Report

Hsien-Ming Lee

Souradeep Sinha

Source Code: Driver.java

Dataset: reduced.banknote.data.arff

Performance Report: A3PerformanceResults(banknote).csv

Comparison Report: comp(banknote).csv

Classifier: IBk (k Nearest Neighbor classification)

k = 1

Statistics over all verification methods:

Mean of average resubstitution errors: 0, Mean of average generalization errors: 0.001

Comments: Performs well, given its resubstitution error and generalization error being zero or tending to zero, in all the verification methods. This is consistent with the fact that classifying with the most nearest one neighbor should be close to perfect, as there can be no ties, or misclassifications.

k = 3

Statistics over all verification methods:

Mean of average resubstitution errors: 0.001, Mean of average generalization errors: 0.002

Comments: Performs quite well, even though averages of resubstitution and generalization errors are not as good as of k = 1. Hints at overfitting, when verified by the Resampling method.

k=5

Statistics over all verification methods:

Mean of average resubstitution errors: 0.001, Mean of average generalization errors: 0.003

Comments: Performs quite well, and averages of resubstitution and generalization errors are comparable to lower values of k = 3. Hints at overfitting, when verified by the Resampling method.

k=10

Statistics over all verification methods:

Mean of average resubstitution errors: 0.002, Mean of average generalization errors: 0.004

Comments: Acceptable performance, and averages of resubstitution and generalization errors are a bit higher than lower values of k. Hints at overfitting, when verified by the Holdout and Resampling methods.

Classifier: J48 (Pruned C4.5 decision tree classification)

Minimum leaf size = 2

Statistics over all verification methods:

Mean of average resubstitution errors: 0.001, Mean of average generalization errors: 0.014

Comments: Difference in resubstitution and generalization error means there is a significant proof of overfitting over all verification methods. However, Holdout and Resampling show maximum overfitting, while Cross validation shows equal error rates and LOOCV does not exhibit any errors at all.

Minimum leaf size = 5

Statistics over all verification methods:

Mean of average resubstitution errors: 0.003, Mean of average generalization errors: 0.015

Comments: Lesser difference in resubstitution and generalization error means there is a lesser overfitting over all verification methods. However, Holdout shows maximum overfitting, while Cross validation shows equal error rates and LOOCV does not exhibit any errors at all.

Minimum leaf size = 10

Statistics over all verification methods:

Mean of average resubstitution errors: 0.007, Mean of average generalization errors: 0.02

Comments: Lesser difference in resubstitution and generalization error means there is a lesser overfitting over all verification methods. However, Holdout shows maximum difference, while Cross validation shows equal error rates and LOOCV does not exhibit any errors at all.

Minimum leaf size = 30

Statistics over all verification methods:

Mean of average resubstitution errors: 0.02, Mean of average generalization errors: 0.045

Comments: Best performance with least difference between error rates. This signifies that the minimum size of leaf is closer to the optimal range.

Classifier: Naïve Bayes Classification

Minimum leaf size = 2

Statistics over all verification methods:

Mean of average resubstitution errors: 0.05, Mean of average generalization errors: 0.07

Comments: Acceptable differences between error rates, however, error rates by themselves are higher than the prior classification methods, which is uncanny, because NBC is supposed to give optimal classification despite not considering attribute dependencies.

Classifier: SMO (Support Vector Machine classification)

Kernel: Polynomial, C: 1.0

Statistics over all verification methods:

Mean of average resubstitution errors: 0.006, Mean of average generalization errors: 0.01

Comments: Small difference between error rates. Less overfitting from hyperplanarity. Though generalization error doubles in the case of Holdout and Resampling verification processes.

Kernel: Polynomial, C: 5.0

Statistics over all verification methods:

Mean of average resubstitution errors: 0.004, Mean of average generalization errors: 0.007

Comments: Small difference between error rates with drop in error rates. Less overfitting from hyperplanarity and efficient classification. Though generalization error doubles in the case of Holdout and Resampling verification processes.

Kernel: Radial Basis Function, C: 1.0

Statistics over all verification methods:

Mean of average resubstitution errors: 0.09, Mean of average generalization errors: 0.17

Comments: Increased error rates suggest inefficient classification. Not a good choice overall.

Kernel: Radial Basis Function, C: 5.0

Statistics over all verification methods:

Mean of average resubstitution errors: 0.02, Mean of average generalization errors: 0.04

Comments: Better performer than C = 1.0.

