Math189R SU18
Homework 8
Thursday, May 24, 2018

Feel free to work with other students, but make sure you write up the homework and code on your own (no copying homework *or* code; no pair programming). Feel free to ask students or instructors for help debugging code or whatever else, though.

The starter files for problem 1 and extra credit can be found under the Resource tab on course website. Please print out all the graphs generated by your own code and submit them together with the written part, and make sure you upload the code to your Github repository.

1 (K-means) In this problem, we will implement the k-means algorithm and separate 5,000 2D data points into different number of clusters.

Let $X = x_1, x_2 \ldots, x_m$ be the data points, and let k be the number of clusters. The k-means algorithm is summarized as following:

1. Randomly initialize k cluster centers, $\mu_1, \mu_2 \ldots \mu_k$, in the feature space.
2. Calculate the distance between each data points and the cluster centers.
3. Assign each data point to the cluster center c whose distance between this data point is the minimum of all the cluster centers, namely,

$$c_i = \arg\min_j ||x_i - \mu_j||^2$$

4. Update each cluster center to be

$$\mu_j = \frac{\sum_{i=1}^m 1\{c_i = j\}x_i}{\sum_{i=1}^m 1\{c_i = j\}}$$

5. Repeat step 2 - 4 until convergence or exhausted.

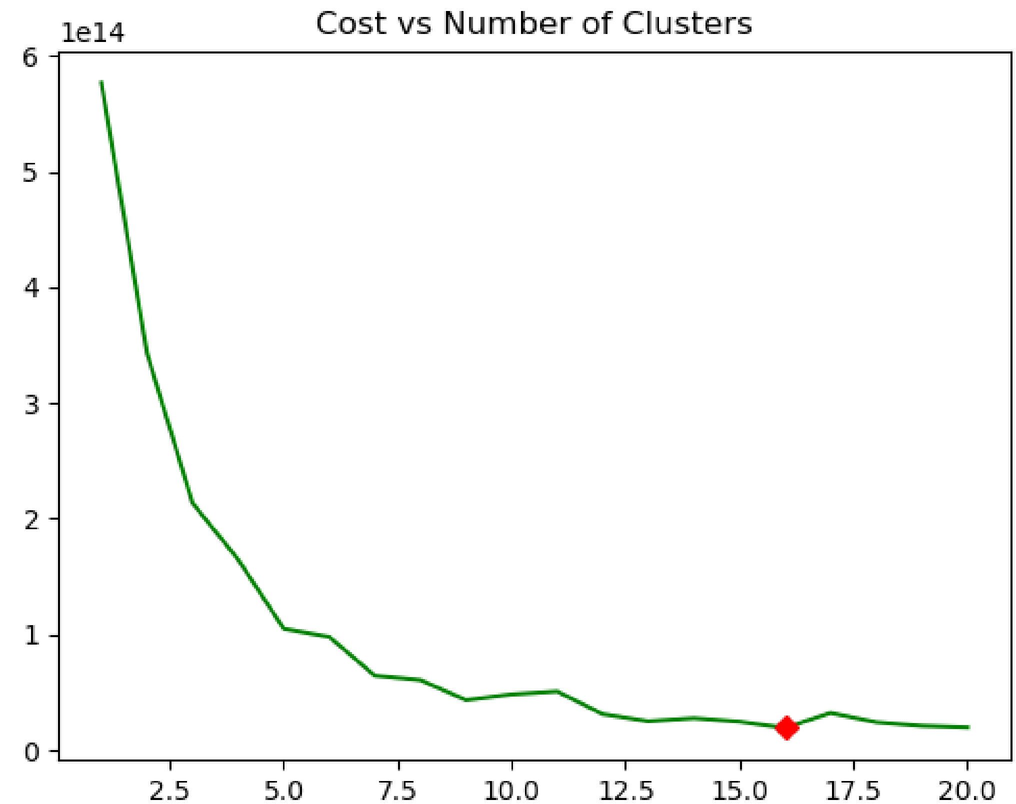
The objective (cost) function is defined as

