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:

1. maximize insights on the data
2. to recognize patterns or structures
3. to identify important variables
4. to build a parsimonious model
5. to uncover underlying assumptions and distributions
6. to uncover or detect outliers

b.) Distinguish between: i. EDA and Summary

The process structure for EDA is:

$Problem \rightarrow Data \rightarrow Analysis \rightarrow Model \rightarrow Conclusion$

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.

ii. Classical and Bayesian

The difference lies in their process structure as below: Classical: $Problem \rightarrow Data \rightarrow Model \rightarrow Analysis \rightarrow Conclusion$ Bayesian

$Problem \rightarrow Data \rightarrow Model \rightarrow Prior\ distribution \rightarrow Analysis \rightarrow Conclusion$

c.) What are the assumptions of EDA?

1. Fixed scale or variation
2. Fixed location
3. Fixed distribution
4. Randomness

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

1. Run sequential plots

T

hese plots are used for identify outliers and check for the fixed location and fixed scale assumptions, that is a shift in location or variation. They are a linear model fit for the response variable against an index for example time.

2. Lag plots

   T

   he plots are used to check for randomness in the data set. Where the graphicaloutput shows any patterns or structure, then the data has no randomness.

3. Histogram

   A

   graphical used to check for the fixed distribution assumption. It identifies the centre of the data, spread, skewness, presence of outliers and presence of different models in the data set.

4. Normality plot

   T

   his plot is used to check for normality of the variables in the data set. Departures from the line of best fit indicate absence of normality.

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 *response* $=$ *deterministic component* $+$ *random* component becomes *response* $=$ *constant* $+$ *error*.
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.

$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$

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

1. The location may be drifting.
2. The single location estimate may be meaningless (if the process is drifting).
3. The choice of location estimator (e.g., the sample mean) may be sub-optimal.
4. The usual formula for the uncertainty of the mean:

   $s(\bar{Y}) = \frac{1}{\sqrt{N(N-1)}}\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$

   may be invalid and the numerical value optimistically small.
5. The location estimate may be poor.
6. The location estimate may be biased.

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 $s(\bar{Y}) = \frac{1}{\sqrt{(N-1)}}\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$ fro N measurements Y1, Y2, . . . , YN.

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

1. The variation may be drifting.
2. The single variation estimate may be meaningless (if the process is drifting).
3. The variation estimate may be poor and biased.

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:

1. All of the usual statistical tests are invalid.
2. The calculated uncertainties for commonly used statistics become meaningless.
3. The calculated minimal sample size required for a pre-specified tolerance becomes meaningless.
4. The simple model: $y = constant + error$ becomes invalid.
5. The parameter estimates become suspect and nonsupportable.

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:

$T(x) = \frac{X^{\lambda}-1}{\lambda}$\$ where \$$\lambda$ is the transformation parameter.

The likelihood for a given λ 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:

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

$$Y = \begin{cases} \frac{(X^{\lambda}-1)}{\lambda}, & \lambda \neq 0 \\ \ln X, & \lambda = 0; X > 0 \end{cases}$$

The expression $Y = \ln X$ simply reflects the limit of the transformation formula when X is positive and λ approaches zero. Given an observed data set $x_1, x_2, \ldots, x_n$) the transformation parameter must be estimated. Approaches to estimating λ are usually based on the assumption that the transformed values $y_1, y_2, \ldots y_n$ are normally distributed.

Using a maximum likelihood estimation:

$L_m(\lambda) = -\frac{1}{2n \ln \hat{\sigma}_z^2}$

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:

1. Grubb's test

   T

   his is the recommended test when testing for a single outlier.

2. Tietjen-Moore Test

   T

   his is a generalization of the Grubbs' test to the case of more than one outlier. It has the limitation that the number of outliers must be specified exactly.

3. Generalized Extreme Studentized Deviate (ESD) Test

   T

   his test requires only an upper bound on the suspected number of outliers and is the recommended test when the exact number of outliers is not known.

4. Dixon test

   T

   his is is also employed when testing for a single outlier. It is based a value being too large (or small) compared to its nearest neighbor.

Question 3

Explain the following:

1. Principal Component Analysis

   T

   his is a dimension reduction technique used to reduce dimensionality and extract important variables in a data set. It aids to combat multicollinearity. Principal component analysis can be generalized as correspondence analysis to handle qualitative variables and multifactor analysis to handle heterogeneous variables. It is the most common dimension reduction technique as it simplifies the description of the data set. It computes new variables called 'principal components' that are a linear combination of the original variables. Principal components are selected on the basis of variance contribution as a proportion of the total variation of the data set.

2. Correspondence Analysis

T

his is a technique for describing contingency and certain binary tables. This description essentially takes the form of a graphic representation of associations among rows and among columns. It aims to find the lower dimensional sub-space which most accurately approximates the original distribution of points. It is best adopted for cross tabulations.

