Harrison Kiang

hkiang2

Programming Exercise 2.1 - “Pima Indians Diabetes”

Note: priors are excluded as they are about the same, i.e., about half the samples are positive, the other half are negative.

a)

A Naive Bayes classifier was built explicitly in “a.R”. 10 trials were run, shown below:

> trscore

[1] 0.7528455 0.7788618 0.7691057 0.7430894 0.7528455

[6] 0.7447154 0.7642276 0.7560976 0.7626016 0.7528455

> trscoremean

[1] 0.7577236

> tescore

[1] 0.8039216 0.6601307 0.6732026 0.7581699 0.7712418

[6] 0.7385621 0.7320261 0.7189542 0.7254902 0.7516340

> tescoremean

[1] 0.7333333

Average accuracy across training data is about 75.77% and average accuracy across training

data is about 73.33%

b)

Attribute 3 (Diastolic blood pressure), attribute 4 (Triceps skin fold thickness), attribute 6 (Body mass index), and attribute 8 (Age) treat 0 values as missing or NA in R and are filtered from the data. 10 trials were run, shown below:

> trscore

[1] 0.7317073 0.7447154 0.7414634 0.7626016 0.7528455

[6] 0.7512195 0.7593496 0.7382114 0.7544715 0.7479675

> trscoremean

[1] 0.7484553

> tescore

[1] 0.7385621 0.7320261 0.6928105 0.7058824 0.7254902

[6] 0.7385621 0.6470588 0.7320261 0.6993464 0.6862745

> tescoremean

[1] 0.7098039

Average accuracy across training data is about 74.85% and average accuracy across training

data is about 70.98%.

c)

klaR and caret were used here with 10 resampling iterations to ensure cross-validation for each run on training and test data. The confusion matrix accuracies for 10 trials are shown below:

> trscore

[1] 0.7495935 0.7642276 0.7593496 0.7837398 0.7804878

[6] 0.7902439 0.7934959 0.7821138 0.7983740 0.7869919

> trscoremean

[1] 0.7788618

> tescore

[1] 0.7908497 0.7647059 0.7843137 0.7712418 0.7777778

[6] 0.7385621 0.7385621 0.6993464 0.7581699 0.7581699

> tescoremean

[1] 0.7581699

Average accuracy across training data is about 77.89% and average accuracy across training data is about 75.82%.

d)

SVM was used to classify the data. 10 sample iterations were run, results are shown below:

> trscore

[1] 0.7593496 0.7674797 0.7658537 0.7658537 0.7691057 0.7609756 0.7544715 0.7544715 0.7642276 0.7723577

> trscoremean

[1] 0.7634146

> tescore

[1] 0.7516340 0.7254902 0.7712418 0.7189542 0.7189542 0.7516340 0.7843137 0.7647059 0.7516340 0.7058824

> tescoremean

[1] 0.7444444

Average accuracy across training data is about 76.34% and average accuracy across training data is about 74.44%.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Accuracy/method | a (NB w/ NA) | b (NB w/out NA) | c (klaR & caret) | d (SVM) |
| training data | 75.77% | 74.85% | 77.89% | 76.34% |
| test data | 73.33% | 70.98% | 75.82% | 74.44% |