Interview Practice_Data Analyst

By Shaswati Ghosh on 9th May,2018

1. Describe a data project you worked on recently.

The latest project I worked on was finding person of interest(fraud) from Enron email dataset. This dataset was having record of 146 people and was comprised of email and financial data of mostly senior management of Enron. The goal was to come up with a predictive model that could spot an individual as a "Person of Interest (POI). I used scikit-learn & various machine learning techniques to predict "POI"based on their financial & email-data.

At first stage, I have used scikit-learn SelectKBest to find best 12 influential features. Then I scaled all features using max-min scalers to avoid the magnitude difference. Then I have used pipeline and tuning process for parameter tuning in order to improve the fitting to test data set.

As the data set was very small, to avoid data biasing I have used 'StratifiedShuffleSplit'validation with 1000 fold to get the result with test data set. In this case *Recall* and *Precision* are better than *Accuracy* to evaluate the performance because of the content of the dataset. Goal is to identify most of POI persons. Even if someone is wrongly picked, later investigation can make them free of charge. But none(in reality as less as possible) of true POI should be falsely made as non-poi. So having a `good RECALL` and a `low PRECISION` is ideal.</p>

2.You are given a ten piece box of chocolate truffles. You know based on the label that six of the pieces have an orange cream filling and four of the pieces have a coconut filling. If you were to eat four pieces in a row, what is the probability that the first two pieces you eat have an orange cream filling and the last two have a coconut filling?

According to conditional probability:

Probability of 1st choice to be orange: 6/10

Probability of 2nd choice to be orange: (6-1)/(10-1) = 5/9

Probability of 3rd choice to be orange: 4/(10-2) = 4/8

Probability of 4th choice to be orange: (4-1)/(10-3) = 3/7

Total Probablity = (6/10) *(5/9)* (4/8) * (3/7)

```
In [1]: float(float((float(6)/float(10))) * float((float(5)/float(9))) * float((float(
        4)/float(8))) * float((float(3)/float(7))))
```

```
Out[1]: 0.07142857142857142
```

Hence the total probability is 0.0714

Follow-up question: If you were given an identical box of chocolates and again eat four pieces in a row, what is the probability that exactly two contain coconut filling?

Total 4 chocolates out of 10 can be picked in 10C4 ways. Possible combination of 2 orange out of 6 is 6C2
Possible combination of 2 coconut out of 4 is 4C2

Hence the probality of exactly 2 being coconut out of picked 4 chocholet is

(6C2 * 4C2) / 10C2 = 6/14

3.Given the table users: construct a query to find the top 5 states with the highest number of active users. Include the number for each state in the query result. Example result:

In [ ]:
```
import sqlite3

sqlite_file = 'test.db'    # name of the sqlite database file
con = sqlite3.connect(sqlite_file) ## Connect to the database
curs = con.cursor() # Get a cursor object

curs. execute("""SELECT state, count(id) as count
FROM users
WHERE active EQ 1
GROUP BY state
ORDER BY count DESC LIMIT 5;""")
```

4.Define a function first_unique that takes a string as input and returns the first non-repeated (unique) character in the input string. If there are no unique characters return None. Note: Your code should be in Python.

In [3]:
```
def first_unique(string):

    characters = []
    counter = {}

    for char in string:
        if char in counter:
            counter[char] +=1
        else:
            counter[char] =1
            characters.append(char)

    for x in characters:
        if counter[x] == 1:
            return (x)
            #return (x, characters, counter)
    return 'None'
```

```
In [4]: first_unique('aabbcdd123')
Out[4]: 'c'

In [5]: first_unique('a')
Out[5]: 'a'

In [6]: first_unique('112233')
Out[6]: 'None'
```

5.What are underfitting and overfitting in the context of Machine Learning? How might you balance them?

Underfitting defines the situation when the model does not follow the trend in data. That means, even for training data set performance score is not good.

Overfitting happens when the model algorithm follows the training data set so well that it fails to generalize the trend for test data. Overfitted model shows very good result on training dataset, but poor results for test data set.

One of the main reason for overfitting/ under fitting is choice and count of features. Too many features or presence of noise in data can lead to overfitting of data. Similarly choosing features having very less effect on performance or less number of feature can lead to underfitting of data. Use of right number of correct features can balance underfitting and overfitting.

To avoid this, we can use cross validation, splitting the data and conduct multiple trials.

6.If you were to start your data analyst position today, what would be your goals a year from now?

My goals would be:

I want to do coding fluently in Python and R; learn Apache Spark and Apache Storm to use in my work.

I would like to accquire deep knowledge in machine learning and use them to solve realistic business problems of data analysis world. I will deepen my knowledge in data analytics, prediction models and user experience design. I will improve my ability to use metrics and analytics to measure and optimize the design.

I will work hard to become a senior data scientist.