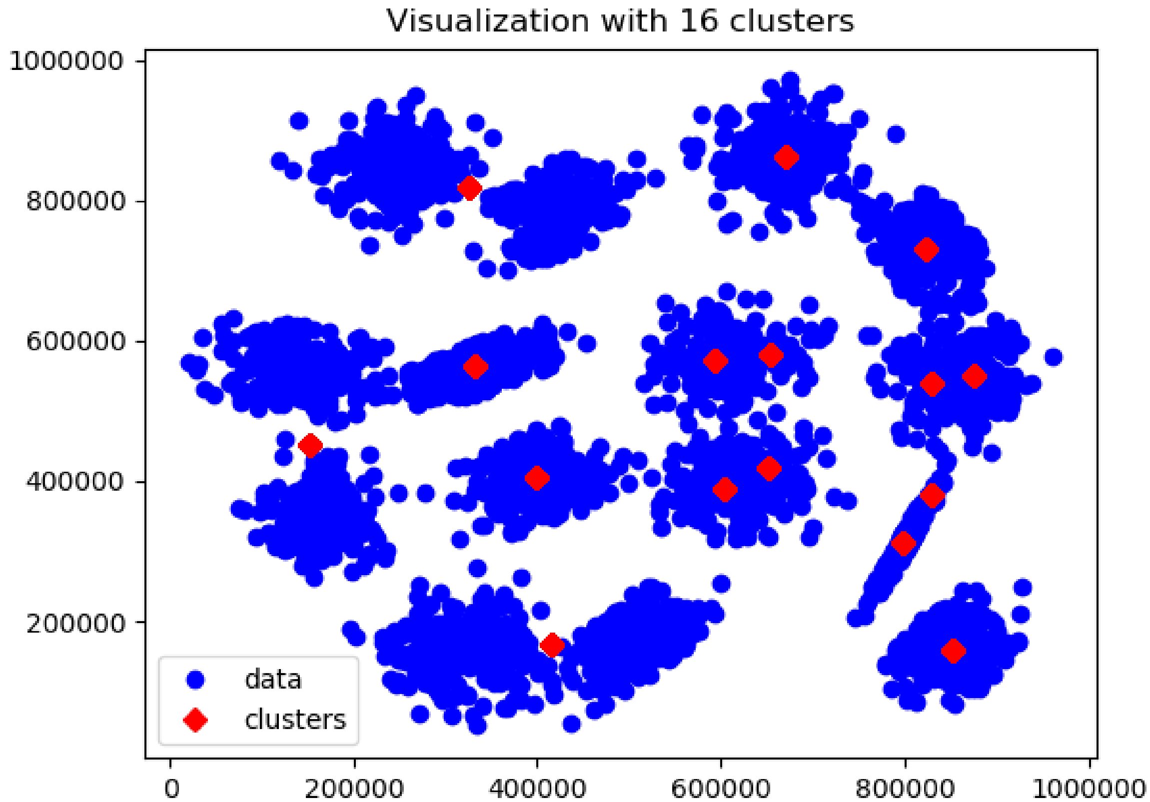
$$J(c, \mu) = \sum_{i=1}^m ||x_i - \mu_{c_i}||^2$$

In this assignment, you will first implement the k-means cost function and the algorithm. Then, for $k = 1, 2 \ldots, 20$, find the number of clusters with the optimal cost and produce a plot of the relationship between the cost and the number of clusters. Then, visualize the data points and the cluster centers on the optimal number of clusters.

Notice that the k-means algorithm might yield different results based on the randomness of the initialization of cluster centers.

