CSCI 567 Assignment 3
Fall 2016

Snehal Adsule
2080872073
adsule@usc.edu

October 17,2016

1 Problem 1

1.1 1 (a)

Closed Form Given that $\hat{\beta}_\lambda = argmin_\beta \{ \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \mid\mid \beta \mid\mid_2^2 \}$
Differentiating wrt β

$$\frac{\delta\hat{\beta}_\lambda}{\delta\beta} = \frac{2}{n}\{\sum_{i=1}^{n}(y_i - x_i^T\beta)(-x_i^T) + \lambda\beta\} = 0$$
$$=> \frac{2}{n}\{-X^TY + X^T\beta X + \lambda\beta\} = 0$$
$$=> \beta(X^TX + \lambda) = YX^T$$
$$=> \hat{\beta} = (X^TX + \lambda)^{-1}X^T.Y$$

using $Y = X\beta^* + \epsilon$
$\hat{\beta} = (X^TX + \lambda)^{-1}X^T.(X\beta^* + \epsilon)$
The guassian distribution for the noise is, $\epsilon N(0, \sigma^2)$.
Using affine transformation distribution of y can be written as

$$\text{Thus, } \hat{\beta} = (X^TX + \lambda)^{-1}X^T.(X\beta^* + \epsilon)$$
$$\text{And, } YN(X\beta^*, \sigma^2 I)$$
$$=> \hat{\beta}_\lambda = ((X^TX + \lambda)^{-1}X^TX\beta^*, (X^TX + \lambda)^{-1}X^TX(X^TX + \lambda I)^{-1})$$

1.2 1 (b)

Bias Term

$$\begin{aligned} E[x^T\hat{\beta}_\lambda] - x^T\beta^* \\ = x^T(E[\hat{\beta}_\lambda] - \beta^*) = x^T((X^TX+)^{-1}X^TX\beta^*\beta^*) \\ = x^T((X^TX+\lambda)^{-1}X^TX - I)\beta^* \end{aligned}$$

next

1.3 1 (c)

Variance Term

$$\begin{aligned} E[(x^T(\beta_\lambda E[\beta_\lambda]))^2] = x^T(X^TX+\lambda)^{-1}X^TX(XX^T+\lambda I)^{-1}x \\ = ||X(XX^T+\lambda)^{-1}x||_2^2 \end{aligned}$$

1.4 1 (d)

We can observe that , with Part b. and Part c. of the bias and variance tradeoff if λ increases, the bias term also increases while the variance term decreases. And when λ is small, the bias term si expected to be smaller and the variance term will be larger, comparatively.

2 Kernel Construction

2.1 2. (a)

To prove that , $k_3(x,x') = a_1k_1(x,x') + a_2k_2(x,x')$ where$a_1, a_2 \geq 0$

Since $k_1(x,x')$ is positive definite, $\forall y \in \mathbf{R}$,

$$\begin{aligned} y^TK^{(1)}y \geq 0 \\ \text{where } K_{ij}^{(1)} = k_1(x_i, x_j') \end{aligned}$$

Similarly,

$$\begin{aligned} y^TK^{(2)}y \geq 0 \\ \text{where } K_{ij}^{(2)} = k_2(x_i, x_j') \end{aligned}$$

Adding ,the above two equations, we get

$$\begin{aligned} y^T(K^{(1)} + K^{(2)})y \geq 0 \ \forall y \in \mathbf{R} \implies \\ y^TK^{(3)}y \geq 0 \ \forall y \in \mathbf{R} \\ \text{where} K_{ij}^{(3)} = k_3(x_i, x_j') \end{aligned}$$

2.2 2. (b)

To prove , $k_4(x,x') = f(x)f(x')$ $K^{(4)}_{ij} = k_4(x_i,x_j) = f(x_i)f(x'_j)$
Since $f(x)$ is a real valued function, consider $K^{(4)}$

$$K^{(4)} = \begin{bmatrix} f(x_1)f(x'_1) & f(x_1)f(x'_2) & \cdots & f(x_1)f(x'_n) \\ \vdots & & & \\ f(x_n)f(x'_1) & f(x_n)f(x'_2) & \cdots & f(x_n)f(x'_n) \end{bmatrix}$$

$$K^{(4)} = F(\vec{x})_{n\times 1}F(\vec{x})^T_{1\times n}$$

where

$$F(x)^T_{1\times n} = \begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots f(x_n) \end{pmatrix}$$

Therefore, $y^T K^{(4)} y = y^T F(x)F(x)^T y = y^T F(x)(y^T F(x))^T = ||y^T F(x)||_2^2 \geq 0$
We can say , $k_2(.,.)$ is a valid kernel function!.

