Gustav Nolgren
Uppsala universitet
2023-09-08

Assignment 1 Submission Sheet

TASK 1: Reading the Data

1. The average length of sepal length is -5.705508.
   The standard deviation is 303.788948.
2. Instances of each class:
   a. Virginica: 3000
   b. Setosa: 3000
   c. Versicolor: 500

TASK 2: Data Cleaning

1. If we were to keep the missing values, in this case -9999, the upcoming tests would be completely wrong since these will impact the calculations so drastically. For example, the average sepal length becomes negative, which is of course impossible.
2. **After declaring missing values**
   Average sepal length: 3.527595.
   Standard deviation: 2.102492.
3. **After removing outliers**
   Average sepal length: 3.519643.
   Standard deviation: 2.018383.
4. Looking at the standard deviation in 2.2, it would be extremely unlikely to have sepal lengths of over 20. Thus, it would be reasonable to think that the data added a zero too many perhaps.
5. Since it is easier to find missing data more often than not, I would say it is better to remove that first and see how the outliers look after that and remove them, if necessary, as well.
6. Removing outliers in this case might be unwise. If you were to do research on a social network, people with very few or very many connections will be of high interest. However, with extremely high degrees of centrality it might be very unlikely that it really is the case.

TASK 3: Data Transformation

1. **After min-max normalization**
   Average sepal length: 0.438652.
   Standard deviation: 0.325546.
2. **After standardization**
   Average sepal length: ~0
   Standard deviation: 1.000077
3. To explain 95% of the variance, 2 components have been selected.
4. 98.07% of the variance.
5. The first PCA component is defined as:
   sl: 0.356113
   sw: -0.079923

pl: 0.856786
pw: 0.36430

6. After scaling the petal length to range from 1-100, only 1 PCA component would be selected.
7. With the outlier of 5000 at [0, 'sl'], we get two components.

TASK 4: Sampling

	Simple sampling	Bootstrapping	Stratified proportional	Stratified balanced
Number of iris versicolor	13	5	250	50
Number of iris setosa	75	68	1498	50
Number of iris virginica	62	77	1498	50
Are there repeated identifiers?	No	Yes	No	No
Does the number of iris versicolor included in the sample change if you change the local random seed?	Yes	Yes	No	No