CS 6375

ASSIGNMENT \_\_\_\_5\_\_\_\_\_

Names of students in your group:

HANVITHA GAVINI

PRATHYUSHA KANMANTH REDDY

Number of free late days used: \_\_\_\_\_\_0\_\_\_\_\_\_\_\_\_\_\_\_\_\_   
Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Please list clearly all the sources/references that you have used in this assignment.

**URL of dataset:**

<http://archive.ics.uci.edu/ml/machine-learning-databases/dermatology/dermatology.data>

No.of instances: 366

No.of attributes: 35

Folds: 10

**Preprocessing technique:**

1. Removed all “?” values by replacing them with NA
2. Eliminated the rows with NA
3. Scaled/normalized all the values between 0 and 1

**Evaluation metrics used:** Accuracy, Precision

**PseudoCode:**

numFolds = 10

split the data into k folds d[1…k]

for i in 1 to numFolds

// create training dataset by combining all folds except d[i]

train = {d[1] + d[2] + … + d[i-1] + d[i+1] + … + d[k]}

// create test dataset using d[i]

test = d[i]

for all classifiers c

// create a model of type c using train

model <- createModel(c, train)

// find accuracy of model of type c on test

for classifier c: accuracy[i] <- findAccuracy(model, test)

//create confusion matrix and find precision

for classifier c: other\_parameter[i] <- findPrecision(model, test)

next c

next i

print average accuracies

print precisions

**Results Table:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Classifier#** | **Classifier Name** | **Folds** | **Parameter 1** | **Parameter 2** | **Parameter 3** | **Parameter 4** | **Average Accuracy** | **Precision** |
| 1 | Decision Trees | 10 | Minsplit=3 | Minbucket=1 | Cp=0.001 | Maxdepth=30 | 94.69 | 87.05 |
|  |  |  | Minsplit=4 | Minbucket=2 | Cp=0.004 | Maxdepth=10 | 88.83 | 70.61 |
|  |  |  | Minsplit=2 | Minbucket=1 | Cp=0.001 | Maxdepth=30 | 82.12 | 68.44 |
| 2 | Perceptron | 10 | Threshold=0.1 | Error function=”ce” | Actual function=sigmoid |  | 55.26 | 28.58 |
|  |  |  | Threshold=0.02 | Error function=”ce” | Actual function=sigmoid |  | 56.68 | 27.11 |
|  |  |  | Threshold=0.003 | Error function=”ce” | Actual function=sigmoid |  | 56.39 | 27.51 |
| 3 | Neural Net | 10 | Hidden layers = (4,5,6) | Algorithm=”rprop+” |  |  | 66.48 | 28.04 |
|  |  |  | Hidden layers = (4,5,6) | Algorithm=”backprop” | Learning rate=0.3 | Threshold=0.1 | 31.01 | 281.6 |
|  |  |  | Hidden layers = (4,7,5,6) | Algorithm=”rprop+” | Learnin rate=0.1 | Threshold=0.1 | 66.77 | 34.9 |
| 4 | Deep Learning | 10 | Hidden layers = c(10,10,10,10,10,10,10,10) | Threshold=0.1 |  |  | 69 | 33.11 |
|  |  |  | hidden = c(12,12,112,12,20,30,3) | Threshold=0.05 |  |  | 67.02 | 45.59 |
|  |  |  | Hidden=c(12,12,112,12,20,30,30,10,20,20) | Threshold=0.1 |  |  | 59.39 | 38.58 |
| 5 | SVM | 10 | Kernel=”radial” | Cost=10 | Gamma=0.5 |  | 31.84 | 87.84 |
|  |  |  | Kernel=”polynomial” | Cost=100 | Gamma=0.5 |  | 3.91 | 2.70 |
|  |  |  | Kernel=”linear” | Cost=10 | Gamma=0.5 |  | 5.29 | 3.32 |
| 6 | naïve Bayes | 10 | Laplace=0 |  |  |  | 77.14 | 77.64 |
|  |  |  | Laplace=1 |  |  |  | 77.14 | 77.64 |
| 7 | Logistic Regression | 10 | Trace=F | Maxit=25 |  |  | 93.27 | 1.07 |
|  |  |  | Trace=T | Maxit=20 |  |  | 93 | 1.01 |
|  |  |  | Trace=T | Maxit=30 |  |  | 92.7 | 0.80 |
| 8 | k-Nearest Neighbors | 10 | K=2 | Prob=false |  |  | 97.22 | 91.1 |
|  |  |  | K=4 | Prob=false |  |  | 97.48 | 90.92 |
|  |  |  | K=4 | Prob=true |  |  | 97.47 | 91.08 |
| 9 | Bagging | 10 | Coob=true |  |  |  | 83.84 | 34.91 |
|  |  |  | Coob=true | Nbag=10 |  |  | 79.66 | 41.80 |
|  |  |  | Coob=false |  |  |  | 80.19 | 34.97 |
|  |  |  | Coob=false | Nbag=40 |  |  | 81.88 | 33.75 |
| 10 | Random Forests | 10 | Importance=true | Proximity=true | Ntree=500 |  | 90.82 | 91.48 |
|  |  |  | Importance=true | Proximity=true | Ntree=100 |  | 91.65 | 92.5 |
|  |  |  | Importance=true | Proximity=true | Ntree=50 |  | 90.26 | 91.27 |
|  |  |  | Importance=true | Proximity=false | Ntree=150 |  | 91.38 | 92.22 |
| 11 | AdaBoost | 10 | Iter=20 | Nu=1 | Type=”discrete” |  | 95.80 | 99.26 |
|  |  |  | Iter=10 | Nu=2 | Type=”discrete” |  | 90.81 | 119.47 |
|  |  |  | Iter=40 | Nu=4 | Type=”discrete” |  | 82.6 | 100.4 |
| 12 | Gradient boosting | 10 | Interaction.depth=2 | Bag.fraction=0.3 | Shrinkage=0.5 |  | 63.19 | 89.22 |
|  |  |  | Interaction.depth=3 | Bag.fraction=0.4 | Shrinkage=0.2 |  | 53.26 | 98.78 |
|  |  |  | Interaction.depth=10 | Bag.fraction=0.1 | Shrinkage=0.8 |  | 75.30 | 123.35 |

