Handwritten hardcopy approved by Dr. Yan Liu

Anjali Krishna Prasad
MORNING SESSION
USCid: 3400313602

Assignment 3

Q1.a)

$$\hat{B}_\lambda = \arg\min_B \left\{ \frac{1}{n} \sum_{n=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \|B\|_2^2 \right\}$$

$$= \arg\min_B \left\{ \frac{1}{n} \sum_{n=1}^{n} (x_i^T \beta - y_i)^2 + \lambda \|B\|_2^2 \right\}$$

Converting the above equation to vectorized form.

$$\hat{B}_\lambda = (BX - y)^T (BX - y) + \lambda B^T B$$

Finding $\nabla \hat{B}_\lambda = 0$

$$\nabla_B = X^T (BX - y) + \lambda B = 0$$

$$X^T BX - X^T y + \lambda B = 0$$

$$B(X^T X + \lambda I) = X^T y.$$

$$B = (X^T X + \lambda I)^{-1} X^T y.$$

↑ closed form solution.

$$B = [X^T X + \lambda I]^{-1} X^T y.$$

we know $y = X^T \beta^* + \epsilon$

$$\hat{\beta}_\lambda = [X^T X + \lambda I]^{-1} X^T (X^T \beta^* + \epsilon)$$

$$= [X^T X + \lambda I]^{-1} [X X^T \beta^* + X^T \epsilon]$$

P B A

where $P = [X^T X + \lambda I]^{-1}$

$A = X^T$

$B = X X^T \beta^*$

$= P(B + A\epsilon)$

Using Affine transformation :

where $X \sim N(0,1)$

$Ax + B \sim N(A\mu + B, A^T \sigma A)$

where ϵ is taken from $N(0,1)$

we get

$$N(X X^T \beta^*, X X^T)$$

As $\mu = 0$ & $\sigma = 1$.

$$\Rightarrow P\left[N(X X^T \beta^*, X X^T)\right]$$

substituting P for its value.

$$[X^T X + \lambda I]^{-1} \left[N(X X^T \beta^*, X X^T)\right]$$

we again use affine transformation to get.

$$\beta \sim N\left[(X^T X + \lambda I)^{-1} X X^T \beta^*, \right.$$
$$\left. (X^T X + \lambda I)^{-T} X X^T (X^T X + \lambda I)^{-1}\right]$$

b) In the previous solution we got

$$B \sim N\left[(X^TX+\lambda I)^{-1}XX^TB^*,\ (X^TX+\lambda I)^{-T}XX^T(X^TX+\lambda I)^{-1}\right]$$

here mean is equivalent to ~~var~~ bias

∴ bias term is

$$(X^TX+\lambda I)^{-1}XX^TB^*$$

c) In a part of this answer we got

$$B \sim N\left[(X^TX+\lambda I)^{-1}XX^TB^*,\ (X^TX+\lambda I)^{-T}XX^T(X^TX+\lambda I)^{-1}\right]$$

Here 2nd term is variance.

∴ variance term = $(X^TX+\lambda I)^{-T}XX^T(X^TX+\lambda I)^{-1}$

d) Without regularization, we can get variance very high. When we use regularization & keep λ small, there is a slight increase in bias but the variance comes down significantly. If we keep increasing λ, bias increases significantly.

We can see in variance equation

$$(X^TX+\lambda I)^{-T}XX^T(X^TX+\lambda I)^{-1}$$

As λ increase variance decreases

$\therefore$ $\lambda \uparrow$ variance $\downarrow$

$\lambda \downarrow$ bias $\uparrow$

Q2a) To prove $k_3(x,x') = a_1 k_1(x,x') + a_2 k_2(x,x')$

where $a_1, a_2 \geq 0$.

Let $\langle x,x \rangle$ be euclidean inner product.

$k_1(x,x')$ is a kernel therefore +ve semi-definite i.e.

$\langle v, k_1(x,x')v \rangle > 0$

a_1 is a scalar

$\langle v, a_1 k_1(x,x')v \rangle = a_1 \langle v, k_1(x,x')v \rangle$

This is +ve semi definite.

$\langle v, a_2 k_2(x,x')v \rangle = a_2 \langle v, k_2(x,x')v \rangle$

This is also +ve semidefin

Combining the two.

$\langle v, \{ a_1 k_1(x,x') + a_2 k_2(x,x') \} v \}$.

By property of inner product.

$a_1 \langle v, k_1(x,x')v \rangle + a_2 \langle v, k_2(x,x')v \rangle$

> 0.

