All machine learning

-Linear Regression

-Logistic Regression: classification, usually when having binary classi prob

-Decision Tree:ask question to data and split them,aim is increase predictiveness(split that increase purity of a node).Does not require to normalize or scale feature. Suitable to work on a mixture of feature data type. Negative:it is prone to overfiting

-Random Forest: ensemble of decision tree , build by bagging,DT are use as parallel astimator.For classify,result is based on majority vote from each decision tree. For regression,prediction of leaf node is the mean value of the target in that leaf. Reduce the risk of overfitting ,high accuracy than 1 DT,time does not bottlenecl(it run parallel). Depend on uncorrelated DT –use bootstrapping(randomly selecting sample from training data with replacement) and randomness(selecting feature ramdomly for each decision tree in random forest). Not a good choice for high dimensional dataset

-GBDT(gradient boosted decision tree): boosting method to combine individial decision tree. Boosting:combine a learning algorithm in series to have strong learner from many sequential week learner(in this case ,week learner is decision tree). Each week learner attempt to minimize the error of previous learner .not like bagging , it not include bootstrap sampling, every time a learner add , it fit on modified version of initial data).Since learner add sequentially , boosting method learn slowly, but it perform better.Efficient on both classify and regression ,more accuracy . GBDT can handle mix of feature and don’t need pre-processing ,but require careful tuning hyperparameter.

-SVM(support vector machine):distinguish class by draw a decision boundary(effective when number of dimension more than number of feature)

-Naïve Bayes:features are independent each other, calculate probability of class given a set of feature,very fast (use when speed need more than high accuracy)

-kNN(k nearest neighbor):solve both classify and regression task,(value of class/datapoint is determined by the data around it (kNN classifer:determines class of data point by majority voting principle;kNN regression take the mean value of five closest point).kNN become slow when number of data point increase,and not memory efficency(cuz model need to store all data point) , it sensitive to outlier.simple,suitable for low dimension data , not really “learn” anything.

-k-means clustering:Unsupervised learning. clustering:a way to group a set of data point in a way that similar data point group together.it try to find underlying structure of data point. It partition data into k cluster in a way that data point have same cluster are similar , and different cluster are futher apart.it built on expectation maximization algorithm.Relatively fast .BUT number of clustet must be predetermined, it don’t work well if there is nonlinear structure seperating group in the data

Partition based,hierarchical clustering(kNN, kmeans) technique highly efficent with normal cluster

Density based technique are more efficent with arbitrary technique/detecting oulier .

- DBSCAN clustering(density based spatial clustering of application with noise):able to find arbitrary shape cluster and cluster with noise. A point belong to a cluster if it close to many point from that cluster(two parameter: eps (distance specifies the neighborhood,two point are neighbor if distance<eps), minPts(minimum datapoint to define a cluster)🡪 core point(at least minPts point in radius eps), border point(reachable from core point ,less than minPts point in radius eps),outlier(not core point, not reachable form any core point) . Not require to specify a number of cluster beforehand ,able to detect outlier. Determine eps is need domain knowledge.

-PCA(principle component analysis): unsupervised algorithm. dimension reduction algorithm(derive new feature from existing feature with keep much information as possible ). It derive new feature(called principle component) by finding the relation among the feature in a data set, the order of PC determined by fraction of variance of original dataset they explain .It aim to explain variance in original data as mush as possible using less feature.advantage of PCA is significant amount of variance of original data is retain using smaller number of feature ,PA is order according to amount of variance they explain. It widely use as a pre-processing step for supervised learning.