Appau Ernest Kofi Mensah

Theoretical Questions

Linear Regression

1. What is regression?

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

Which models can you use to solve a regression problem?

linear regression
regression trees
lasso regression
multivariate regression

So ,LGBM regressor from microsoft and Cat Boost package from yandex for categorical variables have the best models for regression

2. When do we use it?

We use a linear regression model for problems which have only one independent variable and one dependent variable and on the other hand , we use a multiple regression model for problems which have more than one independent variable.

3. Which metrics for evaluating regression models do you know?

After building a model, we evaluate them by using the p values, F-test, R square value and Adjusted R square values. In most cases we perform a cronbach alpha test and a correlation test to check that we meet all the assumptions regression before building a model.

Others include ,mean squared error ,mean squared log error,mean absolute error

4. What is model bias? Model variance? Bias-variance trade-off?

Bias: Bias describes how well a model matches the training set. A model with high bias won't match the data set closely, while a model with low bias will match the data set very closely. Bias comes from models that are overly simple and fail to capture the trends present in the data set.

Variance: Variance describes how much a model changes when you train it using different portions of your data set. A model with high variance will have the flexibility to

match any data set that's provided to it, potentially resulting in dramatically different models each time. Variance comes from models that are highly complex, employing a significant number of features.

Bias model trade off

The Bias-Variance Trade-Off is a commonly discussed term in data science. This is because actions that you take to decrease bias (Leading to a better fit to the training data) will simultaneously increase the variance in the model (Leading to higher risk of poor predictions). The inverse is also true; actions that you take to reduce variance will inherently increase bias.

Validation

1. What is overfitting?

Overfitting refers to a model that models the training data too well. Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data

2. How to validate models?

Model validation is the task of confirming that the outputs of a statistical model have enough fidelity to the outputs of the data-generating process that the objectives of the investigation can be achieved.

In order to validate the model we can use expert judgement ,or based on the kind of machine learning package we are using.Generally we compare the output of the model we built with the ground truth after passing in an validation data set we check the R Squared score of the model and compare the predictions to the actual ground truth

3. Why do we need to split data?

How many parts would you split your dataset?

I would split my data twice, first split would be a training set and an evaluation set, the 2nd split would be on the evaluation set, into a validation set and a testing set

4. Can you explain how cross-validation works?

Cross-validation is a technique that is used for the assessment of how the results of statistical analysis generalize to an independent data set. Cross-validation is largely used in settings

where the target is prediction and it is necessary to estimate the accuracy of the performance of a predictive mode

In our model training process, we keep a sample of our data that we would use to validate our model and check if it generalises and solves our problem

5. What is K-fold cross-validation?

k-Fold Cross-Validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called **k** that refers to the number of groups that a given data sample is to be split into

Classification

1. What is classification?

In machine learning, classification refers to a predictive modeling problem where a class label is predicted for a given example of input data

Which models would you use to solve a classification problem?
Logistic regression model
Non linear support vector machine
Neural networks
Random forest classifier
Decision trees

2. What is logistic regression?

We use logistic models for binary and multi classification problems

3. Is logistic regression a linear model? No its not

Why?logistic regression models are built on a sigmoid fiction which is a nonlinear function

4. How do you evaluate classification models?

By using confusion matrix, precision , f1 score , accuracy ,recall

5. What is accuracy?

Accuracy is an evaluation metric that is applied to classification models. It is computed by counting the number of labels that were correctly predicted, meaning that the predicted label is exactly the same as the ground truth.

6. What is the confusion table?

The confusion matrix or table helps you analyze the impact of the choices you would have to make if you put the model into production

What are the cells in this table?
Counts of
True Positives
True negatives
False positives
False Negatives

7. What is precision, recall and F1-score?

Precision

Precision is the total number of cases that were correctly classified as positive (called true positive and abbreviated as TP) divided by the total number of cases in that prediction (that is, the total number of entries in the row, both correctly classified (TP) and wrongly classified (FP) from the confusion matrix).

F1 score

The F1 score is another important parameter that helps us to evaluate the model performance. It considers the contribution of both precision and recall using the equation:(2 * precision * recall) / (precision + recall)

Recall

Recall is the total number of predictions that were true divided by the number of predictions for the class, both true and false. Think of it as the true positive divided by the sum of entries in the column

Practical Session

Access the practical or technical session here.