Homework 1 Problem 1 – Varun Rao – varunr4

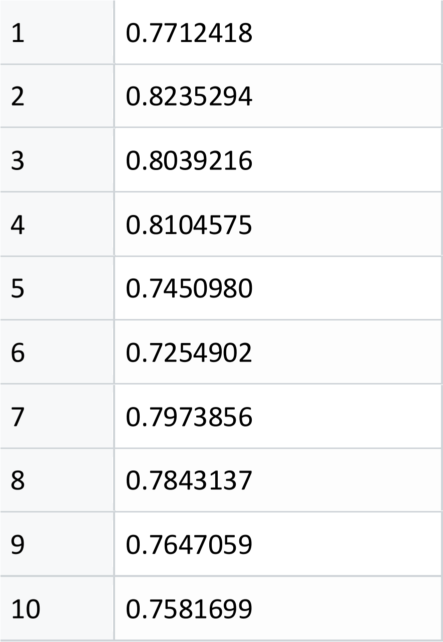
Data Used : http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

Part A – Naïve Bayes written using dnorm

Report the accuracy of the classifier on the 20% evaluation data, where accuracy is the number of correct predictions as a fraction of total predictions.

FileName : hw1-a.R

Accuracy: Overall accuracy : 77.8%



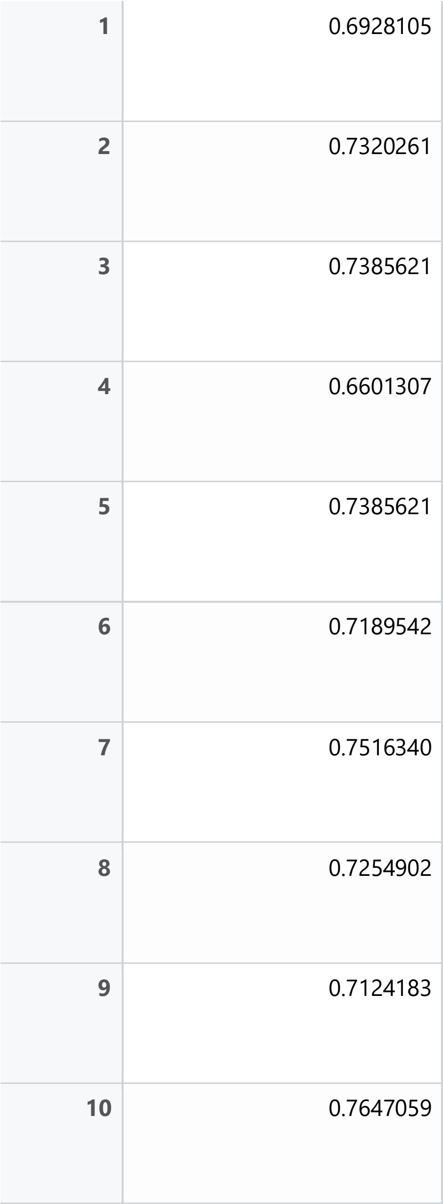
1	0.7712418
2	0.8235294
3	0.8039216
4	0.8104575
5	0.7450980
6	0.7254902
7	0.7973856
8	0.7843137
9	0.7647059
10	0.7581699

Part B – Naïve Bayes written with dnorm and using “NA” values

Report the accuracy of the classifier on the 20% that was held out for evaluation.

FileName: hw1-b.R

Accuracy : 72.35%



1	0.6928105
2	0.7320261
3	0.7385621
4	0.6601307
5	0.7385621
6	0.7189542
7	0.7516340
8	0.7254902
9	0.7124183
10	0.7647059

Part C – Naïve Bayes written using implementation from the caret package.

File Name : hw1-c.R

Accuracy : 79.08%

Part D – Classification using svmlight

File Name : hw1-d.R

Accuracy : 73.2%

Problem 2 –

Data Used from : http://yann.lecun.com/exdb/mnist/

Data file names :

train-images.idx3-ubyte

train-labels.idx1-ubyte

t10k-images.idx3-ubyte

t10k-labels.idx1-ubyte

Packages used :-

imager – resizing

quanteda – for the Bernoulli implementation

h2o – for random forests

Part A – Classifying MNIST using Naïve Bayes

Gaussian – Used the implementation from the caret package

Bernoulli – Used from quanteda package

FileName : hw2a.R (for untouched) & hw2.R (for stretched bounding box)

Accuracy	Gaussian	Bernoulli
Untouched images	50.68%	84.26%
stretched bounding box	83.12%	83.05%

For Untouched pixels the Bernoulli distribution is much better than the Gaussian distribution. Whereas for the stretched bounding box both the probability distributions are almost equal with Gaussian having a slight edge over the Bernoulli distribution.

Part B –

Random forest – random_forest.R and random_forest_untouched.R

Untouched pixels :

	depth = 4	depth = 8	depth = 16
#trees = 10	80.1%	90.8%	95.27%
#trees = 20	82.71%	91.86%	95.97%
#trees = 30	83.17%	92.52%	96.02%

Stretched bounding box :

	depth = 4	depth = 8	depth = 16
#trees = 10	82.26%	92.47%	95.69%
#trees = 20	83.44%	92.68%	96.42%
#trees = 30	83.71%	93.1%	96.67%

Citations :

Packages :

quanteda - https://cran.r-project.org/web/packages/quanteda/index.html

imager - https://cran.r-project.org/web/packages/imager/vignettes/gettingstarted.html

h2o - https://cran.r-project.org/web/packages/h2o/index.html

Links :

https://stackoverflow.com/questions/13172711/replace-na-values-from-a-column-with-0-in-data-frame-r

http://luthuli.cs.uiuc.edu/~daf/courses/AML-18/aml-home.html

https://www.kaggle.com/mlandry/random-forest-example

https://gist.github.com/brendano/39760

I must also cite the Piazza post by Paco Cruz to annotate the code fragment provided by the professor :

https://piazza.com/class/jchzguhsowz6n9?cid=63