CSL407-Machine Learning

HW4(Jaskaran Singh and Vishwash Batra)

(Refer to ReadMe for individual contributions.)

Ans2a)

Let the training set be $X = \{x_t, r_t\}_t \; where \; r^t = \{+1 \; if \; x^t \in C_1 \; or \; -1 \; if \; x^t \in C_2$

Find $w \; and \; w_0 \; such \; that$

$r^t(w^T x^t + w_0) \geq 1 - \varepsilon_t$

$where \; \varepsilon_t = 0 \; for \; correctly \; classified \; example \; far \; from \; the \; margin$

$0 < \varepsilon_t < 1 \; for \; a \; correctly \; classified \; example \; but \; inside \; the \; margin$

$and \; \varepsilon_t \geq 1 \; for \; an \; incorrectly \; classified \; example.$

Soft Error=$\sum \varepsilon_t$

We have to

$$\min \frac{1}{2}||w||^2 + C \sum \varepsilon_t \; subject \; to \; r^t(w^T x^t + w_0) \geq 1 - \varepsilon_t \; \forall \; t$$

Where $\varepsilon_t \geq 0 \; \forall \; t$

Using Lagrange multipliers $\alpha^t \geq 0 \; and \; \beta^t \geq 0$

The primal problem is then

$$L_p = argMin_{w,w_0,\varepsilon_t} Max_{\alpha^t,\beta^t} \frac{1}{2}||w||^2 + C\sum_t \varepsilon_t - \sum_t \alpha^t(r^t(w^T x^t + w_0) - 1 + \varepsilon_t) - \sum_t \beta^t \varepsilon^t$$

Applying KKT conditions,(n is the number of training examples)

1)$\frac{\delta L_p}{\delta w} = 0$

$$w = \sum_{t=1}^{n} \alpha^t r^t x^t$$

2)$\frac{\delta L_p}{\delta w_0} = 0$

$$\sum_{t=1}^{n} \alpha^t r^t = 0$$

3)$\frac{\delta L_p}{\delta \varepsilon_t} = 0$

$$C = \alpha^t + \beta^t$$

Substituting 1),2),3) in Lp we get the dual problem

$$L_d = -\frac{1}{2}\sum_{\substack{1\le t\le n\\1\le s\le n}} \alpha^t \alpha^s r^t r^s (x^t)^T x^s + \sum_{t=1}^{n} \alpha^t + (C - \alpha^t - \beta^t)\sum_{t=1}^{n} \varepsilon^t$$

$$= -\frac{1}{2}\sum_{\substack{1\le t\le n\\1\le s\le n}} \alpha^t \alpha^s r^t r^s (x^t)^T x^s + \sum_{t=1}^{n} \alpha^t$$

Where

$$\sum_{t=1}^{n} \alpha^t r^t = 0$$

Since $\beta^t = C - \alpha^t$ and $\beta^t \geq 0$

Thus, $C - \alpha^t \geq 0$

Thus,$0 \leq \alpha^t \leq C \ \ for\ all\ 1 \leq t \leq n$

Since,$\Phi$ is the feature transformation function that transforms $x^t$ into a high dimensional data point.

The dual problem in the transformed feature space is

$$L_d = -\frac{1}{2}\sum_{\substack{1\le t\le n\\1\le s\le n}} \alpha^t \alpha^s r^t r^s (\Phi(x^t))^T \Phi(x^s) + \sum_{t=1}^{n} \alpha^t$$

The inner product $(\Phi(x^t))^T \Phi(x^s)$ is equal to the kernel function K($x^t, x^s$).

Thus,

$$L_d = -\frac{1}{2}\sum_{\substack{1\le t\le n\\1\le s\le n}} \alpha^t \alpha^s r^t r^s K(x^t, x^s) + \sum_{t=1}^{n} \alpha^t$$

Where

$$\sum_{t=1}^{n} \alpha^t r^t = 0$$

Ans2

c)

Classification accuracy using linear kernel=51.899749

Classification accuracy using polynomial kernel=99.50

Time taken for training linear kernel=0.217371secs(keeps changing even when running with the same seed)

Time taken for training polynomial kernel=0.225488secs(keeps changing even when running with the same seed)

In general, time taken for linear kernel is less than a polynomial kernel.

