Assignment 2

Victor Arvidsson
DV2599 Maskininlärning
Blekinge Tekniska Högskola
Karlskrona, Sweden
viar19@student.bth.se

I. Classifiers

In this assignment three different supervised classification models are compared using the spambase dataset. The three models that were tested are: K-Nearest neighbors, Gaussian Naive Bayes, and Random Forest.

A. K-Neares neigbors

The K-Neares neigbors(KNN) classifier works by using the classes of neighboring data points to the one you want to classify. By using a majority vote, the class for the new sample is chosen. For the tests in this report, a K-value of 5 was used. That means that the classifier will use the 5 closest data points to make the classification.

B. Gaussian Naive Bayes

The Gaussian Naive Bayesian(G-NB) classifier uses Bayesian statistics to calculate two Gaussian probability functions for the two classes. The classifier assumes independence between the data points, which greatly simplifies the calculations and therefore improves performance. Each new data point in the training data updates the posterior probability function for its class. Each predicted data point is classified according to the class that has the highest probability for that data point.

C. Random forest

Random forest(RF) uses decision trees to make predictions. It is an ensemble model that consists of many different decision trees trained on subsets of the data and using bootstrap. The decision trees are used to decide the classification using a majority vote.

II. Friedman and Nemenyi tests

The Friedman test is a non-parametric test used to compare means between data sets. The Nemenyi test is used to find the groups of data that differ from each other after the Friedman test has determined that there is a difference.

A. Friedman test

The Friedman test is used to compare the mean of three of more different data sets. The test is non-parametric, i.e. it works on data that is not distributed according to the normal distribution. For the test, the null hypothesis, H_0, is that there is no difference between the data sets. Therefore our hypothesis, H, is that there is a difference between the data sets. The Friedman test operates of the rank between the data sets, rather than on the values in the data sets. To calculate the Friedman statistic, we calculate three different values:

$$\overline{R} = \frac{1}{nk}\sum_{ij} R_{ij} = \frac{k+1}{2} \tag{1}$$

$$n\sum_{j}(R_j - \overline{R})^2 \tag{2}$$

$$\frac{1}{n(k-1)}\sum_{ij}(R_{ij} - \overline{R})^2 \tag{3}$$

where k is the number of data sets, n is the number of data points in each data set, R_{ij} is the rank of the data point in i-th data point in the j-th data set, and R_j is the average rank for data set j.

Equation (1) is the average rank of all the data. The Friedman statistic is calculated by the ration between equation (3) and (2). In order to assess whether the null hypothesis can be rejected, we compare the Friedman statistic with the critical value given by a pre-computed table [2]. If the Friedman statistics is larger than the critical value, we can reject the H_0 and accept our hypothesis H.

B. Nemenyi test

The Nemenyi test a post-hoc test that analyzes the data sets pairwise to find if there is a difference between them. This should only be used after it has been shown that there is a difference between any of the data sets using, for example, the Friedman test. The Nemenyi test calculates a critical difference using the formula

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6n}}$$

where q_α is defined as the studentized q range, based on α and k, divided by $\sqrt{2}$ [3]. Once we have obtained the critical difference, if the difference between the average rank of two data sets is larger than the critical difference then there is a statistically significant difference between the two data sets.

III. Result

A. Time

Time			
Data set	**KNN**	**G-NB**	**RF**
1	54.1 (2)	15.9 (1)	1115.0 (3)
2	41.8 (2)	10.3 (1)	943.9 (3)
3	40.4 (2)	13.1 (1)	1000.7 (3)
4	49.2 (2)	21.7 (1)	924.8 (3)
5	41.7 (2)	12.7 (1)	865.7 (3)
6	61.4 (2)	20.6 (1)	1018.7 (3)
7	49.9 (2)	14.8 (1)	915.8 (3)
8	44.0 (2)	15.5 (1)	945.6 (3)
9	54.7 (2)	20.4 (1)	1042.9 (3)
10	51.5 (2)	27.5 (1)	987.7 (3)
Avg	48.87 (2)	17.25 (1)	976.08 (3)