3. Multi-Factor Analysis

M

ulti factor analysis is a multivariate data analysis method for summarizing and visualizing a complex data table in which individuals are described by several sets of variables (quantitative and /or qualitative) structured into groups. It takes into account the contribution of all active groups of variables to define the distance between individuals. The number of variables in each group may differ and the nature of the variables (qualitative or quantitative) can vary from one group to the other but the variables should be of the same nature in a given group (Abdi and Williams 2010).

MFA may be considered as a general factor analysis. Roughly, the core of MFA is based on: >+ $_{P}rincipalcomponentanalysis(PCA)whenvariablesarequantitative._{>}+_{M}ultiplecorrespondenceanalysis(MCA)whenv$

MFA is used when several sets of variables have been measured on the same set of observations. MFA comprises three main steps: In the first step, a PCA of each data table is performed and the first step singular value of each table recorded. In the second step divide all the elements of a data table by the table's first singular value, concatenate the data tables into a grand data table and then perform a non-normalized PCA of this grand data table. In the third step, the observation partial factor scores for each table are computed by projecting each data table onto the common space. MFA is part of the multi-table family of PCA related techniques. As such it is a simple, elegant, versatile, and robust technique that can be used to integrate multiple data tables collected on the same set of observations. Its simplicity (both theoretical and computational) makes it an ideal tool for the very large data sets of modern science.

#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: $m(x) = \sum_{i=1}^{I} k_i \times I(\boldsymbol{x} \in D_i)$

Where k_i are constants, I(.) is an indicator function 1 if its argument is true and 0 otherwise, and D_i are disjoint partitions of the training data D such that $\cup_{i=i}^{I} D_i = D$ and $\cap_{i=1}^{I} D_i = \phi$

- *Such models are sometimes called piecewise constant regression models as they partition the predictor space χ 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 D_i we have a path P_i 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:

1. Least Squares: Splits are chosen to minimize sum of square error between observation and mean in each node

2. Least Absolute Deviation: This method minimizes the mean absolute deviation from the median within a node. It is not sensitive to outliers has more robust models.

For Correlation Trees:

1. Missclassification error: Proportion of observations in the node that are not members of the majority class.

2. Entropy index: Also called cross-entropy or deviance measures of impurity. This method is more sensitive than missclassification error to changes in the node.

   b.) Describe the algorithm that is employed in pruning a regression tree. Pruning of regression trees is an essential step for obtaining accurate trees Step1:Use recursive binary splitting to grow a large tree on the training data, stopping only when each terminal node has fewer than some minimum number of observations. Step2:Apply cost complexity pruning to the large tree in order to obtain a sequence of best subtrees, as a function of α. Step3:Use K fold cross-validation to choose α. That is, divide the training observations into K folds. For each $k = 1, \ldots, K$: 3.1: Repeat Steps 1 and 2 on all but the k^{th} fold of the training data. 3.2 Evaluate the mean squared prediction error on the data in the left-out k^{th}fold, as a function of α. Average the results for each value of α, and pick α to minimize the average error. Step4:. Return the sub-tree from Step 2 that corresponds to the chosen value of α.

   c.) Bagging approach is used in classification trees. Explain why bagging may be necessary in building a regression tree and how the algorithm works: Bagging is a general purpose procedure for reducing the variance of a statistical learning method and hence increase prediction accuracy.

   Algorithm:

   (a) We take repeated samples from the single training data set.

(b) We will generate B different bootstrapped training data sets.

(c) We then train our method on the b^{th} bootstrapped training set in order to get $f^{*b}x$

(d) Finally, we average all the predictions, to obtain $\hat{f}_{bag}(X)\frac{1}{B}\sum_{b=1}^{B} f^{*b}(X)$

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:

Set $\hat{f}(x) = 0$ and $r_i = y_i$ for all i in all training data setFor $b = 1, 2, 3 \ldots, B$ repeat: **F**

it a tree f^b with d splits $(d + 1)$ terminal nodes to the training data $(\boldsymbol{X}, r)$ **U**

pdate $\hat{f}$ by adding in a shrunken version of the new tree $\hat{f}(x) = \hat{f}(x) + \lambda\hat{f}^b(x)$ **U**

pdate the residuals: $r_i = r_i - \lambda\hat{f}^b(x_i)$ Output the boosted model.

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.

(a) Bagging: This is a general purpose procedure aimed at reducing the variance of a statistical learning method and hence increasing prediction accuracy.

(b) Boosting: Information build up as each tree is grown using information from previously grown trees.

(c) Random forest:are obtained using a fast divide and conquer greedy algorithm that recursively partitions the given training data into smaller subsets. The enhancement it makes is

(d) Stochastic Gradient Boosting: A technique where small regression or classification trees are built sequentially from residuals like measures from the previous trees. At each iteration a tree is built from a random sub-sample of the data.

#References

(a) Thurstone LL. Multiple Factor Analysis. Chicago, IL: University of Chicago Press; 1947.

(b) Abdi, Hervé, Lynne J. Williams, and Domininique Valentin. "Multiple factor analysis: principal component analysis for multitable and multiblock data sets." Wiley Interdisciplinary reviews: computational statistics 5.2 (2013)

(c) EDA - Engineering Statistics Handbook - NIST