Machine Learning – Exercise 1

Classification

Group 7

1. Classification using K-NN

1.1. What is K-NN?

The k-nearest neighbors algorithm is a non-parametric method used for classification and regression. In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors with the object assigned to the class most common among its k nearest neighbors. K is a positive number, typically small, in order to reduce the number of calculations, and thus the process duration. As the experiments will show, the accuracy of the algorithm is not increasing with the increasing of k. If k=1, then the class of the object would the class of the single neares neighbour.

1.2. Classification on small datasets

1.2.1. Crime-Mapping dataset

The evaluation of the K-NN classifier on this small datasets is made by using the algorithm to predict if there is an official hospital record about the crime, i.e., if the variable phxcommunity should have value of Yes or No. Since the dataset contains also nominal categorical variables, the transformation of categorical data to numerical is needed. The experiment is made using multiple values of parameter k. In order to get most accurate results, the numeric variables need to be normalized.

The evaluation of the accuracy of the algorithm, the presumption, that the greater k does not imply better results, is confirmed. As the figure states, there is no connection between accuracy and k. For example, the algorithm showed better results with k=17 as by using the k=25, but the classification using k=25 made less correctpredictions as by using k=250.

k	acc	duration
1	88.01230	0.0270869731903076
3	87.29508	0.0266668796539307
5	87.70492	0.0269179344177246
7	87.70492	0.0262429714202881
17	85.86066	0.02742600440979
25	85.45082	0.0293509960174561
35	85.14344	0.0351059436798096
50	85.14344	0.0439131259918213
69	85.14344	0.0564539432525635
250	85.96311	0.266210794448853

k	acc	duration
1	0.8852459	0.0238590240478516
3	0.8821721	0.0245380401611328
5	0.8852459	0.0250470638275146
7	0.8780738	0.0270748138427734
17	0.8760246	0.0303258895874023
25	0.8678279	0.0309109687805176
35	0.8627049	0.0370848178863525
50	0.8586066	0.0447700023651123
69	0.8596311	0.0559730529785156
250	0.8596311	0.25897479057312

Figure 1.1 - Tables showing the percentage of correct predictions (acc) and process duration in seconds for each chosen value of k – with/without normalization of variables

