Q1. 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; where as, 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?

Ans: A/B testing is just like two-sample hypothesis testing in which a single variable of two versions is tested. Here, it is tested against the clicked response to Page A against Page B and further determine which page is more effective.

From the given datasets, we can find the mean for how many visitors clicked the button for page A and B.

Case 1:

Mean: - Sign-up button clicked for page A=20/100=0.20

Case 2:

Mean: - Sign-up button clicked for page B=15/70=0.21

From the above the cases, I would say "B" is slightly better than "A". But only mean (for page A and page B) will not give any clue about the statistical significance of the results. More information needed to perform the test to get better results.

Moreover, we can do hypothesis testing for page A and page B.

First, we will perform the null hypothesis test:

The null hypothesis, Ho: mean of page A=mean of Page B

it is incorrect so that we will perform alternative hypothesis. The hypothesis will use p-value to check the strength of the evidence. P value is always between 0 and 1.

a. If p-value (typically ≤ 0.05) which shows strong evidence against the null hypothesis and thus null hypothesis will be rejected.
b. If *p*-value (> 0.05) shows weak evidence against the null hypothesis, so it will fail to reject the null hypothesis.
c. If *p*-values close to the cutoff (0.05), So it will go either way. So, we must draw own conclusions.

To find the p-value for z-test:

1. First write null hypothesis H0 and alternate hypothesis Ha. and decide whether the test is left- tailed, right-tailed, or two-tailed.
2. Identify the level of significance, α and Find the probability that Z is beyond (more extreme than) your test statistic (use z table).
    - If Ha less than alternative, look the test statistic on the Z-table and find its corresponding probability. It will give the - p-value. Test is left -tailed.
    - If Ha greater than alternative, look the test statistic on the Z-table, find its corresponding probability, and subtract it from one. It will give the + p-value. Test is right -tailed

- If Ha not equal to alternative, find the probability that Z is beyond your test statistic and double it. Test is two-tailed. P value=2*Z (obtained from z table)

We will perform the test and find the p values for Page A and Page B and decide which version of Page is more significant.

Q2. 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.

Ans: - There will be three phases for this task:

1. Collecting the tweets stream: - First, we need to call the API (maybe Restful or Soap) so that we can collect the tweets from twitter website. After getting the tweets data we need to format the data into the JSON format which will help in analyzing the tweets better.

2. Data Preprocessing: Twitter data contains lots of noise, so data need to be cleaned up before we use further. Firstly. we need to remove the retweet data as retweet contains same information of tweets data multiple times. Now we will use tokenization so that text will split into separated components. Also removed the signs, emotions, URL, etc. Some of the more frequently used stop words for English include 'a', 'of', 'the', 'I', 'it', 'you', and 'and', and they are generally regarded as functional words which do not carry meaning. We will remove such words. We will also use Stemming algorithms to get rid of derived words like 'developed', 'development', 'developing' is reduced to the stem 'develop'.
3. Tweet data Transformations: – we need to create the feature vectors from the json tweet data so that it can be used for machine learning algorithms. For features transformations, we can use auto-encoders which will help to learn the complex representations and used later for similarity analysis.

4. Machine Learning Algorithm: We can use clustering algorithm (k-means or hierarchical clustering) that will help in making tweets cluster based on its features. After clusters are formed, we can assign the different clusters to the user.

Q3. 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.

Ans: - Overfitting detected when the model is much better on the training set than testing set. It looks like it learned the training data so well it can predict well on the training data but unable to generalize with the unseen data. E.g., if model gives nearly 100 percent accuracy on the training data but

remarkably (e.g. 10%) low on the testing data causes overfitting. There are others reason when model starts to overfit after a point (epochs are higher, depth in a tree is high, too many iterations, too much similar data, etc. could be the reasons) where the learning curve would show that the training scores continue to increase, but the validation scores start to decrease.

Overfitting means that the model learned the training data very well so negatively impact the performance of the model with new datasets. It depends upon the machine learning models, but in general, we need to use resampling technique like k-cross-validation to prevent overfitting which occurs due to training and testing data split. It will help to increase the performance of the model on the new data. Apart from this, there are several techniques for overfitting.

