regression: Dependent variable has three or more possible outcomes but with specified order.
  - Example: Grading scales from 0 to 12.

# Logistic Regression and ML

- Logistic regression attempts to distinguish between classes (or categories).
- It cannot generate information [such as image] of the class that it is trying to predict (e.g. a picture of a cat) [Discriminative model].
- Within ML, the negative log likelihood used as the loss function, using the process of gradient descent to find the global maximum.
- Logistic regression can also be prone to overfitting, when there is a high number of predictor variables within the model.
- Regularization is typically used to penalize parameters large coefficients when the model suffers from high dimensionality.

https://www.ibm.com/topics/logistic-regression

# Logistic Regression

- Logistic regression is used to estimate the relationship between a dependent variable and one or more independent variables.
- It is used to make a prediction about a categorical variable versus a continuous one.
- A categorical variable can be true/false, yes/no, 1/0.
- Logit function transforms the S-curve into straight line.
- Use cases:
  - Fraud detection: Can help to identify data anomalies, which are predictive of fraud.
  - Disease prediction: Can be used to predict the likelihood of disease.
  - Churn prediction: Specific behaviors may be indicative of churn in different functions of an organization.