CSCI 567: Homework 3

Deepika Anand

October 17, 2016

Problem 1. (a)

$$\beta_\lambda = (X^T X +_N)^{-1} X^T y \tag{1}$$

As given y = x$\beta^* + \epsilon$ .Hence combining the two

$$\beta_\lambda = (X^T X +_N)^{-1} X^T (x\beta^* + \epsilon) \tag{2}$$

Using affine transformation of Gaussian Random variable. We can say β_λ is normally distributed with distribution

$$N((X^T X + \lambda I)^{-1}(x\beta^*), (X^T X + \lambda I)^{-1} X^T X (X^T X + \lambda I)^{-1}) \tag{3}$$

Problem 1. (b)

$$x^T E[(\hat{\beta_\lambda} - \beta^*)] = x^T((X^T X + \lambda I)^{-1} X^T X \beta - \beta) \tag{4}$$

This can be re-written as

$$x^T E[(\hat{\beta_\lambda} - \beta^*)] = x^T((X^T X + \lambda I)^{-1} X^T X - I)\beta \tag{5}$$

Problem 1. (c)

Since $x^T(\hat{\beta_\lambda} - E[\hat{\beta_\lambda}])$ is zero mean Gaussian random variable. Hence we can say that variance term is $x^T(X^T X + \lambda I)^{-1} X^T X (X^T X + \lambda I)^{-1} x$
So the variance term can be written as

$$||X(X^T X + \lambda I)^{-1} x||^2 \tag{6}$$

Problem 1. (d)

As complexity increases variance increases and bias decreases

Problem 2. (a)

For any function F(.) if K is valid kernel then,

$$\int_{x,x'} f(x)f(x')k(x,x')dxdx' >= 0 \tag{7}$$

So, k_3 can be replaced with $a_1k_1 + a_2k_2$

$$\int_{x,x'} f(x)f(x')[a_1k_1(x.x') + a_2k_2(x.x')]dxdx' \tag{8}$$

this can be rewritten as

$$\int_{x,x'} f(x)f(x')[a_1k_1(x.x')]dxdx' + \int_{x,x'} f(x)f(x')[a_2k_2(x.x')]dxdx' \tag{9}$$

Since both the component are greater than zero so the addition of the two will be greater than 0, hence we can write k_3 is valid

Problem 2. (b)

As k_4 can be written as NXN matrix for the points f(x_1).....f(x_n)
$k_4 = V^TFF^TV$
Now combing terms it can be written as $k_4 = (F^TV)^TF^TV$
or $k_4 = ||F^TV||^2$
Since square of a scalar quantity is always greater than 0. Hence k_4 is valid kernel

Problem 2. (c)

We need to prove that $x^Tk_5x >= 0$.
Replacing k_5 with $k_1.k_2$

$$k_5 = (k_1diag(x)k_2diag(x))^T \tag{10}$$

$$k_5 = (k_1^{1/2}k_1^{1/2}diag(x)k_1^{1/2}k_1^{1/2}diag(x))^T \tag{11}$$

$$k_5 = (k_1^{1/2}diag(x)k_2^{1/2})^T(k_1^{1/2}diag(x)k_2^{1/2}) \tag{12}$$

Now the equation is of the form A^TA. Which is always greater than 0
Hence k_5 is a valid kernel.

Problem 3. (a)

Say Objective function is

$$J = \sum_n (y_i - w^T x_i)^2 + \lambda ||w||^2 \tag{13}$$

Writtng the equation is matrices form and Using the property that matrix $A^2 = A^T A$, we get

$$J = (Y - XW)^T (Y - XW) + \lambda W^T W \tag{14}$$

$$\frac{\partial J}{\partial W} = \frac{\partial Y^T Y}{\partial W} + \frac{\partial X^T W^T X W}{\partial W} + \frac{\partial X^T W^T Y}{\partial W} + \frac{\partial Y^T X W}{\partial W} + \frac{\partial W^T W}{\partial W} \tag{15}$$

Using the fact that $\frac{\partial Y^T Y}{\partial W} = 0$ and $X^T W^T Y$ and $Y^T X W$ are transformation of each other

$$\frac{\partial J}{\partial W} = 2W^T X^T X - 2X^T Y + \lambda 2W \tag{16}$$

To get minimum equating $\frac{\partial J}{\partial W} = 0$

$$(X^T X)W - X^T Y + \lambda W = 0 \tag{17}$$

Rewriting equation,

$$(X^T X + \lambda I_N)W = X^T Y \tag{18}$$

$$W = (X^T X + \lambda I_N)^{-1} X^T Y \tag{19}$$

Problem 3. (b)

