Questions & Answers for Lecture 1

Q1: What is the similarity and difference between k-NN and k-means?

A1:
Similarity: 1) Both methods have k in the name. In k-means, k refers to the number of clusters, while in k-NN the number of neighbors. 2) Both methods involve computing distances in input space and assigning data points to a set of nearest 'prototype points'.

Difference: 1) k-NN is a supervised machine learning while k-means clustering is an unsupervised machine learning, which means the dataset must be labeled if you want to use k-NN. 2) k-NN can be used for the classification and the regression problems as well. However, it is more widely used in classification problems in the industry. k-means is used for the clustering.

Q2: How to improve the efficiency when implement the k-NN?

A2:
When implementing the k-nearest neighbor method, the simplest method is linear scan. At this time, the distance between the input sample and each training sample is calculated. But when the training set is large, this method is very time-consuming(O(n*d)). The solution is to use the kd tree to improve the efficiency of k-nearest neighbor search.

The kd tree is a tree data structure that stores sample points in a k-dimensional space for quick retrieval. It is a binary tree, which represents a division of k-dimensional space. The process of constructing a kd tree is equivalent to the process of continuously dividing the k-dimensional space with a hyperplane perpendicular to the coordinate axis. Each node of the kd tree corresponds to a k-dimensional super-rectangular area.

With this method, the time complexity of k-NN decrease to O(log(n)*d)

(https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote16.html)