Machine Learning Assignment-3

Sathvik Reddy Junutula (sxj174230)

Objectives: To run ANN, K-NN algorithms to achieve 2 class classification and compare the algorithms (Decision Trees, SVM, Ensemble Methods(Boosting) and report the algorithm that works better on two types of data sets.

Data description and reason for choosing:

1) The first dataset was provided by the instructor and it is used to apply two algorithms. The data is cleaned by removing unnecessary variables (the first two variables that had no role as variables. The variable shares are divided into two classes (high and low) based on the median value. Any other metric could also have been used but as the data is continuous, the median is used as the mean could be biased towards the high sharing values. Post the classification into two class labels as factors, the dependent variables are scaled to use it for both KNN & ANN to achieve the results in best and fast manner.
2) The other dataset is picked from UCI ML repository where the challenge is to identify whether the user defaults his credit card in the next month based on the previous months data. The reason for choosing this dataset is this type of problems are famous in the industry and the weight of the data is like that of the one given by the instructor (except number of variables). Also, the data is mix of continuous and categorical which makes the things a bit challenging. Again, the data is scaled except the class variable and it is then used for ANN and not K-NN.

Experimentation: Both the data sets were run with the two algorithms and the results were presented. The idea, process, learnings and the conclusions are provided individually for each experimentation below.

Artificial Neural Networks:

Keras and Tensorflow for R versions are used to run this experimentation and the following metrics are considered.

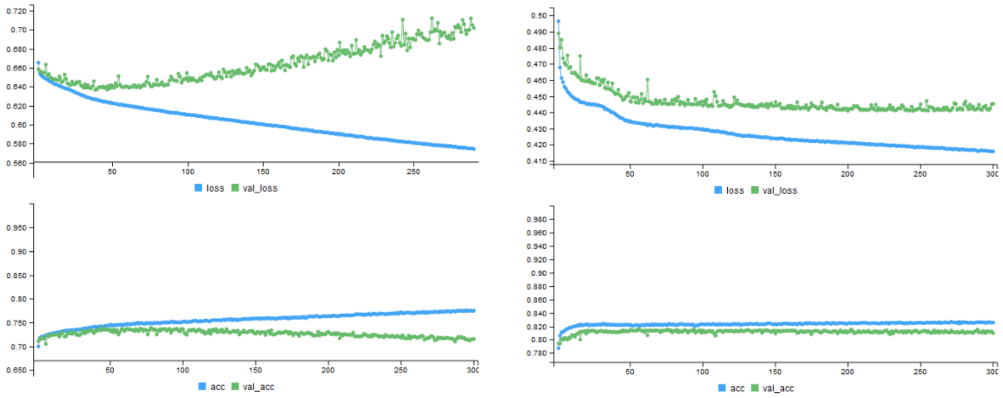
For each of the datasets below are the parameters that were used.

1) 5- Fold Validation (20% of the training data)
2) Batch size of 5 is given during the model building
3) Batch size of 128 is given during model evaluation for both train and test datasets
4) **Activations:**
   As the problem is a classification problem the sigmoid activation is used for the final activation of neuron. And each of Tanh, Relu & Sigmoid activations are used for first and second layer activation for both the data sets.
5) **Loss Evaluation:** For all the model building and evaluations, binary_crossentropy loss was used to assess the results and the optimization metrics for compiling the model used is accuracy.
6) **Layers & Nodes:** For each dataset the same set of Nodes and 2 Layers were used to perform the computations. The number of nodes for each dataset are given below

7) **Epochs:**

For each dataset an experimental epoch model is run with 100 epochs to approximate the best epoch number. For this sigmoid activation is used across all the layers and same Cross-Validation and Batch Sizes are used as mentioned above. The layer density units that were used are 30 & 20 respectively. Upon running the test epoch over the entire training data sets, the model is approximated for epochs by assessing the loss and accuracy for each epoch and a final epoch is reached by considering the epoch where the rate of validation accuracy is constant. They are shown graphically for both the datasets below. Also according to research and online articles, after a point the data overfits as the number of epoch's are increased so tried to use the as low number of epoch as possible.

Plot of Accuracy and Loss for the dataset(s) to finalize the epoch.



