Machine Learning Interview Practice

First, find a job posting. Make sure that you at 70% meet the job requirement.

Chosen Job Description :

- *Data Scientist*
- *Machine Learning Engineer*
- *Data Engineer*

Questions are answered the way i would at an interview.

Questions

Question 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?

Solution 1

- For page A, given ***100 visitors*** , we have ***20 clicks.***
- For page B, given ***70 visitors***, we have ***15 clicks.***
- For simplicity, i denoted ***visitors*** *as* ***“v” clicks*** *as* ***“c”***

A conclusion cannot be made looking at this scenario. That will be too **bais.**
There is a good chance that if the test is to be run again, we might arrive at a different outcome.

So

- the PR(*of clicking a button for page **A** visit*) is ⅕
- The PR(*of clicking a button on page **B** visit*) is 3/14

The PR of no clicks on each page will be difference between 1 and each probability.

Combining the PR together :

- Total numbers of visits : 170 visitors (*i.e the addition of page A visitors ,**100** and page B visitors ,**70***)
- PR(*probability of clicking the button on Page A in total*) is 20/170 which is 2/17
- PR(*probability of clicking the button on Page B in total*) is 15/170 which is 3/34
- Total numbers of clicks is 35

Given this, I think the positioning of button on Page A has a greater probability of being clicked than the position of button on Page B.

Question 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!"
}
```
~~~

Solution 2.

Based on the assumption that i have stream of tweets coming in, there is no data shortage.

Collection of data

It is important to know how to collect the data you want. In this case, we need data collection on Twitter. To collect data on Twitter, my suggestion is to create an account and then go to Twitter API to create an app that allows you to collect Twitter data. Just make sure you don't ask too much, because there is a limit on how many times you can request Twitter data.

If you want to do a one-time collection of tweets, you should use the REST API.
If you want to do a continuous collection of tweets for a specific time period, you should use the streaming API.
It is required that you have a private key and token. For more information about getting information on Twitter, please checkout this Twitter Developer Search Tweets

NOTE :

- *The assumption is that we are not given the demographic information.*

Analysing the data.

We need to group the data we have. Since we have the geographical information and other identifying information we can do some clustering. Below are ways i will group the data :

Location :Location is one important way of grouping the data. One thing i will take into consideration is the Latitude and Longitude. Normally, when we tweet, i believe there is a timestamp and location are well. Given if you disabled the phone GPS or location services, Twitter will still classify tweets using satellite services or use the phone/device IP-address. The only issue here is the accuracy may be nearest in metres.

Time : My intuition here is that tweets could grouped based on time they were tweets, retweeted and more. For instance, if i am interested in knowing what time of the day, do users tweet the most ? I could use the timestamp. I could also use it in combination with location if i want to know what time of the day, week, month do users tweet the most in a region.

Usage : Usage is how often Twitter is used. e.g How many tweets, retweets, mentioned, many more. I also think that what users interact with could be used here. e.g now all users on Twitter tweets every day, but most do interact with the application, website, or view others tweets as well. It also contains how active the user is.

Algorithms: After getting the features, it is important to state the kind of algorithm to be used. In this scenario, i will use a clustering algorithm such as k-Means or gaussian mixture models. We can optimise the algorithm in order to have the best separation between specified groups.

Training, Testing and Communicating Results.

One of the golden rule in machine learning is not to train your model with a testing set. It is very important to divide your dataset into sets. That is you will have your training dataset, testing dataset, and validation set. Normally, i use divide the dataset using a percentage of 70,20,10. Presenting the result would be done using different data visualisation libraries like seaborn, Matplotlip etc.

Question 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.

Solution 3

First, it is important to know what Overfitting is. Overfitting is a kind of model that does not generalise well. Instead it is memorising the dataset. According to Wikipedia, overfitting in statistics is:

the production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably.

However, In machine learning Overfitting is a model that does really well on the training set. It learns the data so well including the noise to extent that it negativity impact the performance of the model on new data.

How to detect overfitting: Your model is likely to be overfitting when it does really well on the training set and performs really bad on the testing set. That is the accuracy of the model of the training set is greater than the testing set accuracy. This is a signal of overfitting.

How to prevent overfitting : There are different ways to prevent overfitting. There are two important techniques that you can use when evaluating machine learning algorithms to limit overfitting:

- Use a resampling technique to estimate model accuracy.
- Hold back a validation dataset.

- Regularisation

Question 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 behaviour?*

Solution 4.

One thing that would be beneficial here, is to use Unsupervised Learning. It will enable us to understand to the patterns. One thing i will do here, will be to cluster some certain similar commands to see how the users combine them together. This helps to see what commands are used with another. With this insight, we can go ahead to create a menu option.

Question 5.

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

Solution 5.

Below i talk about when regularisation is important or not, I think it is important to describe what regularisation is. So **What is regularisation** ?

Regularisation is used to prevent overfitting. In Deep Learning,

regularisation is any modification we make to a learning algorithm that is intended to reduce its generalisation error but not its training error.

-- *Deep Learning by Ian Goodfellow*

According to Him, there are many regularisation strategies. Some put extra constraints on a machine learning model, such as adding restrictions on the parameter values. Some add extra terms in the objective function that can be thought of as corresponding to a soft constraint on the parameter values. If chosen carefully, these extra constraints and penalties can lead to improved performance on the test set.

When is regularisation important ? Regularisation is important when you think you model is overfitting. The goal of learning problem is to find a function that fits or predicts the outcome (label) that minimises the expected error over all possible inputs and labels.

It does not make sense when ? Regularisation does not make sense you have previously normalised your previously normalised your range to be 0 to 1. In this case, regularisation would be less useful.

Question 6.

Your neighbourhood 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?

Solution 6.

My suggestion will be :

- to build a Recommender System which is an Unsupervised Learning algorithm.
- or cluster customers in segments
- Using reinforcement learning to improve over time.

Recommender System is one of the most popular applications of data science today. They are used to predict the rating or preference that a user would give to an item. This system will look at customers histories, find similarities, and then make suggestions based on what they buy alike.

However, **Clustering** using K-means or Gaussian Mixture Model will look the customers in group rather than individually. The algorithm will cluster customers in groups. A preference is created based on the groupings. Now, we can create an offer based on groups and preference with the insight we can get from the clustering algorithms. In order to continue improving the model , a **reinforcement learning** agent can be developed. The agent will learn through rewards and punishment.

Question 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?

Solution 7.

If i am hired, it is important that i will be joyous.

First, i want to learn from the experience of people who are far ahead of me in this field. The job offers that. The task that is given to me will be done by me of course, but then reviewed by project manager or team lead. The feedbacks are important to me because it will help me develop. It is important to mention that i am interested in artificial intelligent. I want to use machine learning and deep learning algorithm to solve business complex challenges.

This contributes to my long term goal in terms of development. Not only does the learning from expect helps me, but research best practice and innovation deepen my understanding and evolves my thinking. One ultimate goal is to build a smart application that solves a problem. I will make it happen. I also want to contribute to the growing community of developers and researchers.