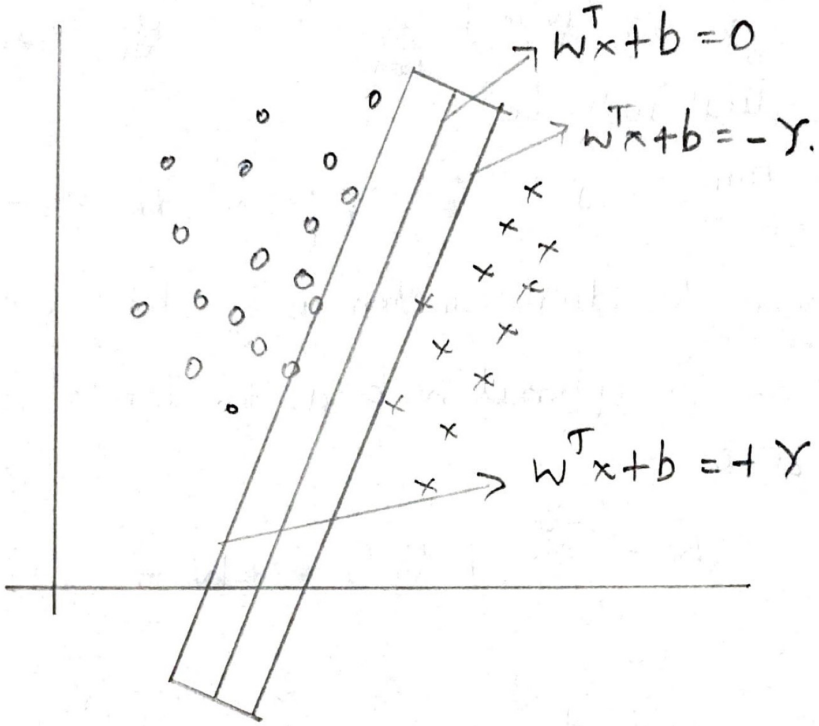
CS5590 – Foundation of Machine Learning
Assignment - 1

Tushar K Raysad (BM21MTECH14004)

1. Support Vector Machines:

In the derivation for the Support Vector Machine, we assumed that the margin boundaries are given by $w.x+b = +1$ and $w.x+b = -1$. Show that, if the +1 and -1 on the right-hand side were replaced by some arbitrary constants $+\gamma$ and $-\gamma$ where $\gamma > 0$, the solution for the maximum margin hyperplane is unchanged. (You can show this for the hard-margin SVM without any slack variables.)

→ Let us choose normalisation such that $w^T x_{(+)} + b = +\gamma$ & $w^T x_{(-)} + b = -\gamma$ for +ve & -ve support vectors respectively.



Hence the Margin will be,

$$+\gamma * \frac{w^T x + b}{\|w\|} + (-\gamma)\frac{w^T x + b}{\|w\|} = \frac{2\gamma}{\|w\|}$$

∴ We get margin $\rho = \frac{2}{\|w\|}$.

We should maximise, $\rho = \frac{2}{\|w\|}$ such that

$w^T x_i + b \geq \gamma$ if $y_i = +1$; $w^T x_i + b \leq -\gamma$ if $y_i = -1$.

As $\min \|w\| = \max \frac{1}{\|w\|}$

$\frac{1}{2} w^T . w$ is to be minimised for all (x_i, y_i)

such that $y_i(w^T x_i + b) \geq \gamma$

Lagrangian for the above function will be

$$L(w,b,\alpha) = \frac{1}{2}\|\vec{w}\|^2 + \sum_{i=1}^{n} \alpha_i \left(Y - y_i(\vec{w}\,\vec{x_i} + b)\right)$$

So, the primal for the above function will be,

$$\min_{\vec{w},b} \max_{\alpha_i \geq 0} \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i \left(Y - y_i(\vec{w}.\vec{x_i} + b)\right)$$

