Assignment 2

1. Clustering
   1. No.

First we need to manually choose the K, if we do not have a key insight of the dataset, choosing K maybe different to find a better model.

More importantly, we will introduce randomness when initializing the K points as centroids, and this may lead to stuck to different local minima.

* 1. Yes.

Result is independent of the initialization, since we will start at the state of “each point being a cluster, recursively merge them, using a deterministic measurement”.

* 1. No.

Because we always choose Euclidean distance to determine the clusters in K-means. When facing non-linearly separable dataset, it cannot deal with non-convex shape, and yields many misclustering results.

1. Classification
   * 1. False, it indicates underfitting, since the model learn a little knowledge of data.
     2. True, the acc gap between training and test data is larger.
     3. True, we want the gap to be 0 to justify the performance doesn’t change.
   1. I will choose 15. The performance is relatively stable from 10 to 20. If we choose another subset as test data, the performance maybe better, since we cannot guarantee the data is balanced at the first choice. Further, 15 neurons also provide a safe band away from overfitting and underfitting.
   2. Hold-out: K-fold

Regularization: Dropout, Early stopping, Weight Sharing

Dataset: More data, data augmentation