Dr. Eick

COSC 6335 Data Mining Fall 2022

Problem Set1

Sixth Draft

Last Updated: September 22, 6p

Task1: Exploratory Data Analysis for an Abalone Dataset



Task1 Due: Sunday, Sept. 25, 11:59p (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.

Assignment1 Tasks:

**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)

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. Compute the mean value and standard deviation of the 7 numerical attributes**[[1]](#footnote-1). 1 point**
3. 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! **3 points**
4. Create a scatter plot for the Shell weight and Height of your dataset. Interpret the scatter plot**! 3 points**
5. Create histograms for Diameter, Shucked Weight, and Rings attributes for both the male and the female abalones; interpret the obtained 6 histograms. **6 points**
6. 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! **4 points**
7. 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. **6 points**
8. 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? **6 points**
9. 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. **6 points (and up to 4 extra points)**

*Task1 Submission Guidelines and Rubrics*

1. Each interpretation should use complete sentences to describe your findings. Put interpretations next to the figures you interpret!
2. Question 1 to 8 will be peer reviewed on Kritik.
3. Question number 8 will be reevaluated by the TA and Professor
4. In Kritik, all tasks will be uploaded as an individual task
5. Each student will be needed to Create, Evaluate and Give Feedback
6. Creation:
	1. Student will need to upload a report and code file
7. Evaluation:
	1. Each student will evaluate the reports of three other student
8. Feedback:
	1. Each student will give feedback on the evaluation
	2. If a student have any objection on any evaluation they will raise a flag and TA will reevaluate the report
9. Marking:
	1. Creation, Evaluation and Feedback all will carry a percentage of total greading

**Rubrics**:

**Q1:** Compute the mean value and standard deviation of the 7 numerical attributes**[[2]](#footnote-2). 1 point**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Level 0 | Level 1 | Level 2 | Level 3 | Weight |
| Mean Values  | No mean values are correct | Only two out of seven are correct | Only four out of seven are correct | Seven out of seven are correct | 1 |
| Standard Deviation | No standard deviation is correct | Only two out of seven are correct | Only four out of seven are correct | Seven out of seven are correct | 1 |

**Q2**: 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! **3 points**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Level 0 | Level 1 | Level 2 | Level 3 | Weight |
| Covariance Matrix | All six covariance matrix has flaws or all are missing | At least four covariance matrix has serious flaws or are missing | At least two covariance matrix has serious flaws or are missing | All six covariance matrix are correct | 1 |
| Covariance Matrix Interpretation | All six covariance matrix has serious misinterpretation or missing information or cannot represent relationships among the variables | At least four covariance matrix has serious misinterpretation or missing information or cannot represent relationships among the variables | At least two covariance matrix has serious misinterpretation or missing information or cannot represent relationships among the variables | All six covariance matrix are correct, convenient and can capture the relationships among the variables well | 2 |
| CorrelationMatrix | All six correlationmatrix has flaws or all are missing | At least four correlation matrix has serious flaws or are missing | At least two correlation matrix has serious flaws or are missing | All six correlation matrix are correct | 1 |
| CorrelationMatrix Interpretation | All six correlation matrix has serious misinterpretation or missing information or cannot represent relationships among the variables | At least four correlation matrix has serious misinterpretation or missing information or cannot represent relationships among the variables | At least two correlation matrix has serious misinterpretation or missing information or cannot represent relationships among the variables | All six correlation matrix are correct, convenient and can capture the relationships among the variables well | 2 |

**Q3**: Create a scatter plot for the Shell weight and Height of your dataset. Interpret the scatter plot**! 3 points**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Level 0 | Level 1 | Level 2 | Level 3 | Weight |
| Scatter Plot Quality  | No or wrong scatter plot given | Quality of the scatter plot is very poor and lots of missing details within the scatter plot | Quality of the scatter plot is modest and number of missing details within the scatter plot is low | Quality of the scatter plot is very high and the scatter plot perfectly visualize the relationship between the variables | 1 |
| Scatter Plot Interpretation | No interpretation is given or the interpretation is wrong | The given interpretation is very low quality with lots of missing information | The given interpretation is modest with a few missing information | The given interpretation is very advanced  | 1 |

