K-Means (1st Machine Learning Workshop Topic)

Please Download Anaconda while we talk

Housekeeping

- No recording please
- Recommended youtube video
  - Dr. Andrew Ng
  - ELE 888 Intelligent Systems

Machine learning:

- Supervised Learning
- Unsupervised learning
- Regression

K-Means Introduction

- Unsupervised machine learning algorithm
- Group data into k clusters
- K-means Algorithm
  - **Lloyd (1957): the standard method**
  - Forgy (1965): used by initialization
  - Hartigan and Wong (1979): Better performance
  - MacQueen (1967): Another well-known one

K-Means Standard Algorithm

Step 1: Initialize K Centroids (three options)

- Randomly select the initial centroid locations
- Use the first K data points as the centroid
- Randomly select K data points

Step 2: Assign each data point to a cluster based on the proximity to the centroid of the clusters

$$\arg min_k \; Dist(x_i, C_k)$$

Step 3: Recalculate the new centroids

$c_k = \frac{1}{N_k}\sum x_i$ for all x_i belong to cluster k, N_k is the number of data points for the cluster

Step 4: Stop when converge, else go back to step 2

K-means characteristics:

- Only works with numeric data
- Does not guarantee optimal, ie. Within cluster sum of square is not minimum

$$\arg\min \sum_{k=1}^{K} \sum_{x_i \in C_k} (x_i - c_k)^2$$

- Distance measure: Euclidean distance, L2-norm

$$dist = \sqrt{\left(x_i^1 - C_k^1\right)^2 + \left(x_i^2 - C_k^2\right)^2 + .. + \left(x_i^N - C_k^N\right)^2}$$

Please note that there is no difference if you take the square root or not for comparison. We are only looking at the positive values.

- If K is undefined, use Elbow Test to find K.

Reference:

- Forgy, E. W. (1965) Cluster analysis of multivariate data: efficiency vs interpretability of classifications. Biometrics. **2:**768–769.
- Hartigan, J. A. and Wong, M. A. (1979). A K-means clustering algorithm. Applied Statistics. **28**:100–108.
- Lloyd, S. P. (1957, 1982) Least squares quantization in PCM. Technical Note, Bell Laboratories (1957). IEEE Transactions on Information Theory (1982). **28**:128–137.
- MacQueen, J. (1967) Some methods for classification and analysis of multivariate observations. 5th Berkeley Symposium on Mathematical Statistics and Probability. **1**: 281–297.

Next Week

- Implement K-Means on Python
- Compare your K-Means with sklearn.cluster's KMeans.
- Please refer to sample code from scikit-learn.org:
  - **https://tinyurl.com/ybgw82ll**
- Excellent free Python learning websites:
  - Code Academy
  - Google for Education
  - Python official website