

INTRODUCTION TO MACHINE LEARNING

HOMEWORK #4

REPORT

- 1- If we are given with "d" dimension and size of a data set N then we need to have d+1 variables. "d" comes from total dimensions we have and +1 is our bias.

"a quadratic programming problem with d+1 variables"

2-

- Accuracy = 34.4793% (692/2007) (classification)
- In sample error for 10 classifiers respectively:
0.030393622321873443
0.012954658694569009
0.046836073741903336
0.0752366716492277
0.08769307424015944
0.07972097658196313
0.08121574489287493
0.07274539113104135
0.08271051320378675
0.08570004982561036
- Total number of support vector machines: 6179

Highest in-sample error comes from the classifier **4 versus all classifier**.

- 3- Smallest in-sample error comes from the classifier **1 versus all classifier**.

- 5- Model accuracy with C = 0.0001

Accuracy = 27.3543% (549/2007) (classification)

In sample error for 10 classifiers respectively: 0.08320876930742402
0.003487792725460887 0.048829098156452415 0.08271051320378675
0.09965122072745392 0.07972097658196313 0.08470353761833582
0.07324364723467862 0.08271051320378675 0.08819133034379671
Total number of support vector machines: 7204

Model accuracy with C = 0.001

Accuracy = 32.0877% (644/2007) (classification)

In sample error for 10 classifiers respectively: 0.04285002491280518
0.009466865969108122 0.04235176880916791 0.07623318385650224
0.09965122072745392 0.07972097658196313 0.08470353761833582
0.07324364723467862 0.08271051320378675 0.08819133034379671

=====

Model accuracy with $C = 0.01$

Accuracy = 34.4793% (692/2007) (classification)

In sample error for 10 classifiers respectively: 0.030393622321873443
0.012954658694569009 0.046836073741903336 0.0752366716492277
0.08769307424015944 0.07972097658196313 0.08121574489287493
0.07274539113104135 0.08271051320378675 0.08570004982561036

=====

Model accuracy with $C = 0.1$

Accuracy = 39.2128% (787/2007) (classification)

In sample error for 10 classifiers respectively: 0.030393622321873443
0.01245640259093174 0.057299451918286 0.07125062282012955 0.09516691579471849
0.07972097658196313 0.08221225710014948 0.05680119581464873
0.08271051320378675 0.039860488290981565

=====

Model accuracy with $C = 1$

Accuracy = 39.2626% (788/2007) (classification)

In sample error for 10 classifiers respectively: 0.026905829596412557
0.01195814648729447 0.06527154957648232 0.07125062282012955
0.09516691579471849 0.07972097658196313 0.0817140009965122
0.039860488290981565 0.08271051320378675 0.05281514698555057

Maximum C achieves the smallest in-sample error.

6- Model accuracy with $C = 0.0001$

Accuracy = 32.7853% (658/2007) (classification)

In sample error for 10 classifiers respectively: 0.037867463876432486
0.006975585450921774 0.07224713502740408 0.0737419033383159 0.0931738913801694
0.07972097658196313 0.08470353761833582 0.06477329347284504
0.08271051320378675 0.07623318385650224

Total number of support vector machines: 6243

Model accuracy with C = 0.001

Accuracy = 34.8779% (700/2007) (classification)

In sample error for 10 classifiers respectively: 0.03089187842551071
0.009466865969108122 0.0787244643746886 0.07324364723467862
0.08918784255107125 0.07922272047832586 0.07922272047832586
0.060787244643746886 0.08271051320378675 0.06776283009466866

Total number of support vector machines: 6006

Model accuracy with C = 0.01

Accuracy = 35.8246% (719/2007) (classification)

In sample error for 10 classifiers respectively: 0.024414549078226207
0.009466865969108122 0.07972097658196313 0.07174887892376682
0.06676631788739412 0.07922272047832586 0.07922272047832586
0.060787244643746886 0.08271051320378675 0.08769307424015944

Total number of support vector machines: 5959

Model accuracy with C = 1

Accuracy = 37.7678% (758/2007) (classification)

In sample error for 10 classifiers respectively: 0.027902341803687097
0.007473841554559043 0.09018435475834578 0.060787244643746886
0.05580468360737419 0.07972097658196313 0.07623318385650224
0.05879422022919781 0.08271051320378675 0.08271051320378675

Total number of support vector machines: 6020

7- C = 1 is selected most often during Cross-Validation runs.