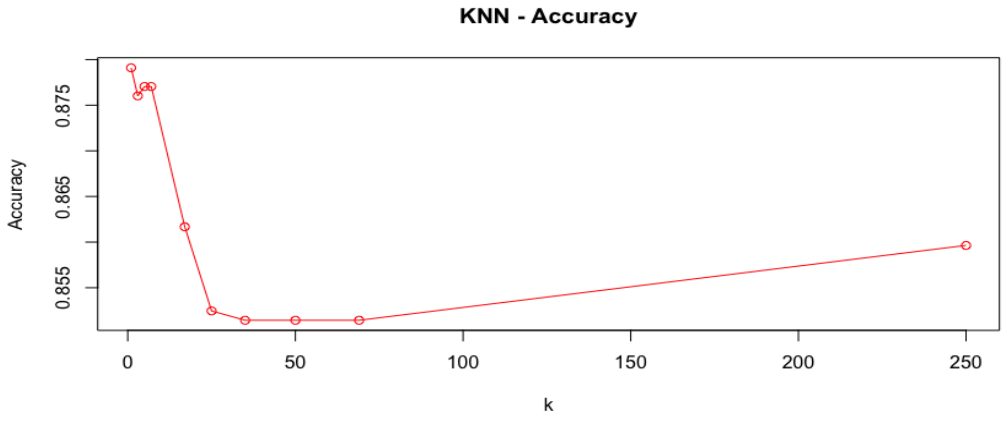

Figure 1.2 - KNN accuracy on Crime-Mapping dataset

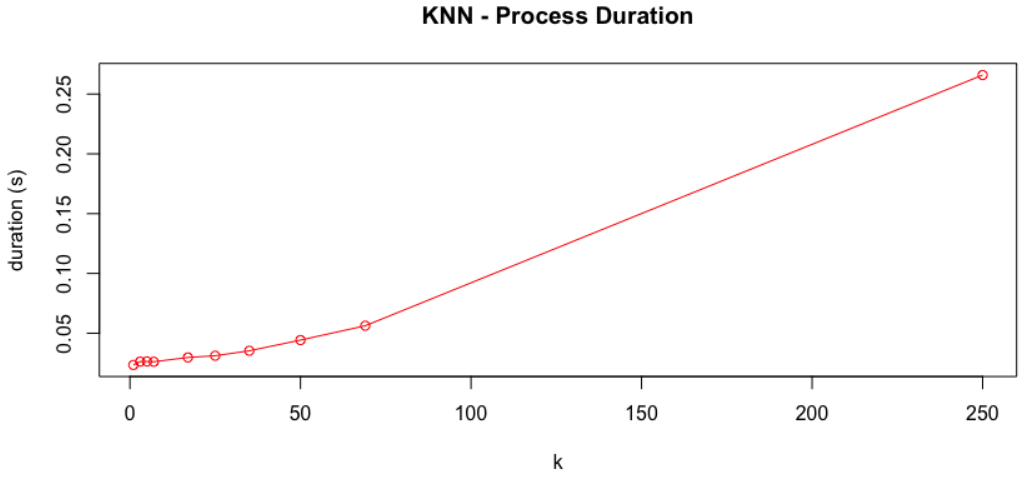

Figure 1.3 - KNN - Process duration

1.2.2. Breast-Cancer dataset

The evaluation of the K-NN classifier on this small datasets is made by using the algorithm to predict if the cancer is benign or malignant, i.e., if the variable class should have value of B or M. The experiment is made using multiple values of parameter k. In order to get most accurate results, the numeric variables need to be normalized.
The evaluation of the accuracy of the algorithm, the presumption, that the greater k does not imply better results, is again confirmed. The classification using k=7 on the normalized dataset results with 100% accuracy.
As the diagramm on figure 2.3 states, the process duration increases with big values of k. However, it is not the statement, as the process duration of the classification using the k=5 is less then when using k=7.

k	acc	duration
1	96.49123	0.00172996520996094
3	96.49123	0.00165987014770508
5	98.24561	0.00198602676391602
7	100.00000	0.00204014778137207
16	96.49123	0.00269484519958496
17	96.49123	0.00348091125488281
25	94.73684	0.00203204154968262
35	94.73684	0.00197601318359375
227	68.42105	0.00429606437683105

k	acc	duration
1	91.22807	0.00374007225036621
3	91.22807	0.00258994102478027
5	91.22807	0.00268793106079102
7	91.22807	0.00367093086242676
16	91.22807	0.0043489933013916
17	91.22807	0.00435209274291992
25	89.47368	0.00270795822143555
35	87.71930	0.00239300727844238
227	68.42105	0.00619602203369141

Figure 2.1 - Tables showing the relation correct predictions (acc) and process duration in seconds for each chosen value of k – with/without normalization of variables