OnlineNewsPopularity

Credit

8)

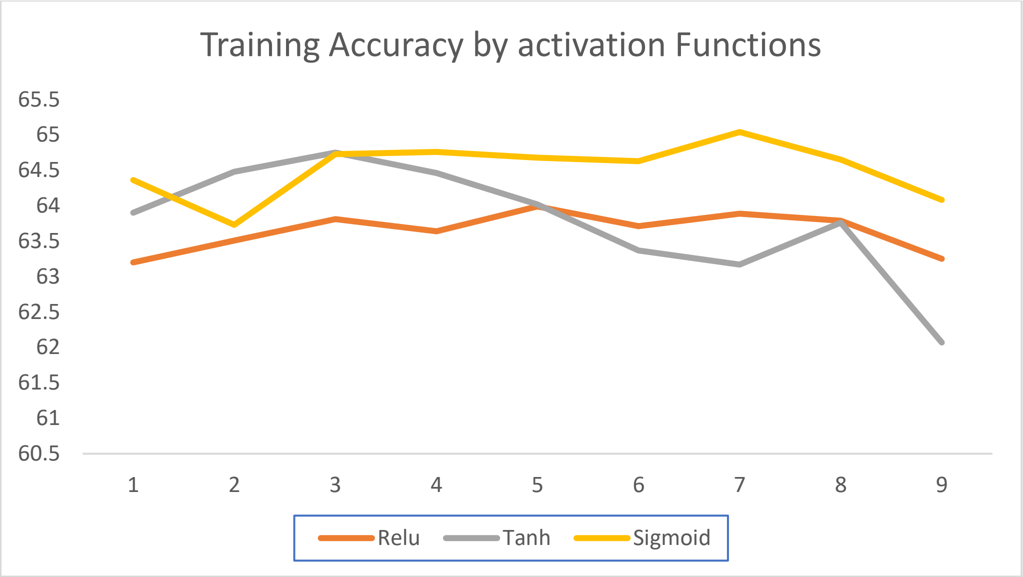
Node Pair ID	1	2	3	4	5	6	7	8	9
Layer1	5	8	11	14	17	20	23	26	29
Layer2	2	4	6	8	10	12	14	16	18

Note: They are mentioned as key-value pair above.

Approach on OnlineNewsPopularity Dataset:

Upon observing the graph above, it can be understood that for OnlineNewsPopularity dataset the loss for CV has a minimum at the nearest value approximately at an epoch of 25 and then it increases. Hence 25 is considered.

Upon experimenting the data with the 25 epochs 2 Layers and the above-mentioned Nodes and above-mentioned activation functions, below are the graph for each activation functions, with their training accuracy for each nodeId.



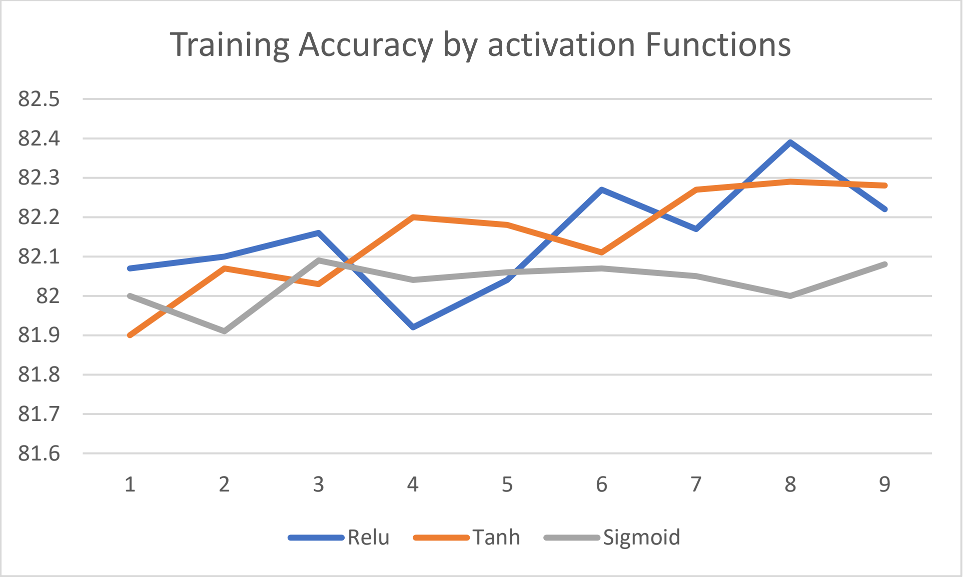
From this the best value of activation function and its respective key value pair for nodes is picked and is tested upon the test data to determine the test accuracy. The picked Key value Pair is **(23,14)** & the activation function is **SIGMOID.**

The best value of test accuracy is found to be **XXXXXX**

Approach on Credit Dataset:

Upon observing the graph above, it can be understood that for OnlineNewsPopularity dataset the loss for CV has a minimum at the nearest value approximately at an epoch of 30 and then it becomes constant. Hence 30 is considered.