2.3 2. (c)

To prove that $k_5(x,x') = k_1(x,x')k_2(x,x')$ $K^{(5)} = K^{(1)} \circ K^{(2)}$ where $\circ$ denotes the Hadamard product. Using the Schur product for$K^{(1)}, K^{(2)}$ we can prove this.

Since, k_1 and k_2 are valid kernel function $\exists v_i w_j$ the eigen vectors of matrix K_1 and K_2 defines such that:
$K^{(1)} = \sum_i \lambda_i v_i v_i^T$ and $K^{(2)} = \sum_j \mu_j w_j {w_j}^T$
Now,

$$\begin{aligned} K^{(5)} &= K^{(1)} \circ K^{(2)} \\ &= \sum_i \lambda_i v_i v_i^T \circ \sum_j \mu_j w_j {w_j}^T \\ &= \sum_{i,j} \lambda_i \mu_j (v_i v_i^T) \circ w_j {w_j}^T \\ &= \sum_{i,j} \lambda_i \mu_j (v_i \circ w_j)(v_j \circ w_j)^T \\ &\geq 0 \end{aligned}$$

As, $(v_i \circ w_j)(v_j \circ w_j)^T = ||v_i w_j||_2^2 \geq 0$

3 Kernel Regression

3.1 3.a

Given that ,$min_w(\sum_i (y_i - w^T x_i)^2 + \lambda ||w||_2^2)$
We can think of it as vector and rewrite is as ,

$min_w(||y - w^T X||_2^2 + \lambda ||w||_2^2)$

$$\begin{aligned}
f(w) &= min_w(||y - Xw||_2^2 + \lambda ||w||_2^2) \\
&= (y - Xw)^T(y - Xw) + \lambda w^T w \\
&= (y^T - w^T X^T)(y - Xw) + \lambda w^T w \\
&= y^T y - y^T X w - w^T X^T y + w^T X^T X w + \lambda w^T w \\
&= y^T y - (X^T y)^T w - w^T X^T y + w^T X^T X w + \lambda w^T w \\
\frac{\partial f(w)}{\partial w} &= -X^T y - X^T y + 2\lambda w + (X^T X w + (X X^T w)) = 0 \\
&= 2\lambda w + 2X^T X w - 2X^T y = 0 \\
\text{w}(\lambda I_D + X^T w) &= X^T y
\end{aligned}$$

=> w* = $(X^T X w + \lambda I_D)^{-1} X^T y$, where I_D denotes DxD identity matrix

3.2 3.b

After applying the non linear feature mapping , the solution should be similar $min_w(||y - w^T \Phi||_2^2 + \lambda ||w||_2^2)$

=> w = $(\Phi^T \Phi + \lambda I_D)^{-1} \Phi^T y$
Using the identity:

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

and assuming matrix inversion is valid

$$\big((\lambda I_D + \Phi^T \Phi)^{-1}\big)\Phi^T y = \Phi^T \big(\Phi \Phi^T + \lambda I_N\big)^{-1} y$$

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

3.3 3.c

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

can be written as

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

$$
\begin{aligned}
\hat{y} &= y^T\left((\Phi\Phi^T + \lambda I_N)^{-1}\right)^T \Phi^T\Phi(x) \\
&= y^T\left((\Phi\Phi^T + \lambda I_N)^T\right)^{-1} \Phi^T\Phi(x), Using, (A^{-1})^T = (A^T)^{-1} \\
&= y^T\left((\Phi^T\Phi + \lambda I_N)\right)^{-1} \Phi^T\Phi(x) \\
&= y^T(K + \lambda I_N)^{-1}\kappa(x)
\end{aligned}
$$

Where $K_{ij} = \Phi_i^T\Phi_j$ and $\kappa(x) = \phi^T\phi^T(x)$ (given)

3.4 3.d

We can say that kernel ridge regression is $O(n^3)$ for n data points, considering the multiplication and inversion of matrices. However, linear regression can be presented as quadratic programing and hence is $O(n^2)$. Kernel $N \times N$ compared to $D \times D$(for ridge regression without kernel) as in Part (b). In cases where $d < n$ this leads to an extra operations for computing K .