∴ The dual will be,

$$\max_{\vec{\alpha_i} \geq 0} \min_{\vec{w},b} \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i \left[(\vec{w}x_i + b)y_i - Y\right]$$

∴ Solving the dual function using KKT condition and Solving for optimal w, b as function of α and equating to zero,

$$\frac{\partial L}{\partial w} = w - \sum_{j=1}^{n} \alpha_j y_j x_j \Rightarrow w \sum_j \alpha_j y_j x_j$$

$$\frac{\partial L}{\partial b} = -\sum_j \alpha_j y_j \Rightarrow \sum_j \alpha_j y_j = 0.$$

Substituting these values in dual we get,

$$\Sigma \alpha_j - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j (x_i \cdot x_j)$$

The above eqn is the same equation that we get when we consider the margin equation to be $w^T x + b = +1$ and $w^T x + b = -1$.

Hence, if we substitute Y in place of 1 we will get the same solution.

2. Support Vector Machines:

Consider the half-margin of maximum margin SVM defined by ρ, i.e, $\rho = \frac{1}{\|w\|}$.

Show that ρ is given by:

$$\frac{1}{\rho^2} = \sum_{i=1}^{n} \alpha_i$$

where α_i are the Lagrange Multipliers given by the SVM dual. (as on slide 30 of the SVM lecture uploaded on Piazza).

Given :- $\rho = \frac{1}{w}$ To find $\frac{1}{\rho^2} = \sum_{i=1}^{n} \alpha_i$

We have $\frac{1}{\rho^2} = \|w\|^2$.

Consider the Dual (in Slide 30).

$$\max_{\alpha_i \geq 0} \min_{\vec{w}, b} \frac{1}{2}\|\vec{w}\|^2 - \sum \alpha_i \left[(\vec{w}x + b)y_i - 1\right].$$

we get $w = \sum_j \alpha_j y_j x_j$ & $\sum \alpha_j y_j = 0$.

Substituting these values we get,

$$L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

Such that $\alpha_i \geq 0$ $\quad \sum_{i=1}^{n} \alpha_i y_i = 0$

Consider $\bar{Y}(x) = w^T x + b$.

$$Y(x) = \sum_{d=1}^{n} \alpha_d y_d (x_d \cdot x) + b.$$

To satisfy the KKT condition the equation should hold 3 properties. They are

$\alpha_i \geq 0$, $y_i \bar{Y}(x) - 1 \geq 0$

$\alpha_i (y_i \bar{Y}(x) - 1) = 0$.

So we get $y_i Y(x) = 1$.

Consider Lagrangian $L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{2} \alpha_i \left[(\vec{w} \cdot \vec{x}_i + b) y_i - 1\right]$

The second term in this lagrangian will vanish as

$y_i Y(x) = 1$.

So $L(w, b, \alpha) = \frac{1}{2}\|w\|^2$.

Using this value and $w = \sum \alpha_i y_i x_i$ in the dual we get

$$\frac{1}{2}\|w\|^2 = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\|w\|^2$$

$$\Rightarrow \|w\|^2 = \sum_{i=1}^{n} \alpha_i$$

$$\therefore \boxed{\frac{1}{\rho^2} = \sum_{i=1}^{n} \alpha_i}$$

3. Kernels: (5 marks) Let k1 and k2 be valid kernel functions. Comment about the validity of the following kernel functions, and justify your answer with proof or counter-examples as required:

(a)k(x, z) = k1(x, z) + k2(x, z)

(b) k(x, z) = k1(x, z)k2(x, z)

(c) k(x, z) = h(k1(x, z)) where h is a polynomial function with positive coefficients.

