Hakeem Bin Usman
kwabenahabius96@gmail.com
Answers to Theoretical Questions
16/05/2020

Linear Regression

1. What is regression? Which models can you use to solve a regression problem?

Regression is a statistical technique that is used to model the relationship of a dependent variable with respect to one or more independent variables

Models can you use to solve a regression problem:

- simple linear regression
- multiple linear regression
- logistic regression

2. What is linear regression? When do we use it?

Linear regression is a type of predictive analysis that attempts to model the relationship between a dependent variable(y) and one or more more independent variables by fitting a linear equation. It is used to predict the response y from the variables.

It is when used when:

- The relationship between the variables is linear.
- The data is homoscedastic, meaning the variance in the residuals (the difference in the real and predicted values) is more or less constant.
- The residuals are independent, meaning the residuals are distributed randomly and not influenced by the residuals in previous observations. If the residuals are not independent of each other, they're considered to be autocorrelated.
- The residuals are normally distributed. This assumption means the probability density function of the residual values is normally distributed at each x value.

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

The various metrics used to evaluate the results of the prediction are :

1. Mean Squared Error(MSE)
2. Root-Mean-Squared-Error(RMSE).
3. Mean-Absolute-Error(MAE).
4. R^2 or Coefficient of Determination.
5. Adjusted R^2

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

Model bias is when a model is not complex enough, hence it cannot capture the structure in the data, no matter how much data we give it.

Model variance is when a model is highly complex for the amount of training data,so the model learns parts of the noise as well as the true problem structure.

Bias-variance trade-off is a balance in the tension between the error introduced by the bias and the variance.

Validation

1. What is overfitting?

Overfitting is a modeling error that occurs when a function is too closely fit to a limited set of data points in that it fits the training dataset very well but has poor fit with new datasets. Overfit models usually have high variance and low bias.

2. How to validate models?

Models can be validated by:

- checking the assumptions made in constructing the model
- examining the available data and related model outputs
- applying expert judgment

3. Why do we need to split data? How many parts would you split your dataset?

We need to split data in order to get a good generalization from a learning algorithm especially when the data is a lot.

Dataset would be split in three parts, that is 60% training, 20% validation, 20% assessment.

4. Can you explain how cross-validation works?

Cross-validation is an example of a sample-recycling method in which we try to estimate the test error using the same data that we trained on.

How it works:

- Splitting the initial dataset into separate training and test subsets.
- Hold out a set at a time and train the model on remaining set
- Test model on hold out test

5. What is K-fold cross-validation?

K-fold cross-validation is a type of cross-validation that tests a model in a series of iterations. It is where a given data set is split into a K number of sections where each section is used as a testing set at some point

Classification

1. What is classification? Which models would you use to solve a classification problem?

Classification is the process of predicting the class or target or category of given data points.
Models to solve a classification problem:

- Logistic regression
- Support Vector Machine
- Random Forest
- K-Nearest Neighbours
- Stochastic Gradient Descent
- Naïve Bayes
- Decision tree

2. What is logistic regression? When do we need to use it?

Logistic regression is a predictive analysis that describes the relationship between a binary dependent variable and a set of independent variables. It is used when the dependent variable has only two values, such as success or failure, 0 and 1 or Yes and No.

3. Is logistic regression a linear model? Why?

Yes it is a linear model. This is because outcome depends on the sum of the inputs and parameters. But logistic regression is a generalized linear model. Generalized linear models have a linear component, but the model itself is nonlinear.

4. How do you evaluate classification models?

To **evaluate the performance of the** classification models, means you want to know how good the model is in predicting the outcome of new observations test data that have not been used to train the model. The following metrics and methods are used for assessing the performance of classification models, including:

- Classification accuracy: which shows how many of the predictions are correct.
  How: By dividing the number of correct predictions by the number of all predictions.
- Confusion matrix: which is a 2x2 table showing four parameters, including the number of true positives, true negatives, false negatives and false positives. Although this is not a metric to evaluate a model, it provides insight into the predictions.

Other major performance metrics are:

- Sensitivity and specificity
- Precision and recall
- F1 score
- ROC(Receiver Operating Characteristics) curve and AUC(Area Under the Curve)

5. What is accuracy?

Accuracy is the proportion of predictions our model got right.
That is, Accuracy = Number of correct predictions / Total number of predictions

6. What is the confusion table? What are the cells in this table?

The confusion table is a table with two rows and two columns.It reports the number of false positives, false negatives, true positives, and true negatives. Those are the cells in the table.

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

Precision measures how good our model is when the prediction is positive.
Recall measures how good our model is at correctly predicting positive classes.
F1-score is the weighted average of precision and recall.