Dr. Eick

COSC 6335 *“Data Mining”* Fall 2024

Problem Set1

First Draft

Last Updated: August 28, 5p

Task1: Exploratory Data Analysis for an Abalone Dataset



Task1 Due: (electronic Submission)

**Learning Objectives**:

1. Learn how to manage and preprocess datasets and how to compute basic statistics and to create basic data visualizations (using R or other tools)
2. Learn how to interpret popular displays, such as histograms, scatter plots, box plots, density plots,…
3. Get some practical experience in exploratory data analysis
4. Learn how to create background knowledge for a dataset
5. Learn to distinguish expected from unexpected results in data analysis and data mining—in general, this task is quite challenging, as it requires background knowledge with respect to the employed data mining technique, and also practical experience.

**Abalone** is a shellfish considered a delicacy in many parts of the world. The abalone shell and the meat is of value. The goal of this project is to perform exploratory data analysis for the Y*Abalone dataset* which is a modification of the Abalone Dataset (<http://archive.ics.uci.edu/ml/datasets/Abalone>). The original Abalone dataset is a 9D dataset and YAbolone is a 10D dataset with an ordinal Age attribute added; YAbalone has the the following attributes:

Sex / nominal / -- / M, F, and I (infant)

Length / continuous / mm / Longest shell measurement

Diameter / continuous / mm / perpendicular to length

Height / continuous / mm / with meat in shell

Whole weight / continuous / grams / whole abalone

Shucked weight / continuous / grams / weight of meat

Viscera weight / continuous / grams / gut weight (after bleeding)

Shell weight / continuous / grams / after being dried

Rings / integer / -- / +1.5 gives the age in years

Age / ordinal/ ---/ Y, M, and O (age classes ‘young’, ‘medium’ and ‘old’; derived from Rings attribute; see below)

4 Examples in the YAbalone Dataset:

M,0.455,0.365,0.095,0.514,0.2245,0.101,0.15,15,O

M,0.35,0.265,0.09,0.2255,0.0995,0.0485,0.07,7,Y

F,0.53,0.42,0.135,0.677,0.2565,0.1415,0.21,9,M

M,0.44,0.365,0.125,0.516,0.2155,0.114,0.155,10,M

The values of the Age attribute have been computed from the Rings attribute as follows: 0-8🡪Y, 9-13🡪M, 14-29🡪O. In general, we are interested to predict Attributes 9 and 10 using the other attributes; that is, we like to predict the age of abalones based on their physical properties described by attributes 2 through 8. Other things we are interested in are finding relationships between the continous attributes in the dataset, and to understand differences between male and female abalones.

Assignment1 Tasks:

Apply the following exploratory data analysis techniques **using R** or other tools of your liking to your dataset:

1. Create the YAbalone dataset or use the one created by the TA!
2. **Task**: Compute the mean value and standard deviation of the 7 numerical attributes**[[1]](#footnote-1).**

**Deliverable: report all mean and std. 1 point**

1. **Task**: Compute the covariance matrix for each pair of the following attributes: Length, Diameter, Shucked Weight, and Rings (treat this attribute as a continuous attribute); next, compute the correlations for each of the 6 pairs of the 4 attributes. Interpret the statistical findings!
**Deliverable: Two comparative explanation, one for covariance matrix and another for correlation. Please make the explanation and comparison based on the properties of covariance matrix and correlation measure e.g. utilizing what these measures explain about the data. 3 points**
2. **Task**: Create a scatter plot for the Shell weight and Height of your dataset. Interpret the scatter plot**!
Deliverable: One comparative explanation. Utilize scatterplots properties for the explanation. 3 points**
3. **Task**: Create histograms for Diameter, Shucked Weight, and Rings attributes for both the male and the female abalones; interpret the obtained 6 histograms.
**Deliverable: One comparative explanation. Utilize histograms properties for the explanation. 6 points**
4. **Task**: Create box plots for the Whole Weight attribute for the instances of each age class—one for Y, M and O— and a fourth box plot for all instances in the dataset. Interpret and compare the 4 box plots for each attribute!
**Deliverable: One comparative explanation. Utilize boxplots properties for the explanation. 4 points**
5. **Task**: Create supervised scatter plots/supervised density plots for the following 3 pairs of attributes using the Age attribute as a class variable: Diameter&Viscera Weight, Diameter&Shell Weight and Viscera Weight& Shell Weight. Use different colors for the class variable. Interpret the obtained plots; in particular address what can be said about the difficulty in predicting the correct age class and the distribution of the instances of the three classes.
**Deliverable: One comparative explanation. Utilize supervised scatterplots properties for the explanation. 6 points**
6. **Task**: Create a new dataset ZAbalone from the YAbalone dataset by transforming the 7 continuous attributes into z-scores. Fit a linear model that predicts the Rings attribute using the 7 z-scored, continuous attributes as the independent variables. Report the R2 of the linear model and the coefficients of each attribute in the obtained regression function. What do the obtained coefficients tell you about the importance of each attribute for predicting the number of rings of an abolone?
**Deliverable: One comparative explanation. Please first explain how you transformed your dataset and next explain the outputs and your interpretation from the outputs. Utilize** R2 **and** coefficient p**roperties for the explanation. 6 points**
7. **Task**: Write a conclusion (at most 13 sentences!) summarizing the most important findings of this task; in particular address the findings obtained related to predicting the age of an abalone (the values of attributes 9 and 10) using attributes 1-8.

**Deliverable: One paragraph that would summarize most important findings about the data. A good explanation would help user to understand different aspects of data very clearly. 6 points (and up to 4 extra points)**

Remark: About 30-40% of the Task1 points will be allocated to interpreting statistical findings and visualizations!

Remark: About 30-40% of the Task1 points will be allocated to interpreting statistical findings and visualizations!

**Submission Guidelines Task1**: Task 1 will be a Kritik submission. For each task you need to create:

* A separate code file with proper commenting
* A report showcasing your results and discussions

**Note: For the following parts, make overall explanations very brief, but clear and informative using any means . All tools such as boxplot, histogram etc. explains certain properties of data. You need to use a tool and make an explanation about the data that would be very clear for people from any technical background. Try to interpret technical explanation from a tool with the data, so that even a person without any technical background can understand it.**

Rubrics for each task is given below:

**Some examples how tools can be used: [Please follow more tutorials to learn better]**

1. **Mean, std:** [**https://libraryguides.centennialcollege.ca/c.php?g=717168&p=5123683**](https://libraryguides.centennialcollege.ca/c.php?g=717168&p=5123683)
2. **Covariance:** [**https://builtin.com/data-science/covariance-matrix**](https://builtin.com/data-science/covariance-matrix) **Correlation:** [**https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5079093/**](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5079093/)
3. **Scatter plot:** [**https://www.texasgateway.org/resource/interpreting-scatterplots**](https://www.texasgateway.org/resource/interpreting-scatterplots)
4. **Histograms:** [**https://www.labxchange.org/library/items/lb:LabXchange:10d3270e:html:1**](https://www.labxchange.org/library/items/lb%3ALabXchange%3A10d3270e%3Ahtml%3A1)
5. **Boxplots:**[**https://www.simplypsychology.org/boxplots.html**](https://www.simplypsychology.org/boxplots.html)[**https://www.atlassian.com/data/charts/box-plot-complete-guide**](https://www.atlassian.com/data/charts/box-plot-complete-guide)
6. **Linear Regression:**[**https://datatab.net/tutorial/linear-regression**](https://datatab.net/tutorial/linear-regression)[**https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/**](https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/)
1. This is more a verification of that you have the correct dataset! [↑](#footnote-ref-1)