Spam detection

By

Akshay Rajput

Question 1:- Spam Classification using SVM.

This has four parts.

Part a:- Solve Dual objective of SVM using cvx package.

To solve this problem we needed to express the dual objective

$$max_\alpha \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} x_{(i)}^T x_{(j)}$$

in the form

$$\alpha^T Q \alpha + b^T \alpha + C$$

If we reduce in this form then,

$C = 0, b^T = 1, Q = y * y^T * x * x^T$

Now give this input to cvx using the constraints

$0 <= \alpha_i <= 1$ and $\alpha_i * g_i(w) = 0$

Part b:- Calculate w and b and report accuracy on test cases.

We can find w and b as following:-

$$w = \sum_{i=1}^{m} \alpha_i * y^{(i)} * x^{(i)}$$

$$b = -\frac{1}{2}(min_{y^{(i)}=1} w^T * x^{(i)} + max_{y^{(i)}=1} w^T * x^{(i)})$$

An example is correctly classified if $y^{(i)} * (W^T * x^{(i)} + b) > 0$ and hence accuracy is = (correctly classified/total) *100

There were two training set and the result of both is given below:-

small training set:-

Total Support vectors = 153
Accuracy = 91.40%

Large training set:-

Total Support vectors = 452
Accuracy = 98.70%

Part c:- Solve SVM problem using gaussian kernel.

The kernel matrix is calculated as follow:-

$$K = e^{(-1*\gamma*(x-z)^T*(x-z))}$$

$$w = \alpha * y * k(t_x, t_i)$$

$$b = 1 - \alpha_i * y_i * K(x_i, x)$$

small train

Total support vectors = 402
Accuracy = 89.4%

full train

Total support vectors = 1438
Accuracy = 97.6%

Part d:- Use libsvm to solve the dual problem using linear and gaussian problem.

This way is faster than using cvx as libsvm is optimised for SVMs and hence it takes less than one minute to compute.

The result of libsvm is given below:-

linear kernel

small training
optimization finished, #iter = 2069
Total support vectors = 152
Accuracy = 91.3%

full training
optimization finished, #iter = 37859
Total support vectors = 452
Accuracy = 98.7%

gaussian kernel

small train
optimization finished, #iter = 893
Total support vectors = 406
Accuracy = 90.4%

full train
optimization finished, #iter = 6906
Total support vectors = 1434
Accuracy = 98.6%