COMP20008 2020 SM1 Workshop Week 8: Clustering and Linear Regression

Clustering

1. Consider the 1-dimensional data set with 10 data points {1,2,3,...10}. Show the iterations of the k-means algorithm using Euclidean distance when $k = 2$, and the random seeds are initialized to {1, 2}.

2. Repeat Exercise 1 using agglomerative hierarchical clustering and Euclidean distance, with single linkage (min) criterion.

3. After completion of the k-means algorithm, we may compute a quality measure for the resulting clustering, known as SSE (sum of squared errors). SSE is the sum of distances of objects from their cluster centroids

$$SSE = \sum_{i=1}^{k} \sum_{x \in c_i} distance(x, \overline{c_i})^2$$

   - Do you think it is more desirable for a clustering to have high SSE or more desirable for it to have low SSE? Why?
   - As the number of clusters increases, would you expect SSE to increase or decrease? Why?
   - Suggest at least one other method you could use in place of SSE to evaluate the quality of clustering.

Regression and Clustering visualisation

Questions 4 – 7, use the Python notebook *workshop-week8-2020S1-regression-VAT.ipynb*.