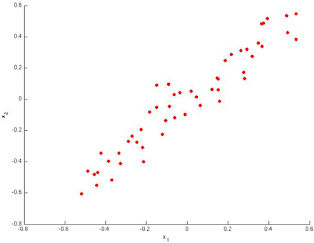
Machine Learning Module-2 Module Bank

PART-A

1. Consider the following 2D dataset



Compute the following:

* 1. Which correspond to possible values that PCA may return for  (the first Eigen vector / first principal component)?
  2. Justify a reasonable way to select the number of principal components k?
  3. Suppose someone tells you that they ran PCA in such a way that “95% of the variance was retained.” What is an equivalent statement to this?
  4. List the recommended applications of PCA.

2.

1. You work for a music streaming service and want to use supervised machine learning to classify music into different genres. Your service has collected thousands of songs in each genre, and you used this as your training data. Now you pull out a small random subset of all the songs in your service. What is this subset called?
2. You work for a hospital that is tracking the community spread of a virus. The hospital created a smartwatch app that uploads body temperature data from hundreds of thousands of participants. What is best technique to analyze the data?
3. The new dataset you have just scraped seems to exhibit lots of missing values. What action will help you minimizing that problem?

3.a.Suppose we have three cluster centroids ,  and.



Furthermore, we have a training example . After a cluster assignment step, what will  be?

b. You want to identify global weather patterns that may have been affected by climate change. To do so, you want to use machine-learning algorithms to find patterns that would otherwise be imperceptible to a human meteorologist. What is the place to start?

4.

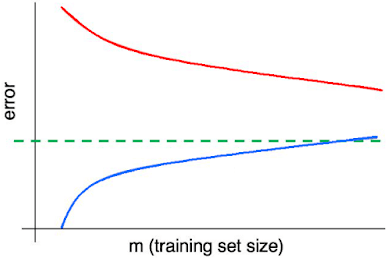
1. Suppose you have trained an anomaly detection system for fraud detection, and your system that flags anomalies when p(x) is less than ε, and you find on the cross- validation set that it is missing many fraudulent transactions (i.e., failing to flag them as anomalies). What should you do?
2. Suppose you have an unlabelled dataset . You run K-means with 50 different random initializations, and obtain 50 different clustering is of the data. What is the recommended way for choosing which one of these 50 clustering has to use?
3. Suppose you have trained an anomaly detection system for fraud detection, and your system that flags anomalies when p(x) is less than ε, and you find on the cross- validation set that it is mis-flagging far too many good transactions as fraudulent. What should you do?

5.

1. You are part of data science team that is working for a national fast-food chain. You create a simple report that shows trend: Customers who visit the store more often and buy smaller meals spend more than customers who visit less frequently and buy larger meals. What is the most likely diagram that your team created?
2. You work for an organization that sells a spam filtering service to large companies. Your organization wants to transition its product to use machine learning. It currently a list of 250, 00 keywords. If a message contains more than few of these keywords do, then it is identified as spam. What would be one advantage of transitioning to machine learning?

6.

1. Your supervisor asks you to create a machine learning system that will help your human resources department classify jobs applicants into well-defined groups. What type of system are you more likely to recommend?
2. Your company wants to predict whether existing automotive insurance customers are more likely to buy homeowners insurance. It created a model to better predict the best customers contact about homeowners insurance, and the model had a low variance but high bias. What does that say about the data model?
3. Suppose you are working machine-learning algorithm to classify/cluster the data samples of any kind of applications. You train a learning algorithm, and find that it has unacceptably high error on the train set. You plot the learning curve, and obtain the figure below. Identify which kind of errors from the following graph.



7. You are given a set of one-dimensional data points: {5, 10, 15, 20, 25, 30, and 35}. Assume that k=2 and first set of random centroid is selected as {15, 32} and then it is refined with {12, 30}.

* 1. Create two clusters with each set of centroid mentioned above following the k- means approach.
  2. Calculate the SSE for each set of centroid.
  3. Discuss the strengths and weakness of the k-means algorithm.

8. During a research work, you found 7 observations as described with the data points below. You want to create 3 Clusters from these observations using K-means algorithm. After first iteration, the clusters C1, C2, C3 has following observations:

C1: {(2, 2), (4, 4), (6, 6)}

C2: {(0, 4), (4, 0)}

C3: {(5, 5), (9, 9)}

If you want to run a second iteration then what will be the cluster centroids” What will be the SSE of this clustering?

1. In a software project, the team is trying to identify the similarity of software defects identified during testing. They wanted to create 5 clusters of similar defects based on the text analytics of the defect descriptions. Once the 5 clusters of defects are identified, any new defect created is classified as one of the types identified through clustering. Explain this approach through a neat diagram. Assume 20 defect data points which are clustered among 5 clusters and k-means algorithm was used.
   1. Can the performance of a learning model be improved. Elaborate your answer.
   2. Someone of your data science team recommends that you use decision trees, naive Bayes and K-nearest neighbour, all at the same time, on the same training data, and then average the results. Apply a suitable algorithm to predict the results with proper justification.
   3. Present the Decision Tree classifier algorithm. Using the algorithms and the following training data, find the class label for a test record {“Sohrab”, CGPA= High, Communication=Bad, Programming Skill=Bad}

|  |  |  |  |
| --- | --- | --- | --- |
| CGPA | Communication | Programming  Skill | Employee |
| High | Good | Good | Yes |
| Medium | Good | Good | Yes |
| High | Good | Bad | Yes |
| High | Good | Good | Yes |
| High | Bad | Good | Yes |
| Medium | Good | Good | Yes |
| Low | Bad | Bad | No |
| Low | Bad | Bad | No |
| Medium | Good | Bad | Yes |
| Medium | Bad | Good | No |
| Medium | Good | Bad | Yes |