**Results with best parameters:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Classifier** | **Best Parameter Used** | **Accuracy** | **Precision** |
| Decision Trees | Minsplit=3, Minbucket=1, Cp=0.001 | 94.69 | 87.05 |
| Perceptron | Threshold=0.02, Error function=”ce”, Actual function= sigmoid | 56.68 | 27.11 |
| Neural Net | Hidden layers = (4,7,5,6), Algorithm=”rprop+”, Learnin rate= 0.1, Threshold=0.1 | 66.77 | 34.9 |
| Deep Learning | Hidden layers = c(10,10,10,10,10,10,10,10), Threshold=0.1 | 69 | 33.11 |
| SVM | Kernel=”radial”, Cost=10, Gamma=0.5 | 31.84 | 87.84 |
| naïve Bayes | Laplace=0 | 77.14 | 77.64 |
| Logistic Regression | Trace=F, maxit=25 | 93.27 | 1.07 |
| k-Nearest Neighbors | K=4, prob=false | 97.48 | 90.92 |
| Bagging | Coob=true | 83.84 | 34.91 |
| Random Forests | Importance=true, proximity=true, ntree=100 | 91.65 | 92.5 |
| AdaBoost | Iter=20, nu=1, type=”discrete” | 95.8 | 99.26 |
| Gradient Boosting | Interaction.depth=10, bag.fraction=0.1, shrinkage=0.8 | 75.3 | 123.35 |

**Analysis:**

We observed that K-NN is the best model for the data set we used.

Random forest, Decision tree and Adaboost provided best accuracy.

SVM gave the least accuracy.

After tuning the SVM model, we found the best cost and gamma as 10, 0.5 respectively and kernel to be radial.

Accuracy is a good metric compared to precision for our dataset for finding best parameters.