CAT II

124384 - Luycer Bosire

8/30/2021

# Question One

a.) Explain what EDA (Exploratory Data Analysis) is.

Exploeatory data analysis (EDA) is an approach or a philosophy for data analysis that applies a variety of techniques, mostly graphical, to:

b.) Distinguish between: i. EDA and Summary

The process structure for EDA is:

EDA is a philosophy that aims at creating a model with a futuristic focus to predict future values.

On the other hand, Summary is a numerical reduction method of a numerical data set with focus on the past with the aim to arrive at a key statistic for example the mean or variance.

1. Classical and Bayesian

The difference lies in their process structure as below: Classical: Bayesian

c.) What are the assumptions of EDA?

d.) Identify and implement relevant EDA techniques for given problems.

e.)Explain what you mean by the fixed location assumption. What are the consequences of violating this assumption?

For the univariate problem, the general model component becomes .  
For this case, the fixed location is simply the unknown constant.

The usual estimate of location is the mean from N measurements Y1, Y2, … , YN.

If the run sequence plot does not support the assumption of fixed location, then:

f.)Explain what you mean by the fixed variation assumption. What are the consequences of violating this assumption?

The usualestimate for variation is the standard deviation fro N measurements Y1, Y2, … , YN.

If the run sequence plot does not support the assumption of fixed variation then:

g.)Explain what you mean by the randomness assumption. What are the consequences of violating this assumption?

The randomness assumption is the most critical but the least tested. If the randomness assumption holds, then the lag plot will be structure-less and random.

If the randomness assumption does not hold then:

# Question 2

a.) Box-cox transformation can be used to remove skewness in data.Describe this approach and explain how maximum likelihood can be used to estimate the parameter transformation.

A Box-cox transformation is performed to find the transformation of the dependent variable that maximizes the correlation between a response and a predictor variable. It is definede as:

$ where $ is the transformation parameter.

The likelihood for a given is inversely proportional to the standard deviation of the corresponding T’s. The likelihood function is maximized when the standard deviation is minimized.

b.) Explain how the modified power transformation can be used to remove kurtosis in a symmetric distribution.

Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. That is, data sets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails. Data sets with low kurtosis tend to have a flat top near the mean rather than a sharp peak. A uniform distribution would be the extreme case.

A general form of the power transformation is given by:

$Y=(X^\lambda - 1)/\lambda, \space \lambda \neq 0$

The expression simply reflects the limit of the transformation formula when X is positive and approaches zero. Given an observed data set the transformation parameter must be estimated. Approaches to estimating are usually based on the assumption that the transformed values are normally distributed.

Using a maximum likelihood estimation:

c.) What are outliers and how may they arise in a practical setting?

An outlier is an observation that appears to deviate markedly from other observations in a sample.

Outliers may arise due to human errors, natural deviations, fraudulent behavior or instrumental errors.Outliers may be due to random variation or may indicate something scientifically interesting.

d.) What techniques are used to identify outliers in a univariate data set.

The box plot and the histogram can also be useful graphical tools in checking the normality assumption and in identifying potential outliers. The lower and upper tails of the normal probability plot can be a useful graphical technique for identifying potential outliers. In particular, the plot can help determine whether we need to check for a single outlier or whether we need to check for multiple outliers.

e.) Distinguish between the following:

# Question 3

Explain the following:

#Question 4

a.) Decision trees are procedures that are employed extensively in machine learning and statistical literature. Briefly explain how these procedures work and also mention aspects of their efficiency and reliability. How the procedure works: Decision tree is a Non-parametric supervised learning method. The algorithm can be used for solving regression and classification problems. A decision tree is a tree-like graph with nodes representing the place where we pick an attribute and ask a question; edges represent the answers the to the question; and the leaves represent the actual output or class label. They are used in non-linear decision making with simple linear decision surface they implement a sequential decision process.

* *Decision tree typically starts with a single node(From its roots) a feature is evaluated and one of the two nodes (branches) is selected*
* *The node branches into possible outcomes*
* *Each node in the tree is basically a decision rule.*
* *Each of the outcomes leads to additional nodes which branch off into other possibilities.*
* *This procedure is repeated until a final leaf is reached, which normally represents the target.*

Aspects of their efficiency and reliability: >+ *The decision tree algorithm recursively partitions the given training data set into subsets using generated understandable rules.* >+ *Decision trees perform classification without requiring much computations* >+ *Decision trees handle both continuous and categorical variables*

b.)A major problem associated with regression trees is instability. Explain what you understand by this problem. Instability of decision tree classification is that small change in the input training sample that may cause dramatically large changes in the output particularly if the change occurs in the top level nodes. The constructed rules may also further be significantly different from the original ones if the input training sample is slightly changed. Moreover, the function approximation provided by standard regression trees is highly non-smooth leading to very marked function discontinuities.

c.) Regression tree can be seen as a kind of additive model or a piece wise constant regression model. Mathematically or using an appropriate example, explain how the work.

Mathematical Representation:

Where are constants, I(.) is an indicator function 1 if its argument is true and 0 otherwise, and are disjoint partitions of the training data such that

* *Such models are sometimes called piecewise constant regression models as they partition the predictor space in a set of regions and fit a constant value within each region.*
* *A propositional logic representation of these regions is presented in the form of a tree.*
* *Each path from the root of the tree to a leaf corresponds to a region.*
* *Each inner node of the tree is a logical test on a predictor variable.*
* *In the particular case of binary trees there are two possible outcomes of the test, true or false. This means that associated to each partition we have a path consisting of a conjunction oflogical tests on the predictor variables.*
* *This symbolic representation of the regression function is an important issue when one wants to have a better understanding of the regression surface.*

d.) Describe the recursive partitioning algorithm and its utility in decision trees Its utility in decision trees is that it is a divide and conquer algorithm that recursively partitions the given training data into smaller subsets, makes the method efficient. Recursive partitioning has three main components: >+ *Way it selects a split test* >+ *Rule to determine when a node is a terminal.* >+ *Rule of assigning a value to each terminal node.*

#Question 5. a.) Distinguish between the criteria that are employed for minimizing node impurity for the two methods. impurity of node is minimized by splitting rule

For Regression Trees:

For Correlation Trees:

d.) Boosting is another approach that is employed used in regression trees. Explain how this algorithm works Boosting is another approach for improving the predictions resulting from a decision tree.

Algorithm:

e.) What difference between bagging and boosting? Unlike bagging where each tree is built on a bootstrap data set which is independent of the other trees in boosting each tree is built or grown from previously grown trees. Each tree is fit on a modified version of the original data set. Boosting works in a similar way as bagging, except that the trees are grown sequentially: 1. Each tree is grown using information from previously grown trees. 2. Boosting does not involve bootstrap sampling; instead each tree is fit on a modified version of the original data set.

f.) Explain the impact of heteroskedasticity in the decision tree fitting process: Heteroscedasticity is the tendency of higher value response to have more variation. Regression trees seek to minimize within node impurity. There will be a tendancy to split nodes with higher variations while instead their observations belong together.

g.) Bagging, boosting, random forests and stochastic gradient boosting algorithms are ensemble methods that are often used to enhance the quality of a decision tree. Briefly describe each approach, clearly highlighting the enhancement that they give.

#References