Project: Answer the Following Interview Questions (7 total)

For each of the questions below, answer as if you were in an interview, explaining and justifying your answer with two to three paragraphs as you see fit. For coding answers, explain the relevant choices you made writing the code.

1. We A/B tested two styles for a sign-up button on our company's product page. **100** visitors viewed page **A**, out of which **20** clicked on the button; whereas, **70** visitors viewed page **B**, and only **15** of them clicked on the button. Can you confidently say that page **A** is a better choice, or page **B**? Why?

   Answers: We need to calculate the click-through-rate , which is number of visitors who clicked the button / the number of visitors who viewed the page. For page A, the click- through- rate is 20/100= 0.2 , while the click-through-rate is 15/70 =0.215 for page B.

   Then we can do a Hypothesis test: null hypothesis : A has a higher click-through rate than B has; alternative hypothesis : A doesn't have a higher click- through rate than B. Then we can calculate the p value by sampling from visitors who clicked pages A or B, if p value < 0.05, we can reject the null hypothesis, which means A has a higher click rate is not true, if p value > 0.05 we fail to reject the null hypothesis, which means A does have a higher rate.

2. Can you devise a scheme to group Twitter users by looking only at their tweets? No demographic, geographic or other identifying information is available to you, just the messages they've posted, in plain text, and a timestamp for each message.In JSON format, they look like this:{

   ```
   "user_id": 3,
   ```

```
 "timestamp": "2016-03-22_11-31-20",
 "tweet": "It's #dinner-time!"
}
```

Assuming you have a stream of these tweets coming in, describe the process of collecting and analyzing them, what transformations/algorithms you would apply, how you would train and test your model, and present the results.

Answers : We can use K-means clustering techniques to classify these tweets by active time in a day or topic. First of all, we can parse JSON format data into python dictionary and then to panda dataframe. Second , clean the data and remove unnecessary info in the data, for example, python’s nltk package can be used for removing ‘is ’, ‘are’, which are no meaning for classification.

Once data is ready, we can do clustering. We can import kmeans module from SKlearn and by setting k =2 , for example and combining features (timestamp and hashtag topics or key words) to get initial groups. As tweets coming more and more, then we can iterate over several values k to get more groups. The performance of k means model on training data can be represented by sum of distance. Testing on new tweets stream to see if the model generalize well. The results should be groups with different centroids.

3. In a classification setting, given a dataset of labeled examples and a machine learning model you're trying to fit, describe a strategy to detect and prevent overfitting.

Answers:

One way to detect overfitting is to compared the performance of training data and test data. If the dataset is fitted perfectly on the training data, which may be measure by a score , for example , close to 1 (R2), but in testing dataset, it performs poorly, the metrics shows a way below are score, let' say less than 0.5 (R2). Then we can say there is overfitting occurred. Overfitting also shows from learning curve. If training curve and testing curve merge at certain point but testing curve (the error) increased dramatically after that , and the gap between training and testing increased, that means overfitting occurred.

One way to prevent overfitting is to do k-fold cross validation. For example, let say k = 3, we split the dataset into 3 folds . The first round is doing the training on fold 1 and 2 and testing on fold 3. The second round is to do training on fold 2 and 3 and testing on fold 1 . The third round train on fold 1 and 3 and test on fold 2. The last step is do the average of the model against each of the folds and then finalize the model.

4. Your team is designing the next generation user experience for your flagship 3D modeling tool. Specifically, you have been tasked with implementing a smart context menu that learns from a modeler's usage of menu options and shows the ones that would be most beneficial. E.g. I often use **Edit** > **Surface** > **Smooth Surface**, and wish I could just right click and there would be a **Smooth Surface** option just like **Cut**, **Copy** and **Paste**. Note that not all commands make sense in all contexts, for instance I need to have a surface selected to smooth it. How would you go about designing a learning system/agent to enable this behavior?

   Answers: We can use reinforcement learning to learn from user's past usage routines and create an agent to collect data through trial and error in this modeling environment. In this example, reinforced with a reward for surface next step and negative reward for other operations. The agent will learn the user's preference by maximizing a cumulative reward. By this learning model, a smart context menu can be learned and created.

5. Give an example of a situation where regularization is necessary for learning a good model. How about one where regularization doesn't make sense?

Answers:

For example, when using decision tree for classification, tuning the depth of trees may increase the model's complexity . Thus adding a penalty term for tree's depth is necessary to prevent the model becoming too complex. When a tree's depth is too large, so its penalty term will be large too and therefore added to the training error, it would not be selected as the optimal model.

When a model is too simple , so there is no necessary to reduce complex

6. Your neighborhood grocery store would like to give targeted coupons to its customers, ones that are likely to be useful to them. Given that you can access the purchase history of each customer and catalog of store items, how would you design a system that suggests which coupons they should be given? Can you measure how well the system is performing?

Answers: I would like to do a clustering to group the customers based on their purchase history. For example, some who like cooking may be interested in vegetable and raw meat, and some who don't like cooking may interested in frozen food, processed food. Thus quantity in those categories of food may help grouping customer into different groups, which is relevant feature selection. If one feature is highly correlated to another, then it may not help to identify different cuterstom groups, this feature is not relevant.

Once we have data and feature ready, we can fit the data into clusters. By calculating the silhouette coefficient of data to the cluster assigned, it shows how similar the data to the cluster.

We can do A/B testing to evaluate the system's performance. We can selects two subsets of customer in the same cluster and send group A coupons according to its cluster and send group B regular coupons, not targeted coupons. Measure their purchases after sending coupons, if number of purchases in group A increased compared to group B's purchases , we can say the system works.

7. If you were hired for your machine learning position starting today, how do you see your role evolving over the next year? What are your long-term career goals, and how does this position help you achieve them?

Answers: If I were hired today, I will enhance my large scale data processing skills, such as spark in one year and be proficient in Uber's proprietary machine learning platform and predictive modeling to power decision making in uber's marketing strategy.

My long term goal is become highly knowledgeable and experienced in transportation system by taking more challenging projects and solving complex problems using data and besides, provide effective and innovative solution for people' ways of transportation , to change our daily life.