SKYLAB AIR

SKYLAB

Introduction

The questions are prepared for the candidates who applied to join **SKYLAB AI Research**. The candidates can use any source to answer the questions. The answers must be written by candidates they/him/herself. In other words, DO NOT copy and paste from the internet. Use your own words and sentences. Evaluate your every answer.

Before answering the questions make sure that you have already familiar with some concepts like supervised learning, unsupervised learning, linear regression, logistic regression, regularization, overfitting, underfitting, k-nearest neighbors algorithm,

Questions

Question1: What is unsupervised and supervised learning? What are the methods of unsupervised and supervised learning?

Question2: Linear Regression is a statistical model. It is a very simple approach for supervised learning. Tell about simple linear regression and multiple linear regression. What are differences of those two method? What questions a linear regression model might answer?

Question3: After building the models, we need to decide how bad our model is. In other words, we need the check the error between actual values and predicted ones. We use a function to do that. What is the name of that function? Also, what is the name of that thing in binary classification setting? (Hint: Give us the actual function names that commonly used in binary classification.)

Question4: How we optimize our linear regression model? Explain the method carefully.

Question5: We use some metrics to evaluate our models. Which metrics used when we evaluate our linear regression model? What is the differences among them. What are pros and cons? Evaluate your answer.

Question6: Overfitting is common problem is statistical learning problems. Especially, in tree methods. We can prevent our model to overfit using some methods. What are the names of those methods in general? Choose one of them and give information about it.

Question7 (Optional): Let $\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith (where $i = 1, 2, ...n$) value of X. Then $e_i = y_i - \hat{y_i}$ represents the ith residual - this is the difference between the ith observed response value and the ith response value that is predicted by our linear model. Then we define residual sum of squares (RSS) as

$$RSS = e_1^2 + e_2^2 + ... + e_n^2$$

or equivalently as

$$RSS = (y_1 - \hat{\beta}_0) - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0) - \hat{\beta}_1 x_2)^2 + ... + (y_n - \hat{\beta}_0) - \hat{\beta}_1 x_n)^2$$

As you know this approach called least squares chooses β_1 and β_2 to minimize RSS. Also define $\bar{y}$ and $\bar{x}$ as $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$. Then using some calculus show that

$$\beta_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Question8: Suppose there is a single predictor and a quantitative response. We fit a linear

regression model to the data, as well as a seperate cubic regression i.e.
$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$.

i. Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

ii. Suppose that the true relationship between X and Y is not linear. However, we don't how far it is from linear. Then do the same thing and consider RSS for both linear regression and cubic regression. What would we expect? Justify your answer.

Question9: We talked about linear regression. We know it used to predict quantitative responses, in other words, continuous outputs. In those problems, our y_i values aren't distinct classes. However, in classification problems, our y_i's will be distinct classes or categories. Consider binary classification case. We want to classify the data into two classes. Assume we used linear regression to classify the data. What problems can occur? Because of these problems we use another well-known method. This method is similar to linear regression expect a few additional steps. Name this method and explain it carefully.

Question10: KNN is very simple supervised learning algorithm. Assume that I know nothing about KNN and I want to learn it. Explain KNN algorithm in a way that teaches me from scratch. While explaining the algorithm you can draw a picture, record a video or you can do whatever you want. Just teach me!