(d) k(x, z) = exp (k1(x, z))

(e) k(x, z) = exp (–\|\|x - z\|\|2 \ σ2)

→ a] $K(x,z) = K_1(x,z) + K_2(x,z)$

We have, $K_1(x,z) = \psi_1(x)\,\psi_1(z)$

& $K_2(x,z) = \psi_2(x)\,\psi_2(z)$

Then,

$$K(x,z) = K_1(x,z) + K_2(x,z)$$

$$= [\psi_1(x)\,\psi_1(z)] + [\psi_2(x)\,\psi_2(z)]$$

$$= [\psi_1(x)\,\psi_2(x)], [\psi_1(z)\,\psi_2(z)]$$

$$= [\psi(x) \cdot \psi(z)]$$

As $K(x,z)$ can be expressed as inner dot product and is a symmetric matrix, we can say that given kernel is valid.

b] $K(x,z) = K_1(x,z)\,K_2(x,z)$

We have, $K_1(x,z) = \psi_1(x)\,\psi_1(z)$

$K_2(x,z) = \psi_2(x)\,\psi_2(z)$

Then, $K(x,z) = [\psi_1(x)\,\psi_1(z)]\,[\psi_2(x)\,\psi_2(z)]$

$= [\psi_1(x)\,\psi_2(x)]\,[\psi_1(z)\,\psi_2(z)]$

$K(x,z) = [\psi(x)\ \psi(z)]$

This gram matrix K is the Hadamard product or element-by-element product of K_1 & K_2. If we consider K_1 and K_2 to be the covariance matrices of $(X_1, \ldots X_n)$ and $(Y_1, \ldots Y_n)$ respectively, then K will be covariance matrix of $(X_1 Y_1 \ldots X_n Y_n)$ which proves that it is symmetric and positive definite matrix. Hence it is a valid function.

C] $K(x,z) = h(K_1(x,z)$ where h is a polynomial function with positive coeffecients.

As each polynomial terms are the product of kernel functions having positive co-effecients, we can prove by applying element-by-element product as shown in questions a and b and by addition of kernels. So $K(x,z) = h(K_1(x,z))$ is a valid kernel function.

d) $K(x, z) = \exp(K_1(x, z))$

We have $e^x = \lim_{x \to \infty} \left[1 + x + \dots + \frac{x_i}{i!}\right.$

From this equation,

$\frac{x_i}{i!}$ terms can be considered as polynomial.

Now we can apply the $K(x, z) = h[K_1(x, z)]$ and apply the limits to the Kernel matrix, we get $K(x, z) = \lim_{i \to \infty} K_i(x, z)$

Since the gram matrix in above sequence is positive semi-definite this is a valid kernel function.

e) $K(x, z) = \exp\left(\frac{-\|x - z\|^2}{\sigma^2}\right)$

$= \exp\left[\frac{-\|x\|^2 - \|z\|^2 + 2x^T z}{\sigma^2}\right]$

$= \left[\exp\left(\frac{-\|x\|^2}{\sigma^2}\right) \exp\left(\frac{-\|z\|}{\sigma^2}\right) \exp\frac{2x^T z}{\sigma^2}\right]$

$= g(x)\, g(z)\, \exp(K_1(x, z))$

Since $K_1(x, z)$ is a kernel function we can say that the product will be kernel function and hence it is valid.

4. SVMs:

In this question, you will be working on a soft-margin SVM. You may find it helpful to review the Scikit Learn's SVM documentation: http://scikit-learn.org/stable/modules/svm.html. We will apply soft-margin SVM to handwritten digits from the processed US Postal Service Zip Code data set. The data (extracted features of intensity and symmetry) for training and testing are available at:

- ***http://www.amlbook.com/data/zip/features.train***
- ***http://www.amlbook.com/data/zip/features.test***

In this dataset, the 1st column is digit label and 2nd and 3rd columns are the features. We will train a one-versus-one (one digit is class +1 and another digit is class -1) classifier for the digits '1' (+1) and '5' (-1). (In the original dataset, only consider data samples(rows) with the label as either 1 or 5, for both train and test settings. Then for training details, you may find this link at http://scikit-learn.org/stable/modules/svm.html helpful.)

