Assignment 1

CS4442B

Albirawi, Zaid
250626065

Feb 12th, 2015

1. Implemented
2.
    a. Implemented
    b. k = 1

        CONF = 421 44
        104 426

        err = 0.1487

        k = 3

        CONF = 412 53
        107 423

        err = 0.1608

        k = 10

        CONF = 412 53
        163 367

        err = 0.2171

    c. 88 89 90

        CONF = 380 85
        130 400

        err = 0.2161

        using k = 1 the error with full features set is 0.1487 and the error using the features 88 89 and 90 is 0.2161. Therefore, the more features, the better.

3.

a. Implemented
b. Training:

CONF = 1009 73
148 932

err = 0.1022

Testing:

CONF = 425 40
65 465

err = 0.1055

c. Training:

CONF = 1058 24
42 1038

err = 0.0305

Testing:

CONF = 463 2
3 527

err = 0.0050

The error ratio is significantly smaller because with the quadratic function we use has way more features.

4.

a. Implemented
b. Implemented
c. k = 100

CONF = 452 13
463 67

err = 0.4784

k = 1000

CONF = 108 357
14 516

err = 0.3729

k = 10000

CONF = 215 250
16 514

err = 0.2673

k = 100000

CONF = 399 66
69 461

err = 0.1357

As the k value grows larger, the result becomes better, however, in 3.a only the result of k = 100000 was able to get close to it.

5.

a. Implemented
b. CONF = 287 178
102 428

err = 0.2814

c. 23, 9, 14, 56, 97

6.

a. Implemented
b. CONF = 154 77
102 129

err = 0.3874

It is almost impossible for us to implement a program that is able to pick out spam emails based on old data. Because the data is outdated, spammers try

not use the same words in their emails to help them escape the spam detectors.