J = $\sum_n (y_i - w^T \phi(x_i))^2 + \lambda ||w||^2$

Differentiate the equation wrt w we get

$$\frac{\partial J}{\partial w} = 2 * \sum_n (y_i - w^T \phi(x_i))\phi(x_i) + 2 * \lambda * w \tag{20}$$

Now equating $\frac{\partial J}{\partial w} = 0$

$$\sum_n y_i - \sum_n w\phi(x_i)^T \phi(x_i) + \lambda w = 0 \tag{21}$$

Writing in the form of matrices

$$\phi(x_i)^T Y = w^*(\lambda I_N + \phi(x_i)\phi(x_i)^T) \tag{22}$$

Hence w^* can be written as

$$w^* = \phi^T(\lambda I_N + \phi\phi^T)^{-1} y \tag{23}$$

Problem 3. (c)

Since $\hat{y} = w^{*T}x$, We just found w^* is the part (b). Taking it's Transformation and putting back in equation. We get

$$\hat{y} = y^T(\phi^T\phi + \lambda I_N)^{-1}\phi^T\phi \tag{24}$$

Replacing $K = \phi^T\phi$

$$\hat{y} = y^T(K + \lambda I_N)^{-1}K \tag{25}$$

Problem 3. (d)

In case when there are N data points each of dimension D. Then Linear ridge regression takes time proportional to $O(N^3 + ND^2)$. N^3 for $(X^TX)^{-1}$ and ND^2 for X^TY Where as ridge regression takes time proportional to $O(N^3)$

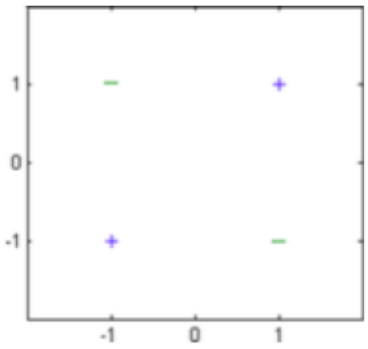
Problem 4. (a)

No, the data is not linearly separable

Problem 4. (b)

The data is not linearly separable in 2D plane but when projected to higher dimension there will be a hyperplane which will be able to separate them. That hyperplane when projected to 2D space will by a quadratic curve. Hence we need to assign weightage such that $x_1x_2 = 1$.
Hence W = [0 0 0 1]

Problem 4. (c)



Problem 4. (d)

K = $\Phi(x)\Phi(x')$
$\Phi(x) = [1 \quad x_1 \quad x_2 \quad x_1x_2]$
$\Phi(x') = [1 \quad x_1' \quad x_2' \quad x_1'x_2']$
Taking Transpose of $\Phi(x)$ and then multiplying with $\Phi(x')$
We get feature transformation, $[1 \quad x_1x_1' \quad x_2x_2' \quad x_1x_1'x_2x_2']$

Problem 5. (a)

The error function J is given as

$$J = ||w||^2 + C \sum_{i=1}^{N} e_n \tag{26}$$

such that $y_n(w^T x_n + b) >= 1 - e_n$ for n = 1, 2.....N
For large values of C, penalizing shrinking the margin heavily. This means decision boundary will separate the data perfectly.

Problem 5. (b)

C = 0, some e_n is allowed that is some misclassification is not penalised heavily while maximizing the margin between the points

Problem 5. (c)

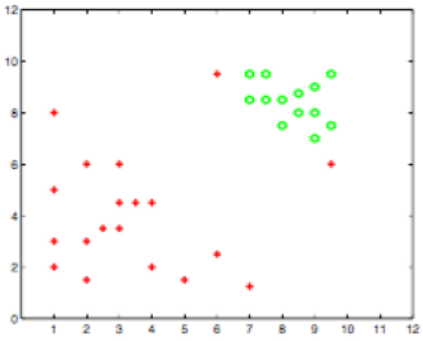
As mentioned in the problem statement itself that "avoid trusting any data point too much". Hence, we chose the case when C 0

Problem 5. (d)

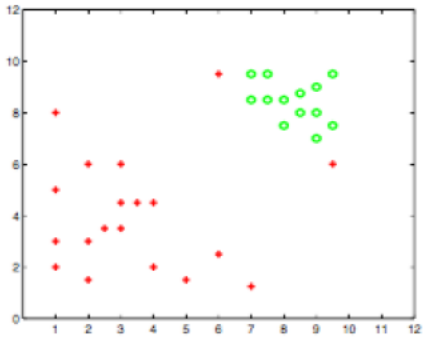
Marking the point which would have been correctly identified by the classifier and hence will not be a support vector.