The screenshot shows the Spyder Python IDE interface. The editor window displays a script named 'temp.py' with the following code:

```
64 for line in file:
65     buffer = line.split()
66     y.append(int(float(buffer[0])))
67     x1 = (float(buffer[1]))
68     x2 = (float(buffer[2]))
69     #.append([float(buffer[1]),float(buffer[2])])
70     x.append([x1,x2])
71
72 with open("test.txt", "r") as file:
73     for line in file:
74         buffer = line.split()
75         y_test.append(int(float(buffer[0])))
76         x1 = (float(buffer[1]))
77         x2 = (float(buffer[2]))
78         #.append([float(buffer[1]),float(buffer[2])])
79         x_test.append([x1,x2])
80
81 problem = svm_problem(y,x)
82 model = svm_train(problem, '-c 0.01 -t 1 -d 2') # C=0.01 Q=2
83 p Labs, p_acc, p_vals = svm_predict(y_test, x_test, model)
84 p Labs = [round(x) for x in p Labs]
85 c0,c1,c2,c3,c4,c5,c6,c7,c8,c9 = calcInSample()
86 print("In sample error for 10 classifiers respectively: ", c0,c1,c2,c3,c4,c5,c6,c7,c8,c9)
87 print("Total number of support vector machines: ", model.get_nr_sv())
88 print("=====")
89 print("Model accuracy with C = 0.0001")
90 model = svm_train(problem, '-c 0.0001 -t 1 -d 2') # C=0.0001 Q=5
91 p Labs, p_acc, p_vals = svm_predict(y_test, x_test, model)
92 p Labs = [round(x) for x in p Labs]
93 c0,c1,c2,c3,c4,c5,c6,c7,c8,c9 = calcInSample()
94 print("In sample error for 10 classifiers respectively: ", c0,c1,c2,c3,c4,c5,c6,c7,c8,c9)
95 print("Total number of support vector machines: ", model.get_nr_sv())
96 print("=====")
97 print("Model accuracy with C = 0.001")
98 model = svm_train(problem, '-c 0.001 -t 1 -d 2') # C=0.001 Q=2
99 p Labs, p_acc, p_vals = svm_predict(y_test, x_test, model)
100 p Labs = [round(x) for x in p Labs]
101 c0,c1,c2,c3,c4,c5,c6,c7,c8,c9 = calcInSample()
102 print("In sample error for 10 classifiers respectively: ", c0,c1,c2,c3,c4,c5,c6,c7,c8,c9)
```

The Variable explorer on the right shows two variables: 'c6' (Float, 1, Value: 0.08470353761833582) and 'c7' (Float, 1, Value: 0.07324364723467862). The Python console on the right displays the output of the script, showing cross-validation accuracy for various C values (0.0001, 0.001, 0.01, 0.1, 1) and in-sample error for 10 classifiers. The output is as follows:

```
Cross Validation Accuracy = 41.8590%
Ecv for C = 0.0001:
Cross Validation Accuracy = 27.8288%
Ecv for C = 0.001:
Cross Validation Accuracy = 32.6018%
Ecv for C = 0.01:
Cross Validation Accuracy = 35.4135%
Ecv for C = 0.1:
Cross Validation Accuracy = 40.8174%
Ecv for C = 1:
Cross Validation Accuracy = 41.9833%
Ecv for C = 0.0001:
Cross Validation Accuracy = 27.8425%
Ecv for C = 0.001:
Cross Validation Accuracy = 32.6156%
Ecv for C = 0.01:
Cross Validation Accuracy = 35.3312%
Ecv for C = 0.1:
Cross Validation Accuracy = 40.6803%
Ecv for C = 1:
Cross Validation Accuracy = 42.0793%
Ecv for C = 0.0001:
Cross Validation Accuracy = 27.8151%
Ecv for C = 0.001:
Cross Validation Accuracy = 32.6156%
Ecv for C = 0.01:
Cross Validation Accuracy = 35.2489%
Ecv for C = 0.1:
Cross Validation Accuracy = 40.7352%
Ecv for C = 1:
Cross Validation Accuracy = 41.9147%
```

8-

9- Model accuracy with C = 0.01

Accuracy = 32.7354% (657/2007) (classification)

In sample error for 10 classifiers respectively: 0.03188839063278525

0.013452914798206279 0.04235176880916791 0.07722969606377678

0.09965122072745392 0.07972097658196313 0.08420528151469855

0.07324364723467862 0.08271051320378675 0.08819133034379671

Total number of support vector machines: 6472

=====

Model accuracy with C = 1

Accuracy = 37.9173% (761/2007) (classification)

In sample error for 10 classifiers respectively: 0.037867463876432486

0.013951170901843548 0.04783258594917788 0.06975585450921774

0.0981564524165421 0.0787244643746886 0.08370702541106129 0.05331340308918784

0.08271051320378675 0.05480817140009965

Total number of support vector machines: 5863

=====

Model accuracy with C = 100

Accuracy = 39.5615% (794/2007) (classification)

In sample error for 10 classifiers respectively: 0.03288490284005979

0.012954658694569009 0.05480817140009965 0.06925759840558046

0.0981564524165421 0.07822620827105133 0.0817140009965122 0.03288490284005979

0.08121574489287493 0.0622820129546587

Total number of support vector machines: 5692

=====

Model accuracy with C = 10000

Accuracy = 39.8605% (800/2007) (classification)

In sample error for 10 classifiers respectively: 0.03288490284005979

0.012954658694569009 0.06178375685102142 0.06776283009466866

0.09765819631290483 0.0737419033383159 0.0787244643746886 0.036372695565520675

0.0802192326856004 0.059292476332835076

Total number of support vector machines: 5694

=====

Model accuracy with C = 1000000

Accuracy = 40.2093% (807/2007) (classification)

In sample error for 10 classifiers respectively: 0.03238664673642252 0.0114598903836572

0.07174887892376682 0.07025411061285501 0.0787244643746886 0.06676631788739412

0.06975585450921774 0.060787244643746886 0.0787244643746886 0.057299451918286

Total number of support vector machines: 5677

In-sample error is lowest at C= 10^6.