STA 4365 HW 3

due Tuesday April 5 by 11:59PM

Problem 1: (5 points) In this problem, you will show that the K-means algorithm is guaranteed to converge to a local minimum. Do this by using the following steps:

(a) Let $\{X_i\}_{i=1}^n$ be observed samples in $\mathbb{R}^p$. Recall the K-means objective

$$\mathcal{L}(\mathcal{W}) = \sum_{k=1}^{K} \frac{1}{|W_k|} \sum_{i,j \in W_k} \|X_i - X_j\|^2.$$

$\mathcal{W} = (W_1, \dots, W_K)$ is a set a clusters, where W_k is the set of i such that X_i is in cluster k. Let $\mathcal{W}^{(t)}$ denote the clusterings after t steps of K-means. Show that

$$\mathcal{L}(\mathcal{W}^{(t+1)}) \leq \mathcal{L}(\mathcal{W}^{(t)})$$

Hint: Rewrite $L(\mathcal{W})$ using the mean vector like we did in class (see Equation 12.18 in Chapter 12 of the ISLRv2 textbook), and show that each step of K-means must decrease the objective.

(b) Argue that the K-means algorithm will eventually exactly converge to a local minimum. *Hint:* Use the fact the the set of possible K-means outcomes is finite.

Problem 2: (7 points) ISLRv2 Chapter 12 Problem 10.

Problem 3: (8 points) In this problem, you will build use unsupervised learning methods to analyze the Fashion-MNIST dataset. The Fashion-MNIST dataset provided here consists of 60,000 images of clothing. Each observation is a 28×28 pixel grayscale image, so each observation has $28^2 = 784$ dimensions. There are 10 types of clothing in the dataset. The first column of the data matrix is the label of the observation, and all other columns are pixel intensities. You can find more details here.

(a) Load the dataset. For this problem, you don't need to rescale the features, because all pixels are already on the same intensity scale. Separate the image and label information and randomly select 5000 observations to use in the model (the selection of random observations is to reduce computational cost).

(b) Learn a K-means model using the K-means++ initialization method with $K = 10$ clusters from your 5000 observations. Visualize 10 images from each cluster.

(c) Learn a 2D t-SNE embedding of the observed data. Use a perplexity of 30, and do PCA pre-processing to reduce the feature size to 50 dimensions before learning the t-SNE model. Using your embeddings, make two plots with identical embedding dimensions color codings: one where color shows the ground truth labels, and one where color shows your K-means cluster labels from the previous part. Describe how well the clusters from the previous part reflect the ground truth labels and the geometric relations identified by t-SNE.