-

Dataset: reduced.census.data.arff

Performance Report: A3PerformanceResults(census).csv

Comparison Report: comp(banknote).csv

Classifier: IBk (k Nearest Neighbor classification)

k = 1

Statistics over all verification methods:

Mean of average resubstitution errors: 0, Mean of average generalization errors: 0.06

Comments: Generalization error much more than resubstitution error, which suggests a significant model overfitting. Not a recommended k value.

k = 3

Statistics over all verification methods:

Mean of average resubstitution errors: 0.03, Mean of average generalization errors: 0.075

Comments: Much less overfitting than earlier value of k.

k=5

Statistics over all verification methods:

Mean of average resubstitution errors: 0.035, Mean of average generalization errors: 0.08

Comments: Performs at par with previous value of k.

k=10

Statistics over all verification methods:

Mean of average resubstitution errors: 0.04, Mean of average generalization errors: 0.076

Comments: Best performance, but an elevated resubstitution error may also suggest underfitting.

Classifier: J48 (Pruned C4.5 decision tree classification)

Minimum leaf size = 2

Statistics over all verification methods:

Mean of average resubstitution errors: 0.036, Mean of average generalization errors: 0.76

Comments: Good performance. Seems a good balance of both the errors.

Minimum leaf size = 5

Statistics over all verification methods:

Mean of average resubstitution errors: 0.04, Mean of average generalization errors: 0.08

Comments: Similar performance as that of previous k

Minimum leaf size = 10

Statistics over all verification methods:

Mean of average resubstitution errors: 0.04, Mean of average generalization errors: 0.08

Comments: Similar performance

Minimum leaf size = 30

Statistics over all verification methods:

Mean of average resubstitution errors: 0.05, Mean of average generalization errors: 0.08

Comments: Evidence of some underfitting.

Classifier: Naïve Bayes Classification

Minimum leaf size = 2

Statistics over all verification methods:

Mean of average resubstitution errors: 0.04, Mean of average generalization errors: 0.07

Comments: Good performance

Classifier: SMO (Support Vector Machine classification)

Kernel: Polynomial, C: 1.0

Statistics over all verification methods:

Mean of average resubstitution errors: 0.04, Mean of average generalization errors: 0.08

Comments:

Kernel: Polynomial, C: 5.0

Statistics over all verification methods:

Mean of average resubstitution errors: 0.04, Mean of average generalization errors: 0.08

Comments:

Kernel: Radial Basis Function, C: 1.0

Statistics over all verification methods:

Mean of average resubstitution errors: 0.066, Mean of average generalization errors: 0.11

Comments:

Kernel: Radial Basis Function, C: 5.0

Statistics over all verification methods:

Mean of average resubstitution errors: 0.046, Mean of average generalization errors: 0.08

Comments:

-

Dataset: reduced.census.data.arff

Performance Report: A3PerformanceResults(census).csv

Comparison Report: comp(banknote).csv

Classifier: IBk (k Nearest Neighbor classification)

k = 1

Statistics over all verification methods:

Mean of average resubstitution errors: 0, Mean of average generalization errors: 0.16

k = 3

Statistics over all verification methods:

Mean of average resubstitution errors: 0.08, Mean of average generalization errors: 0.21

k=5

Statistics over all verification methods:

Mean of average resubstitution errors: 0.1, Mean of average generalization errors: 0.21

k=10

Statistics over all verification methods:

Mean of average resubstitution errors: 0.12, Mean of average generalization errors: 0.021

Classifier: J48 (Pruned C4.5 decision tree classification)

Minimum leaf size = 2

Statistics over all verification methods:

Mean of average resubstitution errors: 0.04, Mean of average generalization errors: 0.186

Minimum leaf size = 5

Statistics over all verification methods:

Mean of average resubstitution errors: 0.08, Mean of average generalization errors: 0.2

Minimum leaf size = 10

Statistics over all verification methods:

Mean of average resubstitution errors: 0.1, Mean of average generalization errors: 0.21

Minimum leaf size = 30

Statistics over all verification methods:

Mean of average resubstitution errors: 0.13, Mean of average generalization errors: 0.22

Classifier: Naïve Bayes Classification

Minimum leaf size = 2

Statistics over all verification methods:

Mean of average resubstitution errors: 0.16, Mean of average generalization errors: 0.27

Classifier: SMO (Support Vector Machine classification)

Kernel: Polynomial, C: 1.0

Statistics over all verification methods:

Mean of average resubstitution errors: 0.135, Mean of average generalization errors: 0.22

Kernel: Polynomial, C: 5.0

Statistics over all verification methods:

Mean of average resubstitution errors: 0.13, Mean of average generalization errors: 0.22

Kernel: Radial Basis Function, C: 1.0

Statistics over all verification methods:

Mean of average resubstitution errors: 0.016, Mean of average generalization errors: 0.26

Kernel: Radial Basis Function, C: 5.0

Statistics over all verification methods:

Mean of average resubstitution errors: 0.015, Mean of average generalization errors: 0.25

-