1. parameter tuning
2. cross validation
3. adding more data
4. early stopping
5. for tree-based models, pre or post pruning of trees

There are few machine learning models examples:

a. K-Means: - use smaller values of k for clustering.
b. Neural Networks: use regularization i.e. dropout to prevent overfitting and useless hidden layers

Q4. 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?

Ans: - I think for these types of task, we need to use reinforcement learning strategy as we can't use rule-based mechanism for each behavior. Q learning algorithm will be helpful for such system to improve the experience. At the start, the user will start with simple context menu and use the online learning algorithm to improve its experience. So, based on the experience, it will help to predict the user action. We can set up a scenario like if the user correctly predicted the user menu, then we will give rewards or else give the penalty for the wrong prediction. This is reinforcement set up, in which agent will sense the states, i.e., "user menu selection" and immediate rewards "whether the predicted menu item was used by user". So, it can be done by using the agent(Q-Learning) which assign the rewards and penalty depending upon whether user correctly selects the menu.

Another way around, one possible way is to use the neural net for predicting the behaviors, but it required the substantial amount of data and time to train the model before using it. So I think the reinforcement learning is the best way in such scenario.

Q5. Give an example of a situation where regularization is necessary for learning a good model. How about one where regularization doesn't make sense?

Ans: - Regularization technique is used to prevent overfitting. This happens when model over learns the training data and thus have a poor prediction on testing data. It causes training error consistently low while validation error is high (high variance but low bias). So, regularization technique is used to overcome the model overfitting.

Regularization doesn't make sense when the model is underfitting. It occurs when the model has low variance but high bias. It is because when it didn't take proper features from the data to train it, resulting in an elementary model. Regularization can further put an adverse effect on the model. In general, Regularization is used to reduce model complexity and is not helpful when the data and model are very simplistic.

E.g., I suppose we want to plot the prices of the houses of some city if we plot the prices across the linear regression line we saw prices are scattered against the linear regression and found out the high error in training. So, avoid this, we increase linearly to nth polynomial so that it will fit the entire datasets (prices of the houses) across the polynomials, which causes nearly zero training loss but model unable to fit across the testing datasets to avoid this we regularization. In simple term, for regularization, we add an extra term in the error so that error can't be reduced to zero and we end up with (n-m)the polynomial regression after training which can perform well with training as well as in testing data.

Error function= (Expected output- actual output) +regularization

Q6. 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?

Ans: We can design a recommendation system for this problem, in which model learns based on the user history and preferences. From this, the model will predict the likelihood of the objects the user which they may buy. One way is to find similarities between the users. So, we can use K nearest neighbor to find "k similar users" to the user. The products will be recommended based on the preference of the chosen k users. One criterion, it will recommend the object most of which are more frequently used by the chosen k users. E.g. Based on the history of test user we can recommend the product to buy. We can use hot-encoding for pre-processing the data before applying to the algorithm.

. We can also use clustering algorithm for such problem.

how well the system is performing:

The model performance can be calculated by finding the Root Mean Square Error overall test users data. We can divide the users into a training and testing set and then use a cost function to evaluate the model.

E.g. For example, we will recommend a user to buy a product if the product was not purchased based on the history of the user, we will count that as score 0. If not score 1. Finally, we will find rmse for the performance of the model.

In Real world, we can perform the A/B testing, from the hypothesis we can find which types of coupons are better for the customers. One Example of the coupons will attract more customers whether it is Dollars Off" or "Percent Off". From the hypothesis, we can come to conclusion which is more significant.

Q7. 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?

Ans: - I will try to employ Machine learning and statistical techniques to create state-of-art data products. I seek to learn more, contribute more and grow more than I am today. I know that I will be rewarded personally, professionally as well as financially if I keep doing that. This position is a perfect fit for my aspirations as it allows me to interact with other Data Scientists, Data Engineers, Business Area Engineers and the UX teams and grow my career. It would be a rewarding experience for me as I would be able to learn and develop my skills. This, in turn, would also lead to personal satisfaction as I would reap the benefits of personal and professional development. This role will give me an excellent opportunity to learn and become proficient in R/Python, Hadoop, and Spark over the next year. It also provides me to explore and enhance new skills over the years.

If I think in terms of long-term goals, I believe, this would be a stepping stone for my subsequent roles. This is an opportunity to hone my skill sets, which would help me add value to my current position and help grow the company. Eventually, I would love to work in automated industrial manufacturing by applying machine learning techniques on large and diverse datasets.