CS 698: Assignment 3

Ronghao Yang
ID:20511820
Session: 4:00pm-5:20pm

November 2, 2017

1 Exercise 1

1.1 Question 1.1

$$\text{let's define } p(x_i) = \frac{1}{1+e^{-y_i w^T x_i}}$$

$$\text{let } g_k = \frac{\partial}{\partial w_k} f(w) = \frac{1}{n^+} \sum_{i:y_i=1} \frac{-e^{-y_i w^T x_i} y_i x_i^k}{1+e^{-y_i w^T x_i}} + \frac{1}{n^-} \sum_{j:y_j=-1} \frac{-e^{-y_j w^T x_j} y_j x_j^k}{1+e^{-y_j w^T x_j}} + 2\lambda w_k$$

$$\text{Therefore, } g(k) = \frac{1}{n^+} \sum_{i:y_i=1} -p(x_i)(\frac{1}{p(x_i)} - 1) y_i x_i^k + \frac{1}{n^-} \sum_{i:y_j=-1} -p(x_j)(\frac{1}{p(x_j)} - 1) y_j x_j^k + 2\lambda w_k$$

$$\nabla f(w) = \begin{bmatrix} g_1 \\ g_2 \\ ... \\ g_d \end{bmatrix}$$

$$\text{let } s_{ks} = \frac{\partial}{\partial w_s} g_k$$

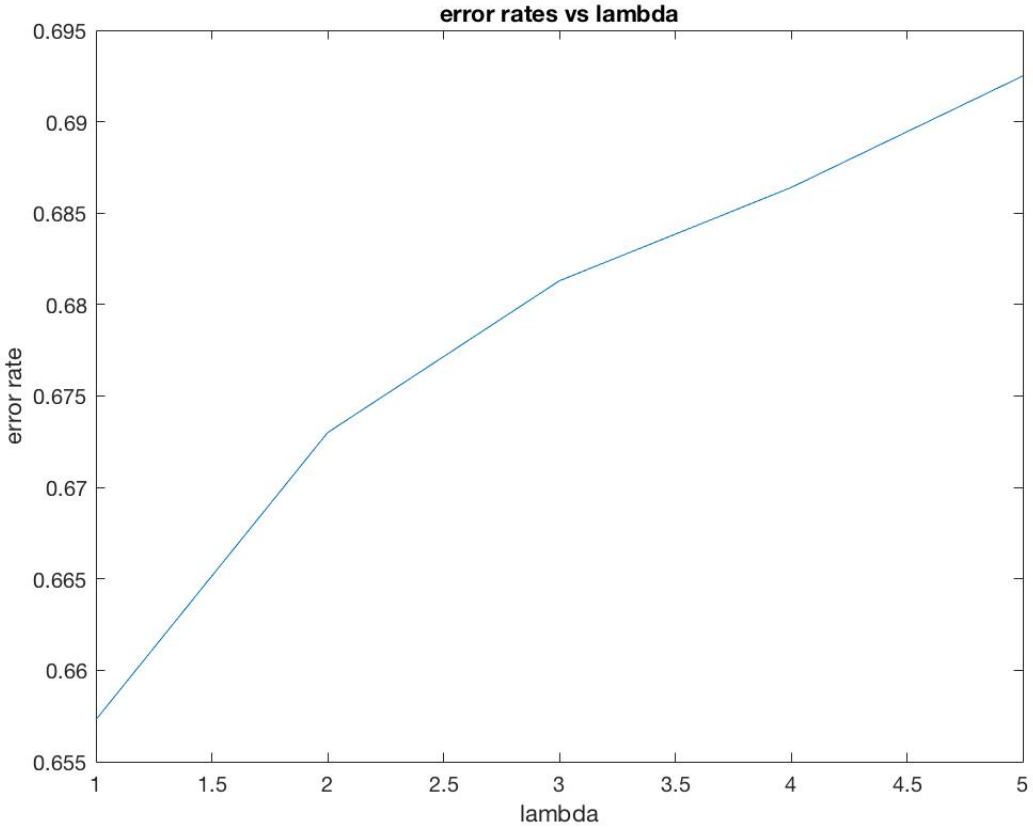
$$\text{Then } s_{ks} = \frac{1}{n^+} \sum_{i:y_i=1} \frac{e^{-y_i w^T x_i} y_i^2 x_i^k x_i^s}{(1+e^{-y_i w^T x_i})^2} + \frac{1}{n^-} \sum_{j:y_j=-1} \frac{e^{-y_j w^T x_j} y_j^2 x_j^k x_j^s}{(1+e^{-y_j w^T x_j})^2} + 2\lambda\epsilon, \text{ where } \epsilon = 1 \text{ if } s = k, \text{ else } \epsilon = 0$$