Problem 5. (e)

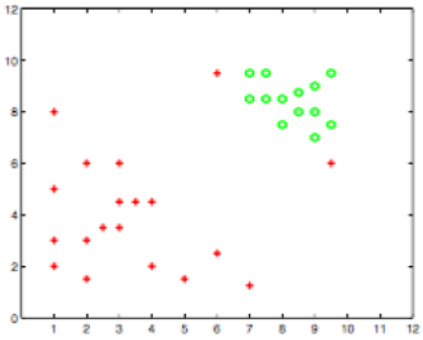
Since C is very large therefore misclassifying a single data point will involve huge penalty and hence the boundary will move significantly.



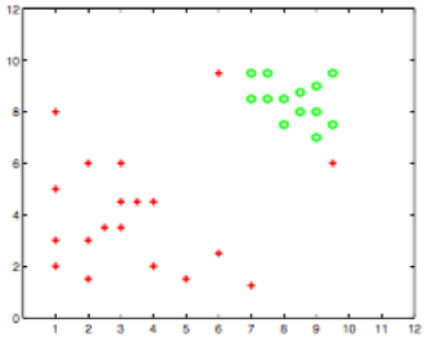
(a) Part 1



(b) Part 2



(c) Part 4



(d) Part 5

Problem 6. (a) (i)

With 100 datasets 10 samples each

Function	Bias2	Variance
g1(x)	0.465	0.000
g2(x)	0.338	0.605
g3(x)	0.284	0.612
g4(x)	0.028	0.643
g5(x)	0.038	0.644
g6(x)	0.051	0.645

Problem 6. (a) (ii)

With 100 datasets 100 samples each

Function	Bias2	Variance
g1(x)	0.469	0.000
g2(x)	0.350	0.000
g3(x)	0.346	0.007
g4(x)	0.003	0.350
g5(x)	0.004	0.351
g6(x)	0.005	0.352

Problem 6. (a) (iii)

Model Complexity impact on bias and variance
As clear from the results. Model complexity has a strong impact on both bias and variance. When complexity is low then bias is high but variance is low however as gradually increase complexity from $g_1(x)$ to $g_6(x)$ then bias decreases but variance increased. Hence, there is always a tradeoff between best value of bias and variance.

Sample Size impact on bias and variance The results show that there is not much difference in the values of bias squared and variance with size of data set. Hence, we can say the bias square or bias itself and variance are independent of sample size/data points

Problem 6. (a) (iv)

lambda	bias2	variance
0.001	0.003	0.345
0.003	0.003	0.345
0.010	0.005	0.347
0.030	0.018	0.360
0.100	0.164	0.506
0.300	1.437	1.779
1.000	15.880	16.223

λ is the regularizer used to prevent overfitting. Hence as lambda increases we see that both bias square and variance increased. However it can be seen that bias increases polynomially whereas variance increases exponentially.

Problem 6. (b) (i)

Linear SVM

C	Time	Accuracy
0.000244	352.1999667	55.75%
0.000977	346.7296667	87.80%
0.003906	229.772	91.30%
0.015625	160.7559667	92.40%
0.0625	120.0366667	93.70%
0.25	103.9117333	94.60%
1	101.8866667	94.20%
4	112.0063333	94.70%
16	167.8740333	94.50%

Linear SVM Average Time taken = 188.353 ms

Code files Description

biasVariance.py - Contains the code for Programming question part a
CommomUtilities.py - Contains code for common functions to be used in SVM programming
CSCI567_hw3_fall16.py - Main file that invokes the code of bias variance and then invokes code for SVMs
libsvm.py - runs SVM on selected parameters with best accuracy
phishing-test.mat - Contains test data set
phishing-train.mat - Contains training data
results.txt - Generated at run time to contain results obtained from SVM
svm.py and svmutil.py - LibSVM's files
testData - Generated on run time, contains data as understood by LibSVM
trainData - Generated on run time, contains data as understood by LibSVM

Problem 6. (b) (ii)

Polynomical SVM

C	Degree	Time	Accuracy
0.015625	1	337.3206667	55.75%
0.015625	2	338.8726667	55.75%
0.015625	3	346.853	55.75%
0.0625	1	302.6866667	88.45%
0.0625	2	317.4586667	88.00%
0.0625	3	334.0913333	76.05%
0.25	1	199.799	91.70%
0.25	2	225.6446667	91.40%
0.25	3	250.0266667	91.35%
1	1	134.2533333	92.35%
1	2	147.2326667	92.75%
1	3	185.3526667	92.55%
4	1	114.1023333	94.10%
4	2	106.9706667	94.45%
4	3	139.6053333	94.95%
16	1	99.827	94.20%
16	2	97.40133333	95.40%
16	3	101.4776667	95.15%
64	1	98.58033333	94.55%
64	2	93.42333333	96.45%
64	3	98.94533333	96.50%
256	1	122.0243333	94.75%
256	2	95.017	96.00%
256	3	98.656	95.60%
1024	1	237.3876667	94.30%
1024	2	103.698	96.10%
1024	3	100.042	96.15%
4096	1	1017.900667	94.60%
4096	2	116.3076667	96.10%
4096	3	110.8546667	96.25%
16384	1	2684.982	94.70%
16384	2	132.0683333	94.75%
16384	3	106.9346667	96.15%