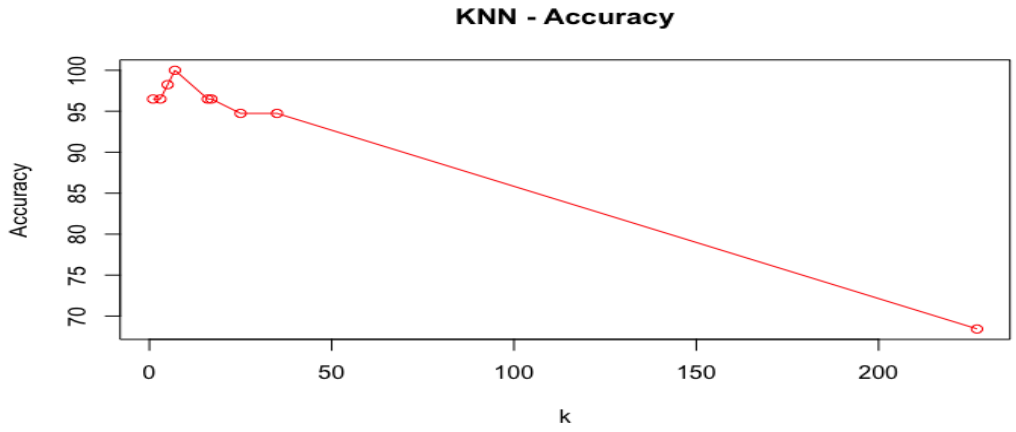

Figure 2.2 - KNN accuracy on Breast-Cancer dataset

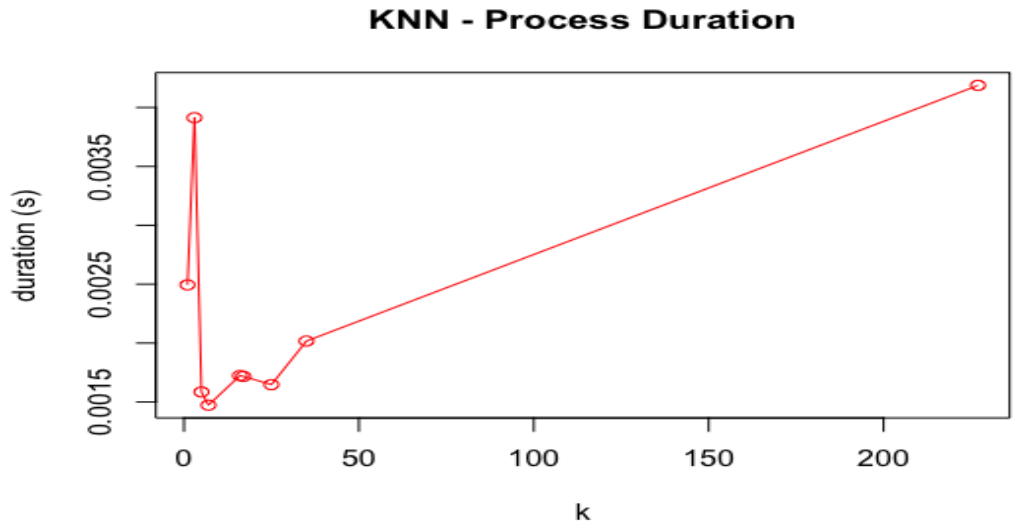

Figure 2.3 - KNN - Process duration

1.3. Classification on large datasets

1.3.1. Amazon reviews

The evaluation of the K-NN classifier on this large dataset is made by using the algorithm to predict which user hase made the review. The experiment is made using multiple values of parameter k. In order to get most accurate results, the numeric variables need to be normalized.
The presumption, that the greater k does not imply better results, is again confirmed. However, due to huge amount of classes (50), the algorithm shows really bad results (see figure 3.2)
As the diagramm on figure 3.3 states, the process duration increases with big values of k.
The normalization of the data showed no significat impact in this example.

k	acc	duration
1	18.000000	6.33231711387634
3	16.666667	5.87442111968994
5	19.333333	5.76916098594666
7	16.666667	5.69636392593384
17	18.666667	5.65344500541687
27	16.000000	5.66564083099365
35	14.666667	5.65485286712646
50	14.000000	5.75175189971924
69	10.000000	6.03331685066223
250	7.333333	6.89574790000916

k	acc	duration
1	18.000000	5.92937588691711
3	17.333333	5.81917095184326
5	18.666667	5.80785202980042
7	17.333333	5.95826697349548
17	17.333333	5.73219990730286
27	17.333333	5.6108660697937
35	16.000000	5.77083802223206
50	14.666667	5.85934686660767
69	11.333333	5.77278399467468
250	7.333333	5.67201995849609

Figure 3.1 - Tables showing the relation correct predictions (acc) and process duration in seconds for each chosen value of k – with/without normalization of variables