For all non zero vectors x

since $a_1 \langle v, k_1(x,x')v \rangle > 0$ and

$a_2 \langle v, k_2(x,x')v \rangle > 0$

Thus by definition of matrix.

$a_1 k_1(x,x') + a^2 k_2(x,x')$ is positive semidefinite.

Proving symmetric nature.

$a_1 k_1(x,x')$ and $a_2 k_2(x,x')$ are symmetric as

$$aA = a[A] = a[A^T]$$

We know $A = A^T$ when symmetric & k_1 & k_2 are kernels so they are symmetric.

To prove: $a_1 k_1(x,x') + a_2 k_2(x,x')$ is symmetric.

Let $a_1 k_1(x,x') = A$

$a_2 k_2(x,x') = B$.

$$(A+B)^T = A^T + B^T$$

But $A^T = A$ & $B^T = B$.

$\therefore (A+B)^T = A+B$.

Hence $a_1 k_1(x,x') + a_2 k_2(x,x')$ is symmetric

$\therefore k_3(x,x')$ is also a kernel.

2 b) $k_4(x,x') = f(x) f(x')$ To prove

$K_4(x, x') = \langle \phi(x), \phi(x') \rangle$

where $\phi : x \to f(x)$

$\phi : x' \to f(x')$.

$K_4(x, x') = \langle f(x), f(x') \rangle$

$\therefore f(x) f(x')$ is symmetric and positive definite.

2c) To prove $K_5(x, x') = K_1(x, x') K_2(x, x')$

$\phi(x)_{ij} = \phi_1(x)_i \cdot \phi_2(x)_j$.

i.e the i,j component of the new feature vector is the product of i^{th} of $\phi_1(x)$ & j^{th} of $\phi_2(x)$.

$$K_5(x, x') = \langle \phi(x), \phi(x') \rangle$$

$$= \sum_{ij} \phi(x)_{ij} \phi(x')_{ij}$$

$$= \sum_{ij} \phi_1(x)_i \phi_2(x)_j \cdot \phi_1(x')_i \phi_2(x')_j$$

$$= \left(\sum_i \phi_1(x_i) \phi_2(x'_i) \right) \left(\sum_j \phi_1(x_j) \phi_2(x'_j) \right)$$

$$= K_1(x, x') K_2(x, x')$$

Hence proved.

3a) $\min_{w} \sum_{n} (y_i - w^T x_i)^2 + \lambda \|w\|_2^2$.

Converting the above equation in vector form, we get.

$$(Xw - y)^T (Xw - y) + \lambda w^T w.$$

Taking gradient wrt w & $\nabla_w = 0$

$$\nabla_w = X^T(Xw - y) + \lambda w = 0.$$

$$X^T X w - X^T y + \lambda w = 0$$

$$w(X^T X + \lambda I_D) = X^T y.$$

$$w = [X^T X + \lambda I_D]^{-1} X^T y.$$

Therefore we can say optimal w can be written as.

$$[X^T X + \lambda I_D]^{-1} X^T y.$$

3b) Applying non linear feature mapping to each sample $x_i \rightarrow \phi_i = \phi_i(x) \in R^T$

s.t $\phi \in R^{N \times T}$

then

$$\sum_n [w^T \phi(x_n) - y_n]^2 + \lambda \|w\|_2^2$$

changing this to vector form

$$\phi^T [\phi w - y] [\phi w - y]^T + \lambda w^T w$$

taking gradient wrt w $\nabla_w = 0$

$$\nabla_w = \phi^T [\phi w - y] + \lambda w = 0$$

$$\phi^T [\phi w] - \phi^T y + \lambda w = 0$$

$$w^* [\phi^T \phi + \lambda I_t] = \phi^T y$$

$$w^* = [\phi^T \phi + \lambda I_t]^{-1} \phi^T y \quad -(1)$$

we have to prove

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

Lets use the following property.