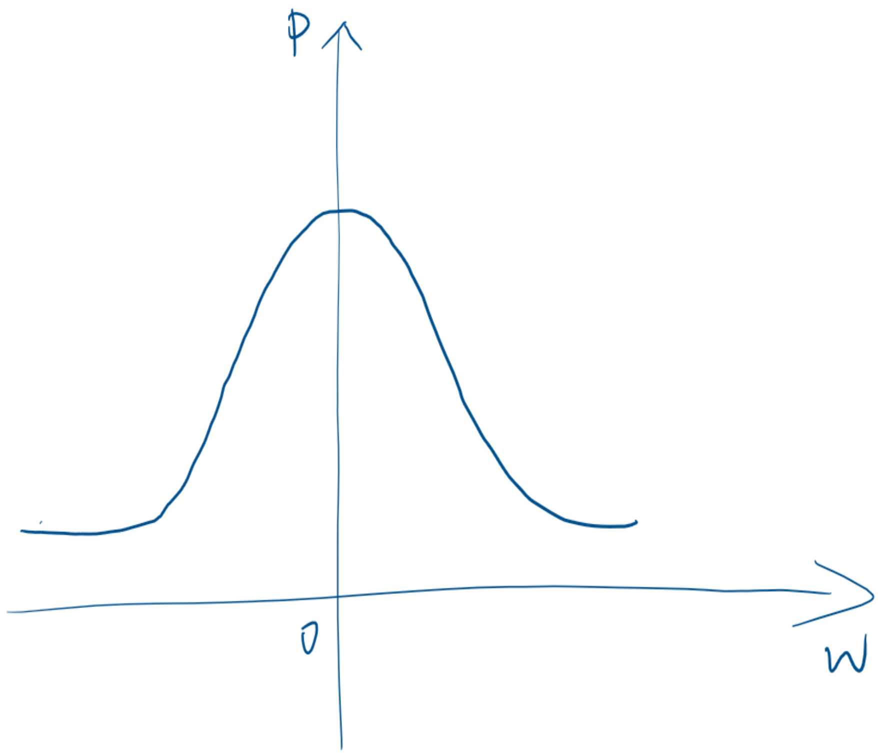
From the graph below, we can see that the optimal cluster number is k = 16. The plot is attached below:

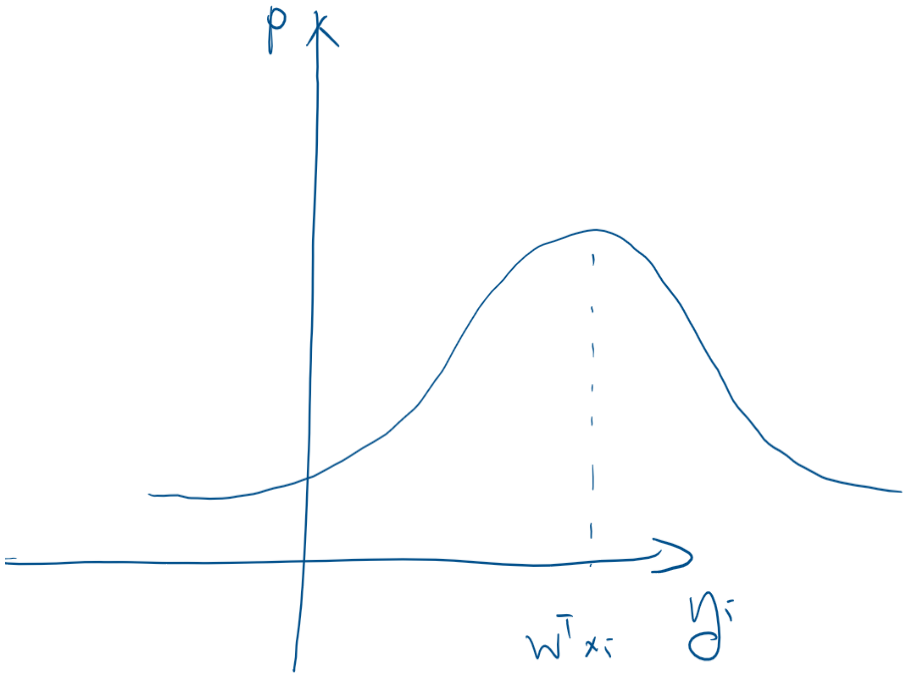


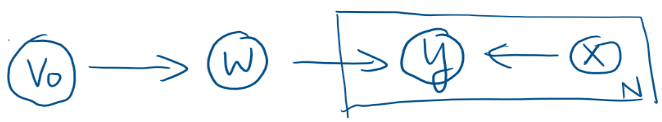

Visualization with 16 clusters
1000000
800000
600000
400000
200000
0
200000
400000
600000
800000
1000000
data
clusters

■

Extra Credit (Linear Regression) Consider the Bayesian Linear Regression model with the following generative process:
(1) Draw $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_0)$
(2) Draw $\mathbf{y}_i \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}_i, \sigma^2)$ for $i = 1, 2, \ldots, n$ where σ^2 is known.
Express this model as a directed graphical model using Plate notation. Is $\mathbf{y}_i$ independent of $\mathbf{w}$? Is $\mathbf{y}_i$ independent of $\mathbf{w}$ *given* $\mathcal{D} = \{\mathbf{x}_i\}$? Support these claims.







y is dependent on w in this case no matter whether we condition on x or not. ∎