Machine learning

Report

Baa Hamza
Ladjerem Abderrahmane

Part 1: using sklearn. tree.DecisionTreeClassifier

Part 1.1: Ntrain+Nval=1000, Nvalid=1000, Ntest=10000

 split the "train+validation" sets. Keep the test set for the **very** end:

- ratio_train = 0.016667
- ratio_valid = 0.016667
- ratio_test = 0.16667

 the validation accuracy for the best choice of max_depth is:

- valid score=0.7045, max_depth = 10

  Now, let's add some PCA as pre-processing

- Using max_depth=5, the best number of PCA components (nComp_PCA) to keep is 7 and we have training score: 0.6715, valid score: 0.6275
- Using max_depth=12, the best number of PCA components (nComp_PCA) to keep is 20 and we have training score: 0.9265, valid score: 0.6935

 The best (max_depth, nComp_PCA) pair:

- (max_depth=60, nComp_PCA =18) where we have training score: 1.0 . valid score: 0.696

  Can you explain the behavior of the optimal max_depth, let's call it m^* , with nComp_PCA, at **small** nComp_PCA ?

  Can you explain the behavior of the optimal max_depth, let's call it , with nComp_PCA, at **large** nComp_PCA ?

 Measure the cross-validation error for this best pair:

- [0.6625 0.695 0.6825 0.675 0.7125]
  mean: 0.685 (+/-: 0.017)

Part 1.2: Ntrain+Nval=2000, Nvalid=2000

- split the "train+validation" sets. Keep the test set for the **very** end:
    - ratio_train= 0.0333339
    - ratio_valid= 0.0333339

 The best (max_depth, nComp_PCA) pair:
    - (max_depth=60, nComp_PCA =20) where we have training score: 1.0 . valid score: 0.6925

 Measure the cross-validation error for this best pair:

    - [0.6625 0.7 0.7025 0.6775 0.715]
      mean: 0.691 (std: 0.019)

Part 1.3: Ntrain+Nval=20000, Nvalid=10000

- split the "train+validation" sets. Keep the test set for the **very** end:
    - ratio_train= 0.333339
    - ratio_valid= 0.16667

 The best (max_depth, nComp_PCA) pair:

    - (max_depth=20, nComp_PCA =27) where we have training score: 0.98055 . valid score: 0.7646

 Measure the cross-validation error for this best pair:

    - [0.731 0.738 0.7215 0.7355 0.717]
      mean: 0.729 (+/-: 0.008)

Part 1.4: The test (with `Ntest=10000`)

- Which model do you prefer, among the 3 "best models" you have found we choice logistic regression
- Using `Ntest=10000` samples that we saved preciously (and NEVER used), the test error:
- With DecisionTree :

  training score: 0.98015 . valid score: 0.76
  test score: 0.7527

- with LogisticRegression :

  training score: 0.8166 . valid score: 0.813
  test score: 0.804

- what is the level of accuracy you can achieve? :

- With Decision Tree:

Balanced accuracy score: 0.760298425462633

- with Logistic Regression:

Balanced accuracy score: 0.8141267241632418