$$(P^{-1} + B^T R^{-1} B)^{-1} B^T R^{-1} = P B^T (B P B^T + R)^{-1}$$

Let $P = I$

$B = \phi$

$R = \lambda I$

Lets keep aside y for sometime

$$(P^{-1} + B^T R^{-1} B)^{-1} B^T R^{-1}$$

$$= (I + \phi^T \frac{1}{\lambda} I \phi)^{-1} \phi^T \frac{1}{\lambda} I$$

as $A^{-1} B^{-1} = (BA)^{-1}$

$$\therefore \left(\phi^T \frac{1}{\lambda} I\right)^{-1} = \lambda I \phi^{-T}$$

$$\Rightarrow \left[\lambda I \phi^{-T} I + \lambda I \phi^{-T} \phi^T \frac{1}{\lambda} I \phi\right]^{-1}$$

$$= \left[\phi^{-T} \lambda I + \phi^{-T} \phi^T \phi\right]^{-1}$$

$$= \left[\lambda I + \phi^T \phi\right]^{-1} \phi^T$$

multiplying the above eqn by y we get ①

$$\left[\lambda I + \phi^T \phi\right]^{-1} \phi^T y \Rightarrow \left[\phi^T \phi + \lambda I_t\right] \phi^T y$$

This means we can use the matrix identity to get the result.
[using P, B & R values]

$$PB^T(BPB^T+R)^{-1}$$

$$= I\phi^T(\phi I\phi^T+\lambda I)^{-1}$$

$$= \phi^T(\phi\phi^T+\lambda I)^{-1}$$

Multiplying the above equation by y we get:

$$= \phi^T(\phi\phi^T+\lambda I)^{-1}y$$

Hence proved.

3c) We know

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

To prove:

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

given:

$$\hat{y} = w^{*T}\phi(x)$$

$$= y^T(\phi\phi^T+\lambda I_N)^{-1}\phi(x)\,\phi^T$$

$$= y^T(K+\lambda I_N)^{-1}k(x)$$

3d) Linear Ridge regression has a complexity of $O(D^3)$

Kernel Ridge regression has the complexity of $O(N^3)$

Linear ridge regression: $(X^T X + \lambda I_D)^{-1} X^T y$.

$X^T X$ is a $D \times N$ and $N \times D$ matrix

$(X^T X + \lambda I_D)^{-1}$ is a $D \times D$ inversion.

$\therefore O(D^3)$ is complexity.

Kernel Ridge regression:

$$\phi^T (\phi \phi^T + \lambda I_N)^{-1} y$$

$\phi \phi^T$ is a $N \times T$ and $T \times N$ matrix

$(\phi \phi^T + \lambda I_N)^{-1}$ is a $N \times N$ matrix inversion

$\therefore O(N^3)$ is complexity of kernel ridge regression.

Linear ridge regression complexity depends on the features & kernel ridge regression depends on no of examples x.

Q4 a) Positive examples are not linearly separable from the negative examples in the original space as there is no line present that can divide the positive and negative examples from each other.

b). We have 2 positive points (1,1), (-1,-1) and we have 2 negative points (-1,1), (1,-1)

Our transformation function is

$\phi(x) = [1, x_1, x_2, x_1, x_2]$. Applying this on all 4 points we get.

Positive points: (1,1) transforms to

[1, 1, 1, 1],

(-1,-1) → [1, -1, -1, 1]

negative points:

(1,-1) → [1, 1, -1, -1]

(-1,1) → [1, -1, 1, -1]

We can clearly see that only terms that decides that if the example is positive or negative is the term $x_1 x_2$

Therefore we can set w coefficients

[0, 0, 0, 1)

4c) If we have a positive example (2, -2), the total five examples can no longer be linearly separated in the feature space $\phi(x)$. As multiplying the transformation we get

$$[1, 2, -2, -4].$$

with the coefficients we got in the previous answer for w.

We get -4. which classifies the example to be negative. But it is actually positive.

4d) $K(x, x) = \phi(x) \cdot \phi(x)^T$

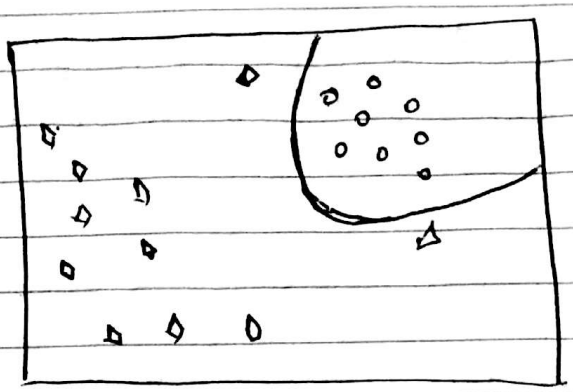
$$\phi(x) = [1, x_1, x_2, x_1 x_2]$$

$$\phi(x)^T = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_1 x_2 \end{bmatrix}$$