Upon experimenting the data with the 30 epochs 2 Layers and the above-mentioned Nodes and above-mentioned activation functions, below are the graph for each activation functions, with their training accuracy for each nodeId.



From this the best value of activation function and its respective key value pair for nodes is picked and is tested upon the test data to determine the test accuracy. The picked Key value Pair is **(26,16)** & the activation function is **RELU.**

The best value of test accuracy is found to be **81.90**

The results from the above experimentation and the best values are summarized below.

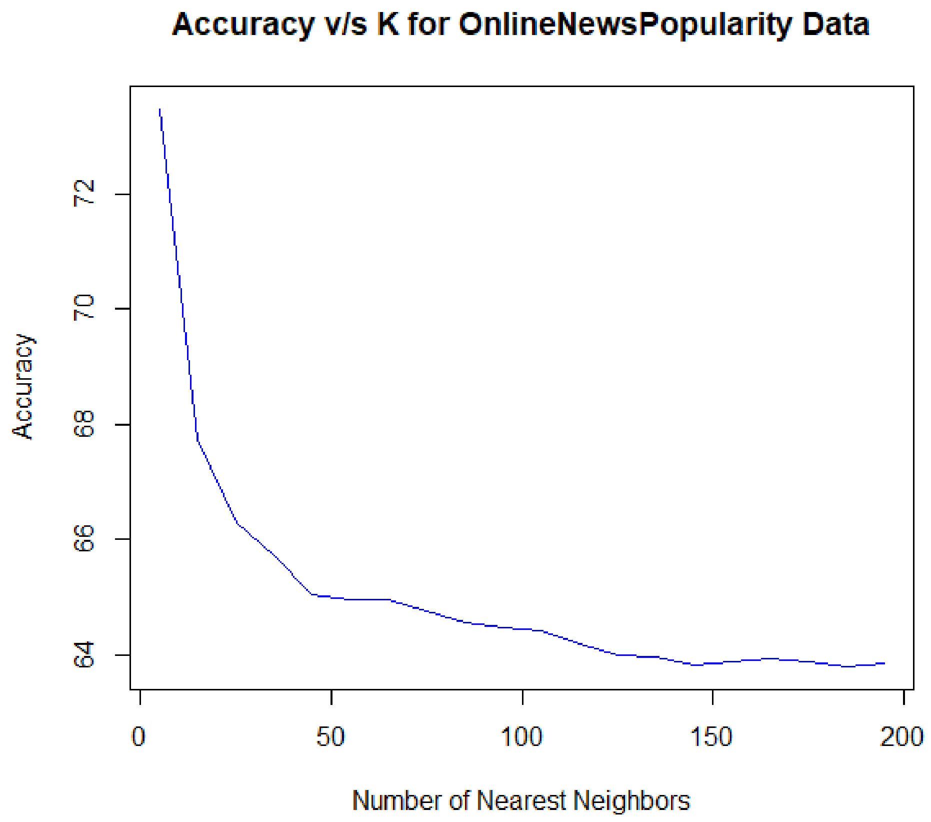
Dataset	Activation Function	Train Accuracy	Test Accuracy
OnlineNewsPopularity	Sigmoid	65.04	64.03
Credit	Relu	82.39	81.90

Experimentation 2: K-Nearest Neighbors :