**Q4:** Create histograms for Diameter, Shucked Weight, and Rings attributes for both the male and the female abalones; interpret the obtained 6 histograms. **6 points**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Level 0 | Level 1 | Level 2 | Level 3 | Weight |
| A3 H-Quality | The two histograms for attribute 3 are not correct or have serious other flaws | The two histograms for attribute 3 misdescribe the differences in the distribution of the two attribute values or have other flaws (e.g. make it hard to compare the 2 histograms) | The two histograms for attribute 3 do not capture the distributions of the two attribute values perfectly but do not have other flaws. | The two histograms for attribute 3 capture the distributions of the two attribute values perfectly and do not have other flaws. | 2 |
| A6 H-Quality | The two histograms for attribute 6 are not correct or have serious other flaws | The two histograms for attribute 6 misdescribe the differences in the distribution of the two attribute values or have other flaws (e.g. make it hard to compare the 2 histograms) | The two histograms for attribute 6 do not capture the distributions of the two attribute values perfectly but do not have other flaws. | The two histograms for attribute 6 capture the distributions of the two attribute values perfectly and do not have other flaws. | 2 |
| A9 H-Quality | The two histograms for attribute 9 are not correct or have serious other flaws | The two histograms for attribute 9 misdescribe the differences in the distribution of the two attribute values or have other flaws (e.g. make it hard to compare the 2 histograms) | The two histograms for attribute 9 do not capture the distributions of the two attribute values perfectly but do not have other flaws. | The two histograms for attribute 9 capture the distributions of the two attribute values perfectly and do not have other flaws. | 2 |
| A3 H-Comparison | The comparison of the commonalities and difference of the 2 histograms for attribute 3 is highly incomplete and/or makes erroneous statements | The comparison of the commonalities and difference of the 2 histograms for attribute 3 is only somewhat complete and valid | The comparison of the commonalities and difference of the 2 histograms for attribute 3 is mostly convincing, complete and valid | The comparison of the commonalities and difference of the 2 histograms for attribute 3 is convincing, complete and valid | 1 |
| A6 H-Comparison | The comparison of the commonalities and difference of the 2 histograms for attribute 6 is highly incomplete and/or makes erroneous statements | The comparison of the commonalities and difference of the 6 histograms for attribute 3 is only somewhat complete and valid | The comparison of the commonalities and difference of the 6 histograms for attribute 3 is mostly convincing, complete and valid | The comparison of the commonalities and difference of the 6 histograms for attribute 3 is convincing, complete and valid | 1 |
| A9 H-Comparison | The comparison of the commonalities and difference of the 9 histograms for attribute 3 is highly incomplete and/or makes erroneous statements | The comparison of the commonalities and difference of the 9 histograms for attribute 3 is only somewhat complete and valid | The comparison of the commonalities and difference of the 9 histograms for attribute 3 is mostly convincing, complete and valid | The comparison of the commonalities and difference of the 9 histograms for attribute 3 is convincing, complete and valid | 1 |
| U A3+6+9 | The assessment of the usefulness of attribute 3,6 and 9 for identifying male and female abalone based on obtained histogram data is poor | The assessment of the usefulness of attribute 3,6 and 9 for identifying male and female abalone based on obtained histogram data is not very convincing | The assessment of the usefulness of attribute 3,6 and 9 for identifying male and female abalone based on obtained histogram data is mostly convincing | The assessment of the usefulness of attribute 3,6 and 9 for identifying male and female abalone based on obtained histogram data is convincing | 1 |

**Q5:** 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! **4 points**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Level 0 | Level 1 | Level 2 | Level 3 | Weight |
| A5 with separate age class –B-Quality | The 3 boxplots for attribute 5 with separate age class are completely wrong | The 3 boxplots for attribute 5 with separate age class have major flaws(e.g. missing values were not removed) | The 3 boxplots for attribute 5 with separate age class have minor flaws | The 3 boxplots for attribute 5 with separate age class are correct | 1 |
| A5 wall instance –B-Quality | The all instance boxplot for attribute 5 is completely wrong | The all instance boxplot for attribute 5 has major flaws(e.g. missing values were not removed) | The all instance boxplot for attribute 5 has minor flaws | The all instance boxplot for attribute 5 is correct | 1 |
| A5 with separate age class –B-Comparison | The comparison of the commonalities and differences of the 3 boxplots for attribute 5 with separate age class is highly incomplete and/or makes erroneous statements | The comparison of the commonalities and differences of the 3 boxplots for attribute 5 with separate age class is only somewhat complete and valid | The comparison of the commonalities and differences of the 3 boxplots for attribute 5 with separate age class is mostly convincing, complete and valid | The comparison of the commonalities and differences of the 3 boxplots for attribute 5 with separate age class is convincing, complete and valid | 1 |
| A5 with separate age class –B-Comparison with whole instance boxplot | The three set comparisons of the commonalities and differences of each age class boxplot for attribute 5 with separate age class against whole instance boxplot are highly incomplete and/or makes erroneous statements | The three set comparisons of the commonalities and differences of each age class boxplot for attribute 5 with separate age class against whole instance boxplot are only somewhat complete and valid | The three set comparisons of the commonalities and differences of each age class boxplot for attribute 5 with separate age class against whole instance boxplot are mostly convincing, complete and valid | The three set comparisons of the commonalities and differences of each age class boxplot for attribute 5 with separate age class against whole instance boxplot are convincing, complete and valid | 1 |
| U A5 | The assessment of the usefulness of attribute 5 for identifying different abalone age group based on obtained boxplot data is poor | The assessment of the usefulness of attribute 5 for identifying different abalone age group based on obtained boxplot data is not very convincing | The assessment of the usefulness of attribute 5 for identifying different abalone age group based on obtained boxplot data is mostly convincing | The assessment of the usefulness of attribute 5 for identifying different abalone age group based on obtained boxplot data is convincing | 1 |

