Assignment 1 Submission Sheet

Data Mining 1DL360 / 1DL370

This document contains questions to help you reflect about the operations applied to the data during this assignment. You have to fill it in and submit it on Studium.

TASK 1: Reading the Data

1. What are the average length of sepals (sl) and their standard deviation?
   Average: -5.706 // Deviation: 303.789

2. How many instances are there for each class?

   - Virginica: 3000
   - Setosa: 3000
   - Versicolor: 500

TASK 2: Data Cleaning

1. Why is it important to let the system know which values are missing?
   To let the system know with values are not available or correct to use

2. What are the average length of sepals (sl) and their standard deviation after declaring missing values?
   AVERAGE: 3.528 ; DEVIATION: 2.102

3. What are the average length of sepals (sl) and their standard deviation after removing outliers?
   AVERAGE: 3.547 ; DEVIATION: 2.023

4. Do you think the outliers you have removed were noise (that is, wrong measurements) or unusual but correct observations?
   I think it removed unusual but correct observation because it has so unusual values (measure bellow 0) so it has probably affected the mean used for removing outlines

5. Would you first handle missing data and then remove outliers, or the other way round? Why?
   First handle missing data and then remove outlines to get rid first of wrong data (we know it´s wrong because it does not make sense to a length be negative)

6. Assume your observations (records) represent people in a social network, and one variable stores their degree centrality. Would you remove outliers in this case? Why?
   No because the values can be very variable

TASK 3: Data Transformation

1. What are the average length and standard deviation of sepals after min-max normalization?
   Average: 0.443 // Deviation: 0.326

2. What are the average length and standard deviation of sepals after standardization?
   Average: 0.000 // Deviation: 1.000

From here on, continue with min-max normalization instead of standardization.

3. How many components have been selected after applying PCA? 3

4. How much variance is captured by the first two components of PCA? 0.845

5. How is the first PCA component defined as a combination of the original attributes?
   Id: 1.000 // pl: 0.015 // pw: -0.001 // sl: 0.004 // sw: 0.004

6. How many PCA components would have been selected after one attribute expressed on a larger range? 1

7. How many components would have been selected with an outlier included? 3

TASK 4: Sampling

	Simple sampling	Boot-strapping	Stratified proportional	Stratified balanced
Number of iris versicolor	13	13		
Number of iris setosa	76	59		
Number of iris virginica	61	78		
Are there repeated identifiers?	No	No		
Does the number of iris versicolor included in the sample change if you change the local random seed?	Yes	Yes		

TASK 5: Classification (Optional task for RapidMiner)

How have the unlabeled records been classified?

Record	Species
1	
2	
3	
4	
5	
6	

Compare the original values (iris_unlabeled) with the values in the labeled dataset (iris_data + iris_labels). Does this classification look correct? If not: find what the problem was, describe it below, and fix it!