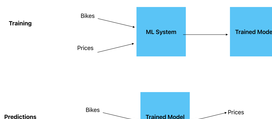
Machine Learning Fundamentals

What is Machine Learning ?

Learn from Examples

Inputs & Desired Output

How to conversion between input to output



Traditional Software Engineering

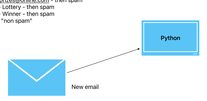
Spam Detection

Spam

Non - Spam

- Is sender = prizes@online.com - then spam
- If title says - Lottery - then spam
- If title says - Winner - then spam
- Otherwise - "non spam"

- Is sender = promotions@online.com - then spam
- If title says - 100% discount
- If title says - Great chance
- promotions.com
- Deposit
- Otherwise - "non spam"



Spam

Non - Spam

Types of Machine Learning

Supervised

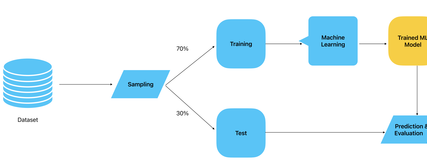
Labelled Dataset (X&y)

Un-Supervised

Features (X)

Reinforcement

Flow of Machine learning



Linear Regression

Supervised ML :- Regression

When predicting a Real value (y)

Logistic Regression

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

When predicting a Binary value (y)

Gradient Descent

Take the gradient of loss function wrt parameters

w(new) = w(initial) - (learning_rate)*grad_w

Evaluation Metrics for Classification

Confusion Matrix

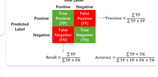
Y_true	prediction_probability	Threshold = 0.5	Threshold = 0.75
1	0.9	1	1
1	0.65	1	0
1	0.7	1	0
0	0.4	0	0
1	0.85	1	1
0	0.95	1	1
0	0.1	0	0
1	0.82	1	1

AOC - TPR Vs FPR
AUC - Area Under the Curve

Precision

Recall

F1- Score



Overfitting Vs Underfitting

Overfitting

- Data in uncleaned & contains noise
- Model is having high variance
- Model is too complex

- Using K fold Cross Validation
- Regularisation technique
- Training with more data
- Ensemble Technique

Underfitting

- Data in uncleaned & contains noise
- Model has high bias
- Model is too simple

- Increase the number of feature
- Increase model complexity
- Increase the duration of training

Cross Validation

- Using K fold Cross Validation
- Hold-out
- Leave-one-out
- Leave-p-out

Hyper-parameter Tuning Techniques

- Manual Method
- GridSearchCV
- Random Search CV

KNN Algorithm

- No parameters to learn during the training
- Lazy Learning

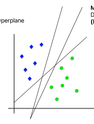
Support Vector Machine(SVM)

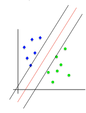
Find Hyperplane that best separates the two classes

- Linear SVM
- Non-Linear SVM

Support Vectors:
Data Points that are closest to the hyperplane

Margin:
Distance between Hyperplane and Support vectors
(Hard margin & soft margin)





Kernel Functions

- Polynomial Kernel
- Sigmoid Kernel
- RBF Kernel
- Bessel Function Kernel
- Anova Kernel

Ensemble Learning

Wisdom of crowd

Ensemble - group of predictors

1. Voting Classifier
2. Bagging Classifier
3. Random Forest
4. Boosting

Unsupervised Learning

Clustering

- Exclusive Clustering - K Means
- Overlapping Clustering - fuzzylc-means clustering
- Hierarchial Clustering

KMeans Clustering

- Step 1: Select the Number of Clusters, k. ...
- Step 2: Select k Points at Random. ...
- Step 3: Assign Each Observation ...
- Step 4: Calculate the Centroid of Each Cluster. ...
- Step 5: Assess the Quality of Each Cluster. ...
- Step 6: Repeat Steps 2-5.

Inertia/SSE

Silhouette Score

Hierarchical Clustering

- Agglomerative Clustering (Bottom level)
- Division Cluster (Split from Root - Top Level)

1. *Complete Linkage Clustering - Max distance two clusters*
2. *Single Linkage Clustering - Min possible distance between two clusters*
3. Mean Linkage Clustering - *mean of pairwise distance for points of 2 clusters*
4. Centroid Linkage Clustering - *distance between 2 cluster centroids*