$$K(x, x) = 1 + x_1^2 + x_2^2 + x_1^2 x_2^2$$

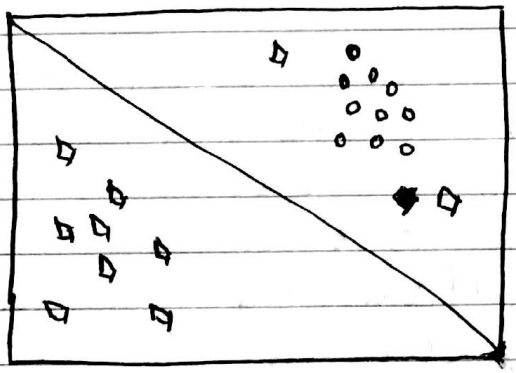
$K(x, x)$ is a polynomial kernel.

Q5 a)



When $C \to \infty$, we get hard margin in SVM which results in overfitting. Therefore I drew a parabolla to separate the 2 classes without misclassifying any points.

b)

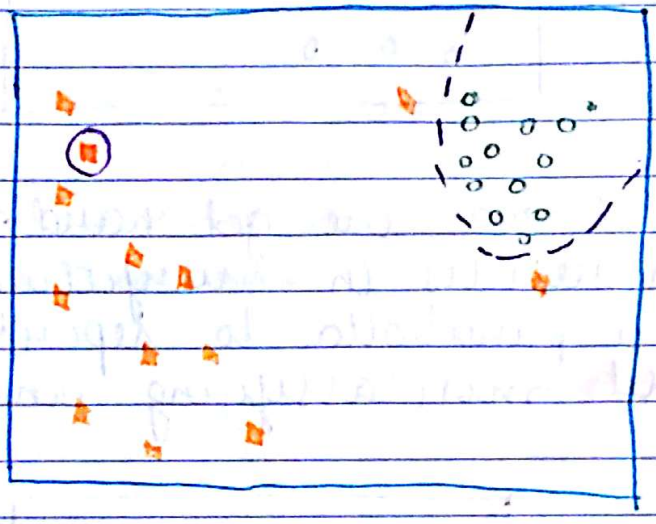


For ~~C~~ $C \gtrsim 0$, we get soft margin that means misclassification is allowed in this case. For a very tiny value of C, we should get misclassified examples.

c) I expect the second one ($C \gtrsim 0$) to work better the the 1st one ($C \to \infty$) as it is more generic. We can clearly see overfitting in the first ~~case~~ case.

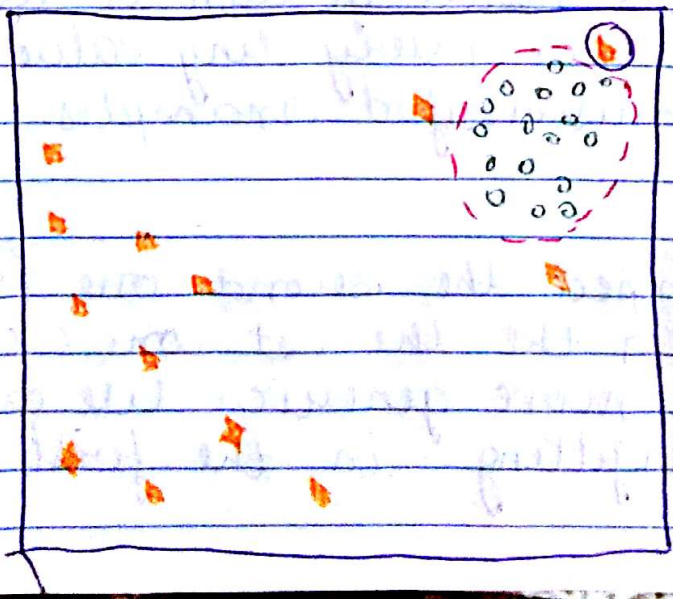
In the second case, only 2 points are misclassified but the model looks more generic.

d)



The point which I have circled, is the new point I have added which will not change the decision boundary when $C \to \infty$. As when $C \to \infty$, we get hard margin. Hard margin restricts misclassification and according the the hard margin in this case, the new point will not be misclassified.

e)



The circled point is the new point I have added which changes the decision boundary for very large $C \to \infty$. When $C \to \infty$, it is a hard margin that does not allow any misclassifications. Therefore according to the old decision boundary, the red point I added will be misclassified. Therefore the decision boundary will change as shown in the figure.

Programming question:

Extension by 1 week for programming question approved by Dr. Yanjun on medical grounds.