Ans4)

a)

| Box Constraint Parameter C | Gaussian Kernel function width g | Average Cross Validation Accuracy(in %) |
|---|---|---|
| c=0.001000 | g=0.001000 | cv=69.860000 |
| c=0.001000 | g=0.010000 | cv=75.320000 |
| c=0.001000 | g=0.100000 | cv=75.460000 |
| c=0.001000 | g=1.000000 | cv=13.140000 |
| c=0.001000 | g=10.000000 | cv=40.020000 |
| c=0.001000 | g=100.000000 | cv=44.460000 |
| c=0.001000 | g=1000.000000 | cv=10.820000 |
| c=0.010000 | g=0.001000 | cv=69.860000 |
| c=0.010000 | g=0.010000 | cv=75.340000 |
| c=0.010000 | g=0.100000 | cv=75.460000 |
| c=0.010000 | g=1.000000 | cv=13.140000 |
| c=0.010000 | g=10.000000 | cv=15.180000 |
| c=0.010000 | g=100.000000 | cv=44.600000 |
| c=0.010000 | g=1000.000000 | cv=10.900000 |
| c=0.100000 | g=0.001000 | cv=73.000000 |
| c=0.100000 | g=0.010000 | cv=89.320000 |
| c=0.100000 | g=0.100000 | cv=88.760000 |
| c=0.100000 | g=1.000000 | cv=13.140000 |
| c=0.100000 | g=10.000000 | cv=14.180000 |
| c=0.100000 | g=100.000000 | cv=44.800000 |
| c=0.100000 | g=1000.000000 | cv=10.940000 |
| c=1.000000 | g=0.001000 | cv=88.820000 |
| c=1.000000 | g=0.010000 | cv=93.800000 |
| c=1.000000 | g=0.100000 | cv=96.540000 |
| c=1.000000 | g=1.000000 | cv=38.040000 |
| c=1.000000 | g=10.000000 | cv=11.880000 |
| c=1.000000 | g=100.000000 | cv=35.340000 |
| c=1.000000 | g=1000.000000 | cv=10.960000 |
| c=10.000000 | g=0.001000 | cv=92.620000 |
| c=10.000000 | g=0.010000 | cv=95.500000 |
| c=10.000000 | g=0.100000 | cv=96.620000 |
| c=10.000000 | g=1.000000 | cv=41.420000 |
| c=10.000000 | g=10.000000 | cv=12.260000 |
| c=10.000000 | g=100.000000 | cv=35.420000 |
| c=10.000000 | g=1000.000000 | cv=10.960000 |
| c=100.000000 | g=0.001000 | cv=93.760000 |
| c=100.000000 | g=0.010000 | cv=95.320000 |
| c=100.000000 | g=0.100000 | cv=96.620000 |
| c=100.000000 | g=1.000000 | cv=41.420000 |
| c=100.000000 | g=10.000000 | cv=12.260000 |
| c=100.000000 | g=100.000000 | cv=35.420000 |
| c=100.000000 | g=1000.000000 | cv=10.960000 |
| c=1000.000000 | g=0.001000 | cv=92.740000 |
| c=1000.000000 | g=0.010000 | cv=95.320000 |

| | | |
|---|---|---|
| c=1000.000000 | g=0.100000 | cv=96.620000 |
| c=1000.000000 | g=1.000000 | cv=41.420000 |
| c=1000.000000 | g=10.000000 | cv=12.260000 |
| c=1000.000000 | g=100.000000 | cv=35.420000 |
| c=1000.000000 | g=1000.000000 | cv=10.960000 |

The models giving the highest cross validation accuracy for the *mnist* dataset are

c=10.000000,g=0.100000,cv=96.62%(Taking this model for comparison with neural network from hw3)

c=100.000000,g=0.100000,cv=96.62%

c=1000.000000,g=0.100000,cv=96.62%

b)

One can use Micro averaged F-measure to compare between the best models of ANN and SVM.

One having the higher F-measure is a better classifier.

In case of SVM, we take the best model to be c=10,g=0.1

In case of ANN, we take the best model to be H=500.

On the same validation set:

Fsvm=.962376

Fmlp=.929703

Thus, SVM performs better than MLP on the validation set.