Text Mining

Assignment 2

Kevin Zollicoffer

# 1)

See attached python project. Output from q1() is:

begin q1()

TMPY> Training decision tree

TMPY> Saving decision tree

TMPY> Evaluating classifier

TMPY> Computing evaluation metrics

Precision: [ 0.73724921 0.98850575 1. ]

Recall: [ 1. 0.95868852 0. ]

Accuracy: 0.961324631375

end q1()

# 2)

See attached python project. Output from q2() is:

begin q2()

TMPY> Training decision tree

TMPY> Saving decision tree

TMPY> Evaluating classifier

TMPY> Computing evaluation metrics

Precision: [ 0.73724921 0.98743802 1. ]

Recall: [ 1. 0.97934426 0. ]

Accuracy: 0.975586173556

TMPY> Training decision tree

TMPY> Saving decision tree

TMPY> Evaluating classifier

TMPY> Computing evaluation metrics

Precision: [ 0.73724921 0.97903989 1. ]

Recall: [ 1. 0.9495082 0. ]

Accuracy: 0.947788252357

TMPY> Training decision tree

TMPY> Saving decision tree

TMPY> Evaluating classifier

TMPY> Computing evaluation metrics

Precision: [ 0.73724921 0.98861734 1. ]

Recall: [ 1. 0.96819672 0. ]

Accuracy: 0.968334541939

end q2()

# 3)

# 4)

See attached python project. Output from q4() is:

begin q4()

TMPY> Training naive bayes

TMPY> Evaluating classifier

TMPY> Computing evaluation metrics

Precision: [ 0.73724921 0.98254282 1. ]

Recall: [ 1. 0.97803279 0. ]

Accuracy: 0.970993473532

end q4()

# 5)

See attached python project. Output from q5() is:

begin q5()

TMPY> Training linear model

TMPY> Evaluating classifier

TMPY> Computing evaluation metrics

Precision: [ 0.73724921 0.79359708 1. ]

Recall: [ 1. 0.99967213 0. ]

Accuracy: 0.8080734832

end q5()

# 6)

See attached python project. Output from q6() is:

begin q6()

TMPY> Extracting tfidf features

TMPY> Extracting tfidf features

TMPY> Training linear model

TMPY> Evaluating classifier

TMPY> Computing evaluation metrics

Precision: [ 0.73724921 0.84260028 1. ]

Recall: [ 1. 0.99868852 0. ]

Accuracy: 0.861493836113

end q6()

* Accuracy is higher
* Precision s lower
* Recall is lower

# 7)

Please see above output.

# 8)

See attached python project. Output from q8() is:

begin q8()

TMPY> Extracting frequent-k features

TMPY> Extracting frequent-k features

TMPY> Training decision tree

TMPY> Evaluating classifier

TMPY> Computing evaluation metrics

Precision: [ 0.73724921 0.86772152 1. ]

Recall: [ 1. 0.89901639 0. ]

Accuracy: 0.824510514866

TMPY> Training linear model

TMPY> Evaluating classifier

TMPY> Computing evaluation metrics

Precision: [ 0.73724921 0.73735409 1. ]

Recall: [ 1. 0.99409836 0. ]

Accuracy: 0.734590282814

end q8()

* Accuracy is lower
* Recall is similar
* Precision is tighter
* I would prefer the larger dictionary

# 9)

*k-nn*

# 10)

Using the robust loss function allows for an interpretation of an estimate of the statistical confidence of the classifier’s prediction.