Polynomial SVM Average Time taken = 272.599 ms

Problem 6. (b) (iii)

RBF SVM

C	Gamma	Time	Accuracy
0.015625	6.10E-05	353.9333333	55.75%
0.015625	0.000244141	350.14	55.75%
0.015625	0.000976563	349.1723333	55.75%
0.015625	0.00390625	359.521	55.75%
0.015625	0.015625	354.523	55.75%
0.015625	0.0625	351.0716667	85.05%
0.015625	0.25	356.7546667	64.85%
0.0625	6.10E-05	347.7	55.75%
0.0625	0.000244141	346.1353333	55.75%
0.0625	0.000976563	350.3846667	55.75%
0.0625	0.00390625	355.4193333	56.35%
0.0625	0.015625	309.1126667	88.25%
0.0625	0.0625	244.894	91.75%
0.0625	0.25	301.4093333	88.35%
0.25	6.10E-05	358.1973333	55.75%
0.25	0.000244141	345.7533333	55.75%
0.25	0.000976563	376.551	57.80%
0.25	0.00390625	285.977	88.75%
0.25	0.015625	203.5116667	91.60%
0.25	0.0625	160.42	93.25%
0.25	0.25	206.7416667	95.85%
1	6.10E-05	350.0886667	55.75%
1	0.000244141	354.7456667	58.40%
1	0.000976563	285.695	88.95%
1	0.00390625	189.9943333	91.70%
1	0.015625	143.1996667	93.00%
1	0.0625	121.863	95.40%
1	0.25	150.9446667	96.60%
4	6.10E-05	367.661	58.05%
4	0.000244141	285.4716667	88.90%
4	0.000976563	181.0486667	91.60%
4	0.00390625	137.7243333	92.80%
4	0.015625	105.794	95.10%
4	0.0625	103.09	96.15%
4	0.25	138.0606667	97.05%
16	6.10E-05	281.7436667	88.80%
16	0.000244141	187.11	92.05%
16	0.000976563	132.6806667	93.15%
16	0.00390625	110.128	94.50%
16	0.015625	103.387	94.70%
16	0.0625	94.78366667	97.30%
16	0.25	141.289	96.90%
64	6.10E-05	193.9283333	91.65%
64	0.000244141	131.805	92.90%
64	0.000976563	111.0803333	94.40%
64	0.00390625	98.28666667	94.60%
64	0.015625	99.662	96.05%

C	Degree	Time	Accuracy
64	0.0625	83.72833333	96.15%
64	0.25	135.6706667	96.30%
256	6.10E-05	135.0433333	92.90%
256	0.000244141	114.133	94.70%
256	0.000976563	100.462	94.75%
256	0.00390625	102.5073333	95.00%
256	0.015625	99.29766667	96.15%
256	0.0625	92.76866667	95.80%
256	0.25	140.9293333	96.60%
1024	6.10E-05	112.7036667	94.40%
1024	0.000244141	103.7326667	93.95%
1024	0.000976563	96.63266667	95.10%
1024	0.00390625	109.628	95.95%
1024	0.015625	108.18	96.45%
1024	0.0625	90.82733333	96.40%
1024	0.25	132.834	96.30%
4096	6.10E-05	104.7796667	94.50%
4096	0.000244141	101.1653333	94.35%
4096	0.000976563	116.7513333	94.75%
4096	0.00390625	132.6766667	96.30%
4096	0.015625	111.6596667	96.25%
4096	0.0625	95.827	96.55%
4096	0.25	140.1436667	96.50%
16384	6.10E-05	106.454	94.40%
16384	0.000244141	121.4933333	95.05%
16384	0.000976563	170.3496667	96.35%
16384	0.00390625	165.237	96.45%
16384	0.015625	130.8033333	96.80%
16384	0.0625	93.19033333	96.20%
16384	0.25	133.217	96.40%

RBF SVM Average Time taken = 191.576 ms

Problem 6. (b) (iv)

Selected RBF Kernel with C = 16384 that is 4^7 and gamma = 0.015625 that is 4^{-3}
Accuracy = 95.55%
Number of correct prediction/Size of test data = (1911/2000) 55