EEP 596: Adv Intro ML || Conceptual Assignment 4

Dr. Karthik Mohan and Xiulong

Univ. of Washington, Seattle

Feb 1, 2022

Problem 1

1d k-means

Let's look at points in 1 dimension - I.e. numbers on the real line. Consider the following sequence of points:

$$-5, 10, -3, 4, 7, -1, -2, 3, 5, -0.5$$

Assume we are looking to apply k-means on this data set and start with two centroids, $\mu_1 = -3, \mu_2 = -1$. Going through k-means iteratively, what would be the centroids of the clusters you end up with when you are done? Pick the option below that's the closest to the k-means solution. (Highly recommended for you to work this out on paper and follow the k-means procedure to improve your understanding). a) -3.5,3 b) 1,3 c) -2.25,-1 d) -2.25,6

Problem 2

Clusters

You implement 2 different ML models for clustering on a data set and obtained a set of clusters A and a set of clusters B. Assume that both of them have 3 clusters each. The inter-cluster distances for clusters A and those for clusters B is as follows:

$$\begin{array}{rcl} \mathrm{A} & : & 3, 5, 4 \\ \mathrm{B} & : & 2, 7, 6 \end{array}$$

The max intra-cluster distance for cluster A is 4.75 while that for cluster B is 3. Which ML model based on the Dunn Index would you say produces a better set of clusters - Model A or Model B? Also mention the Dunn Index of the best model.

Problem 3

Agglomerative Clustering

Here we will generate data points in donut shapes and cluster them through k-means and agglomerative clustering. Our data set will be composed of two donuts. For the bigger donut - Generate 100 random data points using the following procedure in Python:

- Each data point can be described using a (r, θ) representation where r is a radius and θ is an angle. This is the polar coordinates representation. Once you have an r and θ, you can generate the feature vector for the point as $[r\cos(\theta), r\sin(\theta)]$.
- For each data point, sample r from a Normal distribution, i.e. $r \sim \mathcal{N}(10, 0.2)$ where $\mathcal{N}$. Also sample a θ $U[0, 360]$, i.e. uniformly at random between 0 and 360 degrees.
- These 100 pairs of (r, θ) that you just generated give you 100 2-d data points and this forms the bigger donut.

Problem 3 contd

Agglomerative Clustering

- Repeat the above procedure to produce 100 data points that form the smaller donut. For the smaller donut, use $r \sim \mathcal{N}(8, 0.2)$.
- Use scatter plot the smaller and larger donut and color them differently using matplotlib.
- Apply k-means and agglomerative clustering (choose a linkage function that makes sense) to the data set.
- Plot the clusters from k-means and those from agglomerative clustering. Which one is able to separate the donuts?
- In your submission, include all of the code for the steps outlined in this problem.
- Hint: Numpy has a standard normal library under `np.random.randn`. This gives you a z. Note that $z = \frac{r-\mu}{\sigma}$.

Problem 4

Product Recommendations at Sambazon

You are tasked with building a recommendation system that recommends new and upcoming products to customers of your company, Sambazon. Your goal is to implement a ML model that can help boost the sales of these new products by 20% over vanilla un-personalized recommendations of these new products. The new products are of 2 different types: One category is fashion where customers preferences keep changing with time and what's popular today maybe obsolete tomorrow. The other category are products that see regular and repeated buys by customers if they find a match. The new products come with text descriptions that are accurate in describing what they do and how they can help the customers.

Problem 4 (contd)

Product Recommendations at Sambazon (contd)

Also since we are talking about brand new products - You can assume that these products have seen zero to very limited sales per day so far on Sambazon and the potential customers that might like these products if identified correctly, number in millions. Given all of what you have seen in our course so far, provide a detailed ML solution that can help address the problem you want to solve. Mention what your data set looks like, what ML approach(es) you would take, what feature engineering would you do, what specific ML models you would use and what are some practical considerations you thought of for your picks. Also mention evaluation metrics that you would use to evaluate your approach. (Note: This kind of a problem is not something hypothetical we are looking at for the sake of this assignment - It might be one that you get at your next ML interview that you may be asked to brainstorm!!)