**Q6:** 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. **6 points**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Level 0 | Level 1 | Level 2 | Level 3 | Weight |
| SS A3,7; A3,8; A7,8 | The 3 supervised scatter plots for the three pairs of attributes are not correct or have serious flaws(e.g. not a supervised scatter plot) | The 3 supervised scatter plots for the three pairs of attributes misdescribe the difference in the distribution of the attributes values or have other flaws (e.g. missing axis lables, legend..) | The 3 supervised scatter plots for the three pairs of attributes do not capture the difference in the distribution of the attributes values perfectly but do not have any other flaws (e.g. size of the point symbol.) | The 3 supervised scatter plots for the three pairs of attributes capture the difference in the distribution of the attributes values perfectly | 1 |
| SS A3,7 Interpretation | The interpretation of the supervised scatterplot for the attribute 3 and 7 is highly incomplete and/or makes erroneous statements | The interpretation of the supervised scatterplot for the attribute 3 and 7 is only somewhat complete and valid | The interpretation of the supervised scatterplot for the attribute 3 and 7 is mostly convincing, complete and valid | The interpretation of the supervised scatterplot for the attribute 3 and 7 is mostly convincing, complete and valid | 1 |
| SS A3,8 Interpretation | The interpretation of the supervised scatterplot for the attribute 3 and 8 is highly incomplete and/or makes erroneous statements | The interpretation of the supervised scatterplot for the attribute 3 and 8 is only somewhat complete and valid | The interpretation of the supervised scatterplot for the attribute 3 and 8 is mostly convincing, complete and valid | The interpretation of the supervised scatterplot for the attribute 3 and 8 is mostly convincing, complete and valid | 1 |
| SS A7, 8 Interpretation | The interpretation of the supervised scatterplot for the attribute 7 and 8 is highly incomplete and/or makes erroneous statements | The interpretation of the supervised scatterplot for the attribute 7 and 8 is only somewhat complete and valid | The interpretation of the supervised scatterplot for the attribute 7 and 8 is mostly convincing, complete and valid | The interpretation of the supervised scatterplot for the attribute 7 and 8 is mostly convincing, complete and valid | 1 |
| SS A3,7, 8 Interpretation | The interpretation of the 3D scatterplot is highly incomplete and/or makes erroneous statements | The interpretation of the 3D scatterplot is only somewhat complete and valid | The interpretation of the 3D scatterplot is mostly convincing, complete and valid | The interpretation of the 3D scatterplot is mostly convincing, complete and valid | 1 |
| U A 3-7-8 | The assessment of the usefulness of 3D-scatterplot compared to 2D scatterplot is poor | The assessment of the usefulness of 3D-scatterplot compared to 2D scatterplot is not very convincing | The assessment of the usefulness of 3D-scatterplot compared to 2D scatterplot is mostly convincing | The assessment of the usefulness of 3D-scatterplot compared to 2D scatterplot is convincing | 1 |

**Q7:** 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? **6 points**

Linear model rubric

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Level 0 | Level 1 | Level 2 | Level 3 | Weight |
| LM-correctness | Reported coefficients and R\*\*2 are completely wrong | Reported coefficients and R\*\*2 contain a lot of error | Reported coefficients and R\*\*2 contain some minor errors or R\*\*2 is incorrect or missing | Reported coefficients and R\*\*2 are correct | 3 |
| LM-Interpretation | The interpretation of the obtained coefficients to assesses attribute usefulness to estimate rings attribute is of very poor quality | The interpretation of the obtained coefficients to assesses attribute usefulness to estimate rings attribute has some major flaws or is incomplete | The interpretation of the obtained coefficients to assesses attribute usefulness to estimate rings attribute is mostly convincing | The interpretation of the obtained coefficients to assesses attribute usefulness to estimate rings attribute is convincing | 3 |

**Q8:**

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. **6 points (and up to 4 extra points)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Level 0 | Level 1 | Level 2 | Level 3 | Level 4  | Weight |
| Conclusion  | No conclusion | The conclusion is poor and it gives very poor insight which variables will be good for predicting the age of Abalone | The conclusion gives a very moderate insight which variables will be good for predicting the age of Abalone, but it can be improved | The conclusion gives a very good insight which variables will be good for predicting the age of Abalone | The conclusion gives a very advanced insight about which variables will be good for predicting the age of Abalone | 6+(4) |

1. This is more a verification of that you have the correct dataset! [↑](#footnote-ref-1)
2. This is more a verification of that you have the correct dataset! [↑](#footnote-ref-2)