From this table, we get the following result

$$\overline{R} = 2$$
$$n\sum_j (R_j - \overline{R})^2 = 20$$
$$\frac{1}{n(k-1)}\sum_{ij}(R_{ij} - \overline{R})^2 = 1.$$

The Friedman statistic is $\frac{20}{1} = 20$, and the critical value for $\alpha = 0.05$, $k = 3$ and $n = 10$ is 6.2. Therefore, we reject H_0 and find that there is a difference between the data sets. For the Nemenyi test, the value of q_a for $k = 3$ and $\alpha = 0.05$ is 2.343. Therefore, we get that the critical difference is

$$CD = 2.343\sqrt{\frac{3 \cdot 4}{6 \cdot 10}} \approx 1.048$$

With this, we find that there is a significant difference between the Random forest and Gaussian Naive Bayes classifiers.

B. Accuracy

Accuracy			
Data set	**KNN**	**G-NB**	**RF**
1	74.0% (3)	84.4% (2)	95.2% (1)
2	77.6% (3)	86.3% (2)	95.2% (1)
3	77.8% (3)	87.8% (2)	93.3% (1)
4	82.2% (3)	86.7% (2)	94.8% (1)
5	81.1% (3)	88.5% (2)	95.9% (1)
6	82.2% (3)	82.8% (2)	95.7% (1)
7	82.0% (3)	83.3% (2)	97.0% (1)
8	82.8% (3)	86.7% (2)	97.4% (1)
9	72.4% (3)	83.5% (2)	89.1% (1)
10	76.3% (2)	71.7% (3)	86.7% (1)
Avg	78.84% (2.9)	84.17% (2.1)	94.03% (1)

From this table, we get the following result

$$\overline{R} = 2$$
$$n\sum_j (R_j - \overline{R})^2 = 18.2$$
$$\frac{1}{n(k-1)}\sum_{ij}(R_{ij} - \overline{R})^2 = 1.$$

The Friedman statistic is $\frac{18.2}{1} = 18.2$, and the critical value for $\alpha = 0.05$, $k = 3$ and $n = 10$ is 6.2. Therefore, we reject H_0 and find that there is a difference between the data sets. For the Nemenyi test, the value of q_a for $k = 3$ and $\alpha = 0.05$ is 2.343. Therefore, we get that the critical difference is

$$CD = 2.343\sqrt{\frac{3 \cdot 4}{6 \cdot 10}} \approx 1.048$$

. The resulting differences can be found in the table below.

	KNN	G-NB	RF
KNN	Not different	Not different	**Different**
G-NB	Not different	Not different	**Different**
RF	**Different**	**Different**	Not different

C. F1-score

F1-Score			
Data set	**KNN**	**G-NB**	**RF**
1	65.3% (3)	82.9% (2)	93.8% (1)
2	71.3% (3)	85.0% (2)	93.9% (1)
3	70.7% (3)	86.4% (2)	91.0% (1)
4	75.7% (3)	85.4% (2)	93.2% (1)
5	76.0% (3)	86.6% (2)	94.6% (1)
6	77.7% (3)	81.9% (2)	94.5% (1)
7	74.6% (3)	82.1% (2)	96.1% (1)
8	78.0% (3)	84.8% (2)	96.6% (1)
9	67.2% (2)	66.9% (3)	86.8% (1)
10	70.3% (3)	71.1% (2)	82.9% (1)
Avg	72.68% (2.9)	81.31% (2.1)	92.34% (1)

Since the resulting ranks are the same as for accuracy, except for data set 9 and 10 being switched, we will get exactly the same results as seen in section 2.B.

References

[1] P. Flach, "Machine Learning: The Art and Science of Algorithms that Make Sense of Data," Cambridge University Press, pp. 106–110, 2012.

[2] P. M. Lee, "Upper Critical Values for the Friedman Test,", Department of Mathematics, University of York, 2005, https://www.york.ac.uk/depts/maths/tables/friedman.pdf

[3] N. Kourentzes, "Critical values for the nemenyi test," https://kourentzes.com/forecasting/2014/05/01/critical-values-for-the-nemenyi-test/, 2014