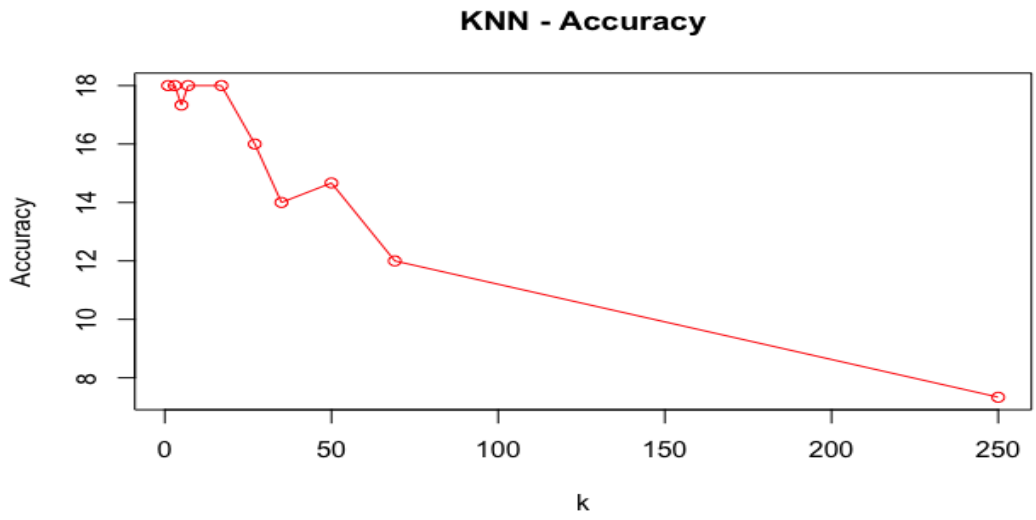

Figure 3.2 - KNN accuracy on Amazon-Reviews dataset

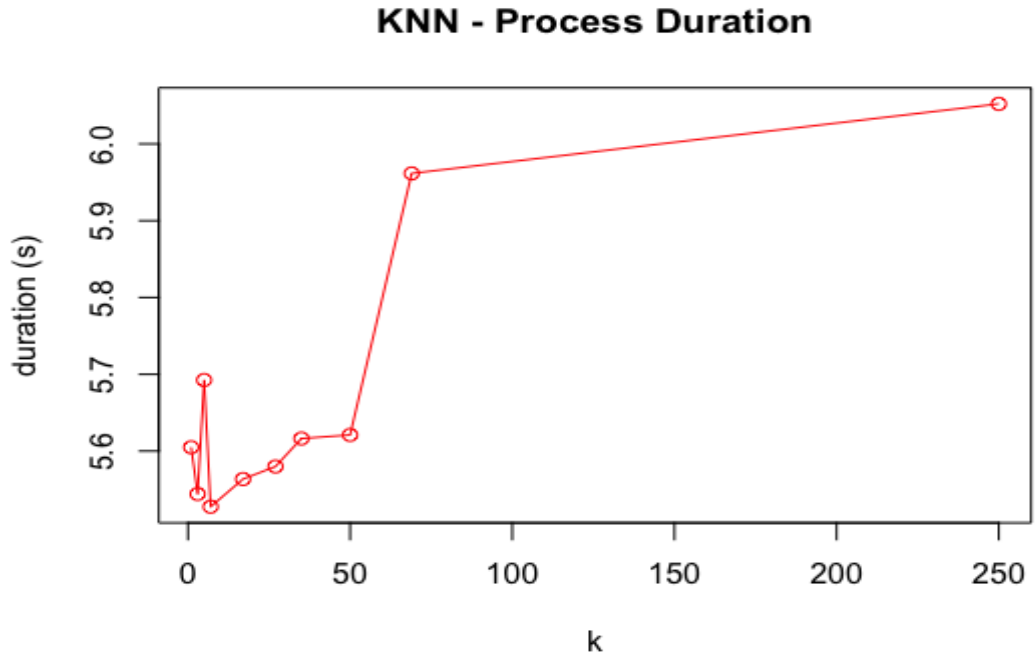

Figure 3.3 - KNN - Process duration

1.3.2. Synthetic Financial Datasets For Fraud Detection

The evaluation of the K-NN classifier on this large dataset is made by using the algorithm to predict if a financial transaction is a fraud or not . The experiment is made using multiple values of parameter k. In

order to get most accurate results, the numeric variables need to be normalized. However, since the dataset is pretty large, the K-NN algorithm did not manage to process the classification using k=1 in 12h.