Northeastern University, Khoury College of Computer Science

CS 6220 Data Mining — Assignment 4

Due: February 1, 2023(100 points)

YOUR NAME
YOUR GIT USERNAME
YOUR E-MAIL

K-Means

The normalized automobile distributor timing speed and ignition coil gaps supplied are from production F-150 trucks over the years of 1996, 1999, 2006, 2015, and 2022. We have stripped out the labels for the five years of data. Each sample in the dataset is two-dimensional, i.e. $\mathbf{x}_i \in \mathbb{R}^2$, and there are $N = 5000$ instances in the data.

Vanilla k-Means

In this part of the homework, we'll take a look at how we can identify patterns in this data despite not having the labels. We'll start with the simplest approach, the k-Means unsupervised clustering algorithm.

1. Implement a simple k-means algorithm in Python on Colab with the following initialization:

$$\mathbf{x}_1 = \begin{pmatrix} 10 \\ 10 \end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix} -10 \\ -10 \end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \mathbf{x}_4 = \begin{pmatrix} 3 \\ 3 \end{pmatrix}, \mathbf{x}_5 = \begin{pmatrix} -3 \\ -3 \end{pmatrix}, \tag{0.1}$$

   You need only 100 iterations, maximum, and your algorithm should run very quickly to get the results. In order to maintain consistency between submissions, use a random seed of 27. You can do this with

```
>> numpy.random.seed(seed=27)
```

2. Scatter the results in two dimensions with different clusters as different colors. You can use **matplotlib**'s **pyplot** functionality:

   ```
   >> import matplotlib.pyplot as plt
   >> plt.scatter(<YOUR CODE HERE>)
   ```

3. You will notice that in the above, there are only five initialization clusters. Why is $k = 5$ a logical choice for this dataset? After plotting your resulting clusters and. What do you notice?

With Production Information

Very often, it is possible to obtain additional information about the collected data. This sometimes allows us to define a new mathematical operators (including distances). In this part of the homework, we'll look at how to use this information to improve our modeling with an understanding of how two features in each sample are related.

A common distance metric is the *Mahalanobis Distance* with a specialized covariance.

$$d(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T P^{-1} (\mathbf{x} - \mathbf{y}) \tag{0.2}$$

where $\mathbf{x}$ and $\mathbf{y}$ are two points of dimensionality m (2 in this case), and $d(\mathbf{x}, \mathbf{y})$ is the distance between them. In the case of the F150 engine components, P is a known relationship through Ford's quality control analysis each year, where it is numerically shown as below:

$$P = \begin{pmatrix} 10 & 0.5 \\ -10 & 0.25 \end{pmatrix} \tag{0.3}$$

4. Implement a specialized k-means with the above Mahalanobis Distance. Scatter the results with the different clusters as different colors. What do you notice? You may want to precompute P^{-1} so that you aren't calculating an inverse every single loop of the the k-Means algorithm.

5. Calculate and print out the *first* principle component of the aggregate data.

6. Calculate and print out the *first* principle components of *each cluster*. Are they the same as the aggregate data? Are they the same as each other?

Naïve Bayes, Bayes Rules

The original performance of acoustic classification for Parkinsons Disease leverages speech recordings from controlled subject responses from variety of questions. The task in the competition was to detect whether or not a person X had Parkinsons disease from a sampling of data. As of 2018, the state of the art classifiers have achieved 90% correct classification on a held out dataset, both for subjects who had Parkinsons and those who did not (at equal rates). So, when classifier Y sees person X, it works correctly 90% of the time.

7. Let's say that we run a clinic. This clinic leverages this classifier, which has 90% accuracy. Also, let us say that we know that our current patient load is that 10% of the population have Parkinsons and 90% of the population do not. Let's also say that we're seeing patient X, and the classification algorithm has detected that they have Parkinson's disease. What's the probability that indeed X has Parkinson's disease?

   Come up with the numerical solution, and show your written work.

Gradient Descent - Logistic Regression

Yann LeCun, the Director of Facebook Research and one of the fathers of deep learning neural networks, got his start with the MNIST Dataset, which is widely regarded as the ”Hello World” of Computer Vision. The MNIST dataset can be downloaded at LeCun's website.
Logistic regression's cost function is typically binary cross-entropy.

$$\mathcal{L}(W,b) = \sum_i y_i \log h_{W,b}(\mathbf{x}) + (1 - y_i) \log \left(1 - h_{W,b}(\mathbf{x})\right)$$

where

$$h_{W,b}(\mathbf{x}) = \sigma\left(W^T \mathbf{x} + b\right)$$

and

$$\sigma(z) = \frac{1}{1 - \exp(-z)}$$

8. Prove that the derivative of the sigmoid function is $\nabla_z \sigma(z) = \frac{\partial \sigma(z)}{\partial z} = \sigma(z)\left(1 - \sigma(z)\right)$.
9. Calculate the gradient of $\mathcal{L}$ with respect to W.
10. Calculate the gradient of $\mathcal{L}$ with respect to b.
11. Build a logistic regression classifier that is able to classify MNIST.