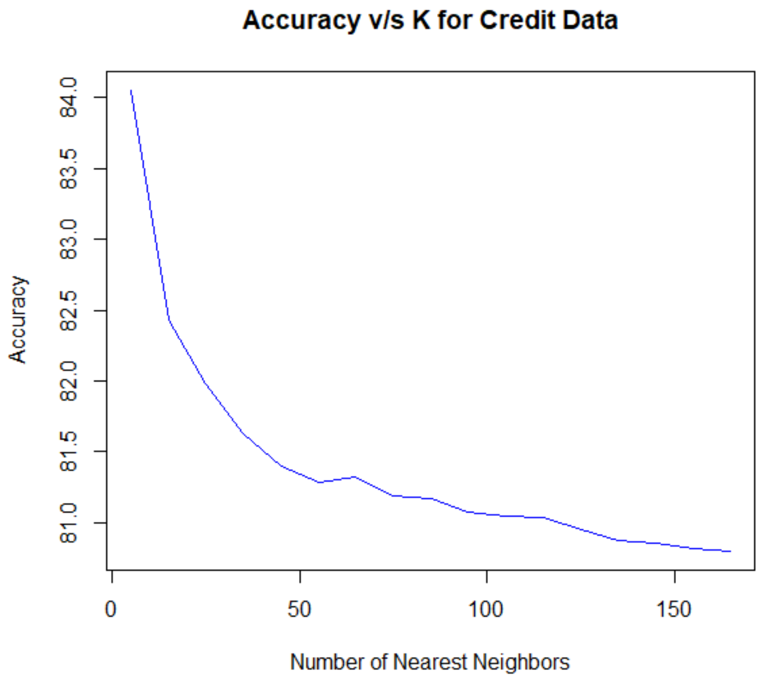
The cleaned data set(s) both Credit and OnlineNewsPopularity were taken and are used for running the algorithm.

The data is divided into 75-25 percentages and considered for training and testing purpose. Initially as the K-NN algorithm isn't that computationally expensive, the number of the nearest neighbours i.e, K values are considered from a range of (min,max) = (5,sqrt(number of total examples)). After considering the online research articles and some expert's opinions online the maximum number of K value is considered as sqrt(number of total examples). Post analysis further down it can be clearly observed that after a certain values of K = the sqrt(number of total example) the data tries to overfit the model and it's in our hand to pick the correct number of K-Value and it should be considered for evaluating the test data by use of the training accuracy graph.

Plot 2.1: Accuracy plot for training dataset for OnlineNewsPopularity data with respect to number of nearest neighbours:



Plot 2.2: Accuracy plot for training dataset for Credit data with respect to number of nearest neighbours:



Interpretations:

a) The accuracy for both the datasets is increasing right after a few iterations and it becomes nearly constant i.e, the rate of change of accuracy rate is very minimal

b) It can be interpreted from the Plot 2.1 & Plot 2.2 that the maximum accuracy is obtained at the nearest neighbour value i.e, K = 150, K = 125 for OnlineNewsPopularity and Credit data respectively.

c) Also, the test accuracy for OnlineNewsPopularity and Credit data are 63.17,80.59 respectively and the whole result is tabulated below and hence it is suggested to use 150 & 125 neighbours for OnlineNewsPopularity and Credit datasets.

Dataset	**Algorithm**	**K**	**Training Accuracy**	**Test Accuracy**
OnlineNewsPopularity	*KNN*	*150*	*63.84*	*63.17*
Credit	*KNN*	*125*	*80.95*	*80.59*

Conclusions:

Below is the quick snapshot of all the Algorithms that were used from the beginning and summarized to help pick which could be the better one along with the comments.

Algorithm	**OnlineNewsPopularity**	**Credit**	**Comments**
SVM	**65**	**81**	*Time Consuming*
Decision Trees	**63**	**81**	*High Supervision*
Ensemble(Boosting)	**66**	**82**	*Intermediate*
K-NN	**63**	**81**	*Unsupervised, Lazy*
ANN	**64**	**82**	*HighSupervision, Intelligent, Blackbox*

from the previous experimentation and analysis, it was advised to use Boosted version of a decision tree when a binary classification problem is to be solved for both the data sets. Now after introcuding KNN and ANN algorithm's K-NN does lack in performance when compared to the best of the previous, but the algorithm of Artificial Neural Networks is almost near to the performance of the Boosted version of the Decision tree. So from the numbers if the domain knowledge is available on the datasets we can clearly go with the Ensemble methods and if we are thrown in the dark with the data, ANN is suggested. But in the end, both the algorithm's take a tough fight.

References:

https://tensorflow.rstudio.com/keras/articles/sequential_model.html

https://www.datacamp.com/community/tutorials/tensorflow-tutorial

https://tensorflow.rstudio.com/keras/articles/about_keras_layers.html