**Question1 KNN and Normalization/Standardization**

Attached file Q1\_train.csv Q1\_test.csv are the training and testing dataset for a binary classification. The last columns of both files contain the labels.

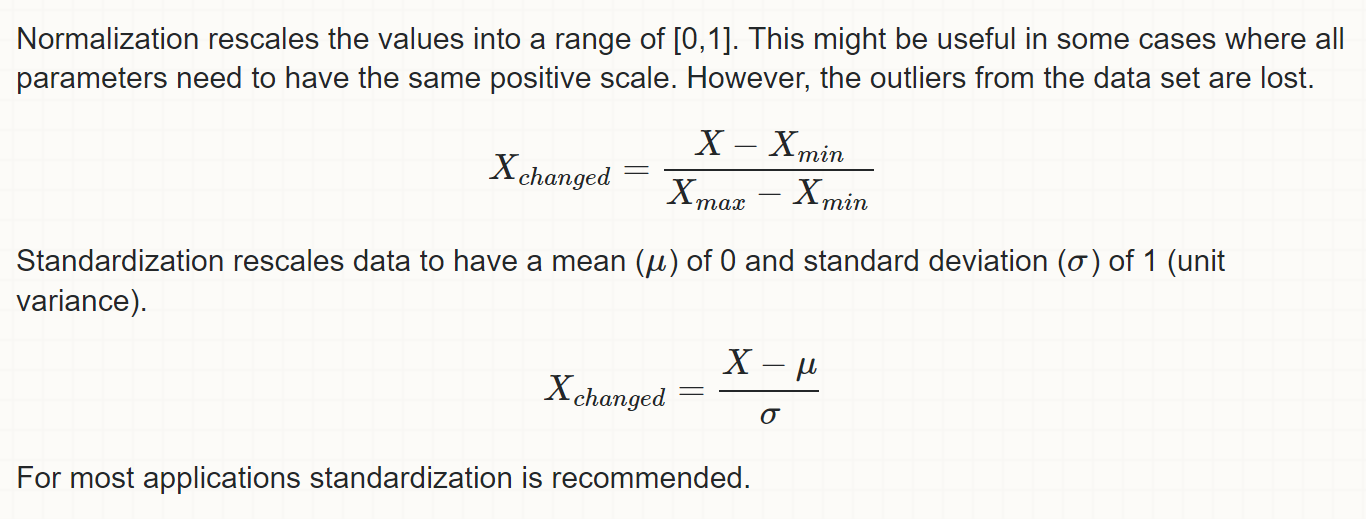
Write a basic K-nearest neighbor classifier program to test the performance on the test dataset using the Euclidean distance. Try different K values and record the performances (accuracy) below.   
Submit the code along with the tables below

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| K= | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| No. of correct predictions |  |  |  |  |  |  |  |  |  |
| Accuracy |  |  |  |  |  |  |  |  |  |

Now, add a normalization step to scale all attributes into the range of [0,1] and try your improved KNN algorithm on the test dataset and record the performance over K=1 to 9. Do the same for standardization.

Performance with scaled attributes via normalization

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| K= | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| No. of correct predictions |  |  |  |  |  |  |  |  |  |
| Accuracy with normalization |  |  |  |  |  |  |  |  |  |
| Accuracy with standardization |  |  |  |  |  |  |  |  |  |



**Question 2: Latest research papers on KNN classifier**

Find and summarize the key new ideas in TWO of the following papers on KNN **using your own words (**you can use figures to explain the ideas):

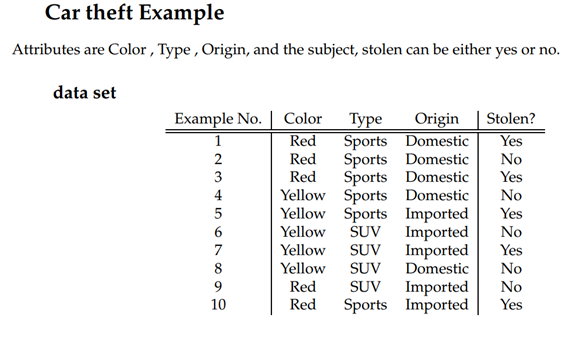
1. Papernot, Nicolas, and Patrick McDaniel. "Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning." *arXiv preprint arXiv:1803.04765* (2018).
2. Jiang, Weiwei. "Time series classification: nearest neighbor versus deep learning models." *SN Applied Sciences* 2, no. 4 (2020): 1-17.
3. Bergman, Liron, Niv Cohen, and Yedid Hoshen. "Deep nearest neighbor anomaly detection." *arXiv preprint arXiv:2002.10445* (2020).
4. Yang, Xi, Xiaoting Nan, and Bin Song. "D2N4: A Discriminative Deep Nearest Neighbor Neural Network for Few-Shot Space Target Recognition." *IEEE Transactions on Geoscience and Remote Sensing* 58, no. 5 (2020): 3667-3676.

Write one of your own ideas on how to improve KNN classifier

**Question 3:** Naïve Bayes Classifier

For the following theft stolen prediction problem, predict the label classify a Red Domestic SUV using a Bayes Classifier (Do it manually by calculating all the probabilities)

P(Yes| Red Domestic SUV) vs P(No| Red Domestic SUV)



Question 4: Use Scikit-Learn or Weka to solve a classification problem using Naive Bayes Classifier

**You need to paste your screenshots of either running output.**

The dataset is for a german bank loan risk prediction problem.   
german\_credit\_data.csv has the raw data provided by the data provider

german.doc has the explanation how they preproposs the raw data to encode it

german.data-numeric.csv has the cleaned converted numeric dataset for use in your algorithm

The last column of the german.data-numeric.csv is the class label.

**You should use the german.data-numeric-train.csv as the training data, and use the german.data-numeric-test.csv as testing data**

**Option 1: use weka to apply Naïve Bayes Classifier to the datasets Q4\_train.csv Q4\_test.csv (**Screenshot required)

<https://scienceprog.com/building-and-evaluating-naive-bayes-classifier-with-weka/>  
<http://weka.sourceforge.net/doc.dev/weka/classifiers/bayes/NaiveBayes.html>

**Option 2: use Scikit-learn Python Library to apply Naïve Bayes Classifier to the datasets Q4\_train.csv Q4\_test.csv (if you know python, this one is preferred)**

<https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/>

(code submission required)