(a)Consider the linear kernel K(xn, xm) = x T n xm. Train using the provided training data and test using the provided test data, and report your accuracy over the entire test set, and the number of support vectors.

(b) In continuation, train only using the first {50, 100, 200, 800} points with the linear kernel. Report the accuracy over the entire test set, and the number of support vectors in each of these cases.

(c) Consider the polynomial kernel K(xn, xm) = (1 + x T n xm) Q, where Q is the degree of the polynomial. Comparing Q = 2 with Q = 5, comment whether each of the following statements is TRUE or FALSE.

i. When C = 0.0001, training error is higher at Q = 5.
ii. When C = 0.001, the number of support vectors is lower at Q = 5.
iii. When C = 0.01, training error is higher at Q = 5.
iv. When C = 1, test error is lower at Q = 5.

(d) Consider the radial basis function (RBF) kernel in the soft-margin SVM approach. Which value of C ∈ {0.01, 1, 100, 104 , 106} results in the lowest training error? The lowest test error? Show the error values for all the C values.

(a) Accuracy over the entire test set = 0.9787735849056604

Number of Support Vectors = [14+14] = 28.

(b) For dataset of 800 samples:

Accuracy= 0.9811320754716981

Number of Support vectors= [7+7] = 14.

For dataset of 200 samples:

Accuracy= 0.9811320754716981

Number of Support vectors= [4+4] = 8.

For dataset of 100 samples:

Accuracy= 0.9811320754716981

Number of Support vectors= [2+2] = 4.

For dataset of 50 samples:

Accuracy= 0.9811320754716981

Number of Support vectors= [1+1] = 2.

(c) i) When C = 0.0001, training error is higher at Q = 5.

- This statement is *FALSE*. Training error is higher at Q = 2.

ii) When C = 0.001, the number of support vectors is lower at Q = 5.

- This statement is *TRUE*.

iii) When C = 0.01, training error is higher at Q = 5.

- This statement is *FALSE.* Training error is higher at Q = 2.

iv) When C = 1, test error is lower at Q = 5.

- This statement is *FALSE*. Test error is lower at Q = 2.

(d) For C=0.01:

training error - 0.0051249199231262 48 ,

test error - 0.016509433962264 12

For C=1:

training error - 0.0044843049327354 39 ,

test error - 0.021226415094339646

For C=100:

training error - 0.0032030749519538215 ,

test error - 0.018867924528301883

For C=10000:

training error - 0.002562459961563124 ,

test error - 0.018867924528301883

For C=1000000:

training error - 0.002562459961563124 ,

test error - 0.02358490566037741

Test error is least at C=0.01

Training error is least at C=10000 and C=1000000

5. SVMs (contd):

GISETTE (https://archive.ics.uci.edu/ml/ datasets/Gisette) is a handwritten digit recognition problem. The problem is to separate the highly confusible digits '4' and '9'. This dataset is one of five datasets of the NIPS 2003 feature selection challenge.

The dataset for this problem is large, so please budget time accordingly for this problem.

(a)Standard run: Use all the 6000 training samples from the training set to train the model, and test over all test instances, using the linear kernel. Report the train error, test error, and number of support vectors.

(b) Kernel variations: In addition to the basic linear kernel, investigate two other standard kernels: RBF (a.k.a. Gaussian kernel; set $\gamma = 0.001$), Polynomial kernel (set degree = 2, coef0 = 1; e.g, $(1 + x^T x)^2$). Which kernel yields the lowest training error? Report the train error, test error, and number of support vectors for both these kernels.

(a) For linear kernel :

training error= 0.0

test error= 0.02400000000000002

Number of support vectors= [542+542] = 1084.

(b) RBF Kernel:

training error= 0.0

test error= 0.5

Number of support vectors= [3000+3000] = 6000.

Polynomial Kernel:

training error= 0.00049999999999999449

test error= 0.02000000000000018

Number of support vectors= [641+691] = 1332.

RBF Kernel yields the lowest training error i.e. 0.