Dr. Eick

COSC 3337

Assignment1: Exploratory Data Analysis and Data Visualization

Individual Assignment

Due: Thursday, September 26, 11p (electronic Submission)

Last Updated: Sept. 25, 5p

Weight: 37%±5% of the points available for the 3 assignments

1. Exploratory Data Analysis for a Vehicle Silhouettes Dataset

 

Download Statlog (Vehicle Silhouettes) Data Set dataset from [http://archive.ics.uci.edu/ml/datasets/Statlog+(Vehicle+Silhouettes)](http://archive.ics.uci.edu/ml/datasets/Statlog%2B%28Vehicle%2BSilhouettes%29) limiting yourself to analyzing to the following subset of the dataset for the tasks 1-5 below; use all examples to create the subset:

1. If your last name starts with A-K, you analyze the COMPACTNESS (average perim)\*\*2/area), ELONGATEDNESS (area/(shrink width)\*\*2), SCALED VARIANCE (2nd order moment about minor axis)/area ALONG MAJOR AXIS attributes (1st , 8th , and 11th attribute) and the class variable.
2. If your last name starts with L-Z, you analyze the you analyze the COMPACTNESS (average perim)\*\*2/area), CIRCULARITY (average radius)\*\*2/area, SCALED VARIANCE (2nd order moment about minor axis)/area HOLLOWS RATIO attributes(1st , 2nd , and 18th attribute) and the class variable.

Apply the following exploratory data analysis techniques **using R** to your dataset:

1. Compute the covariance matrix for the three numerical attributes you are analyzing; also compute the correlation for each of the three pairs of attributes. Interpret the statistical findings! 4 points
2. Create a scatter plot for the last two numerical attributes of your dataset. Interpret the scatter plot! 4 points
3. Create histograms for the first 2 numerical attributes each for the whole dataset and the instances of each of the 4 classes. That is, you create 10 histograms. Interpret the obtained displays! 8 points
4. Create box plots for the COMPACTNESS attribute for the instances of each class and a fifth box plot for all instances in the dataset. Interpret and compare the 5 box plots! 5 points
5. Create 3 supervised scatter plots using 2 of the 3 attributes and the class variable; use different colors for the class variable. Interpret the scatter plots! 9 points
6. Fit a linear model that predicts the dependent variable B and H using all the 18 numerical attributes as independent variables for a dataset *VS-Mod[[1]](#footnote-1)* which is created as follows from the complete raw Vehicle Silhouette Dataset:
	1. Z-score the 18 numerical attributes
	2. Add an attribute B to the dataset that is 1 if the example is a bus, and 0 otherwise.
	3. Add a variable V to the dataset that is 1 if the example is a van and 0 otherwise.

Report the R2 of the obtained linear model and the coefficients of each attribute in the obtained two regression functions. Next, interpret the results! What do the coefficients tell you about the importance of the 18 attributes for the two prediction problems? What about negative and positive coefficients—also assess to which extend the coefficients of two regression functions agree with each other. **15 points**

1. Using the dataset VS-Mod you used in the previous task, create 3 different decision tree models that predict the class attribute B based on the numerical 18 attributes and have 20 or less nodes[[2]](#footnote-2) and create 3 different decision tree models that predicts the class attribute V based on the numerical18 attributes and has 20 or less nodes. Explain how the 3 decision tree models were obtained. Report the training accuracy and the testing accuracy for each decision tree; interpret the learnt decision tree—what do they tell you about the importance of the 18 attributes in the used dataset for the classification problem? Assess the training accuracy obtained. Also compare you findings with the findings you obtained for task 6! **18 points**
2. Write a conclusion (at most 18 sentences!) summarizing the most important findings of task1-7—what did we learn about the dataset? In particular, address the findings obtained related to predicting buses, vans, and all 4 classes using the attributes in the dataset. Also assess the difficulty of your classification task **6 points (and up to 4 extra points)**
3. Are there any other interesting observations about your dataset? (up to 4 extra points)

Remark: About 40% of the Assignment1 points will be allocated to interpreting statistical findings and visualizations! A few extra points will be allocated for really good answers to the questions in green!

5 Examples in the raw Vehicle Silhouette Dataset:

96 55 103 201 65 9 204 32 23 166 227 624 246 74 6 2 186 194 opel

89 36 51 109 52 6 118 57 17 129 137 206 125 80 2 14 181 185 van

99 41 77 197 69 6 177 36 21 139 202 485 151 72 4 10 198 199 bus

104 54 100 186 61 10 216 31 24 173 225 686 220 74 5 11 185 195 saab

101 56 100 215 69 10 208 32 24 169 227 651 223 74 6 5 186 193 opel

Attribute Information for the Vehicle Silhouette Dataset:

**Attribute Information:**

ATTRIBUTES

1. COMPACTNESS (average perim)\*\*2/area

2. CIRCULARITY (average radius)\*\*2/area

3. DISTANCE CIRCULARITY area/(av.distance from border)\*\*2

4. RADIUS RATIO (max.rad-min.rad)/av.radius

5. PR.AXIS ASPECT RATIO (minor axis)/(major axis)

6. MAX.LENGTH ASPECT RATIO (length perp. max length)/(max length)

7. SCATTER RATIO (inertia about minor axis)/(inertia about major axis)

8. ELONGATEDNESS area/(shrink width)\*\*2

9. PR.AXIS RECTANGULARITY area/(pr.axis length\*pr.axis width)

10. MAX.LENGTH RECTANGULARITY area/(max.length\*length perp. to this)

11> SCALED VARIANCE (2nd order moment about minor axis)/area
ALONG MAJOR AXIS

12. SCALED VARIANCE (2nd order moment about major axis)/area
ALONG MINOR AXIS

13. SCALED RADIUS OF GYRATION (mavar+mivar)/area

14. SKEWNESS ABOUT (3rd order moment about major axis)/sigma\_min\*\*3
MAJOR AXIS

15. SKEWNESS ABOUT (3rd order moment about minor axis)/sigma\_maj\*\*3
MINOR AXIS

16. KURTOSIS ABOUT (4th order moment about major axis)/sigma\_min\*\*4
MINOR AXIS

17. KURTOSIS ABOUT (4th order moment about minor axis)/sigma\_maj\*\*4
MAJOR AXIS

18. HOLLOWS RATIO (area of hollows)/(area of bounding polygon)

Where sigma\_maj\*\*2 is the variance along the major axis and sigma\_min\*\*2 is the variance along the minor axis, and

area of hollows= area of bounding poly-area of object

The area of the bounding polygon is found as a side result of the computation to find the maximum length. Each individual length computation yields a pair of calipers to the object orientated at every 5 degrees. The object is propagated into an image containing the union of these calipers to obtain an image of the bounding polygon.

CLASSES (4): OPEL, SAAB, BUS, VAN

1. Density-Based Crime Analysis and Data Visualization



In Part B of this Assignment we will be using the following 4 datasets depicted below that can be downloaded from the directory: <http://www2.cs.uh.edu/~ceick/NO/>



Each dataset contains longitude-latitude pairs of the crime of the mentioned category which occurred in a particular time interval; e.g. Harassment12-17.csv contains locations of harassment crimes which occurred in time slots 12 through 17.

1. Create “heatmap”-style density plots for the 4 datasets! Use the same bandwidth for each display; experiment with different values for the bandwidth and try to identify the most suitable bandwidth for density plots for the 4 datasets. Report how and why you chose the particular bandwidth for your 4 displays. 15 points
2. Create a density contour plots for the Harassment12-17.csv and Harassmen6-11.csv datasets. Interpret the obtained density plots! 6 points
3. Summarize to which extend Harassments and PetitLarcency are collocated in time slots 6-11! Try to produce an analysis method and/or visualization method that reports the strength of the collocation[[3]](#footnote-3) between the two crime types in the given time interval. Next, summarize to which extend the 2 crime types are anti-collocated! Try to produce an analysis method and/or visualization method that report the strength of the anti-collocation[[4]](#footnote-4) between the two crime types. Interpret the display/analysis results! 12 points and up to 7 extra points.
4. Summarize to which extend the distribution of harassment crimes changed between time slots 0-5 and 6-11 and between time slots 6-11 and 12-17. Develop an analysis method and/or visualization method that summarizes the strength and direction[[5]](#footnote-5) of change for a given pair of datasets. Alternatively, you might produce and implement a method that identifies and visualizes regions of increase and decreased density with respect a given pair of datasets. Interpret the obtained displays/analysis results! 15 points and up to 3 extra points,

More sophisticated analysis and visualization approaches for tasks 12 and 13 will get higher scores including awarding extra points. Moreover, “alternative” approaches to find good data visualizations for those two tasks are welcome; you might take a look at: <https://cityvis.io/>

Remarks:

* Points allocated to a particular task are preliminary and subject to change
* Points will be deducted for incomplete submission.

**Submission Guidelines**:

Create a folder and name it as LastName\_StudentId\_HW1. HW1 folder should include:

* R code for the tasks.
* The data files needed to run the R codes.
* The assignment report containing all the plots and results along with the interpretations one question at a time.

Submit the LastName\_StudentId\_HW1 folder in a zipped file (.zip no .rar , .7z …) through Blackboard.

1. The dataset has 20 attributes! [↑](#footnote-ref-1)
2. Intermediate nodes count! [↑](#footnote-ref-2)
3. The strength of collocation between crime type A and crime type B is high in location (x,y), if the corresponding density functions ϕA(x,y) ϕB(x,y) both have high values. [↑](#footnote-ref-3)
4. The strength of anti-collocation between crime type A and crime type B is high in location (x,y) is high if one of the corresponding density functions ϕA(x,y) and ϕB(x,y) has high values, whereas the other density function has low values. [↑](#footnote-ref-4)
5. Did the density go up or down? [↑](#footnote-ref-5)