$$\text{Therefore, } s_{ks} = \frac{1}{n^+} \sum_{i:y_i=1} (p(x_i) - p^2(x_i)) x_i^k x_i^s + \frac{1}{n^-} \sum_{j:y_j=-1} (p(x_j) - p^2(x_j)) x_j^k x_j^s + 2\lambda\epsilon$$

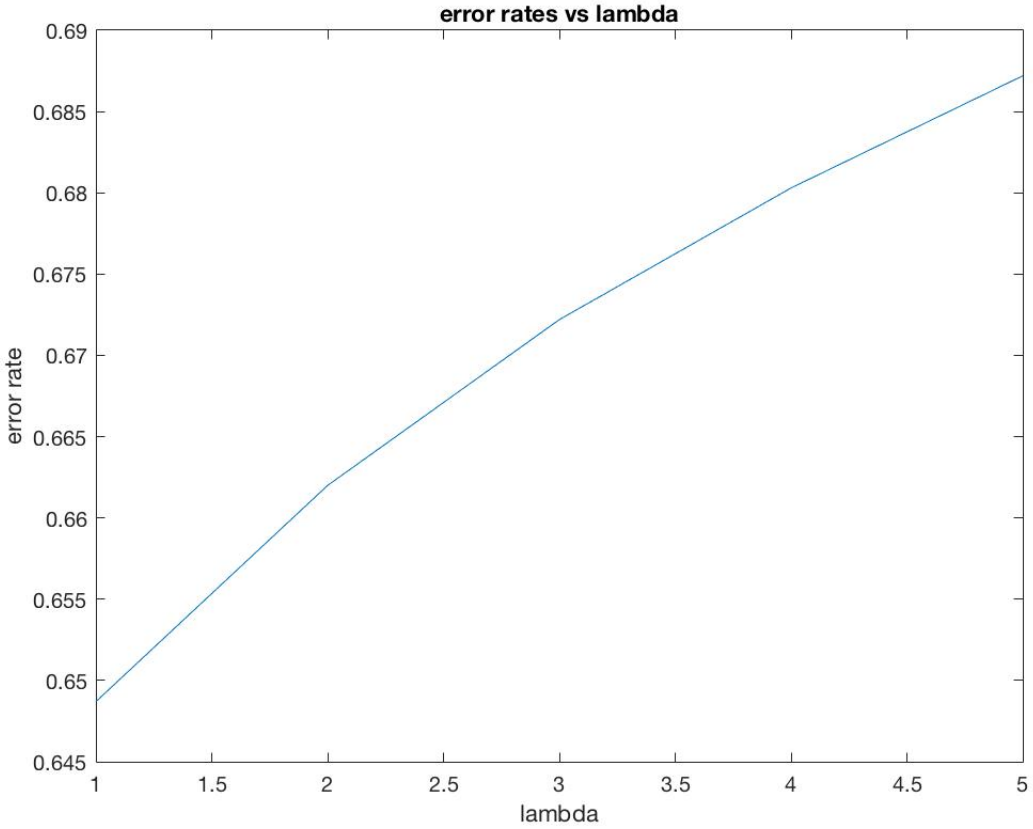
$$\text{where } \epsilon = 1 \text{ if } s = k, \text{ else } \epsilon = 0$$

$$\nabla^2 f(w) = \begin{bmatrix} s_{11} & s_{12} & ... & s_{1d} \\ s_{21} & s_{22} & ... & s_{2d} \\ ... & & & \\ s_{d1} & s_{d2} & ... & s_{dd} \end{bmatrix}$$

1.2 Question 1.2



1.3 Question 1.3



1.4 Question 1.4

For kernelized logistic regression, each x is transformed to $\phi(x)$, and the weight is defined in terms of support vectors

$$w = \sum_i \alpha_i \phi(x_i)$$

Therefore, the sigmoid function is defined as

$$p = \frac{1}{1+e^{-y_i \sum_i \alpha_i \phi(x_i)\phi(x)}} = \frac{1}{1+e^{-y_i \sum_i \alpha_i K(x,x_i)}}$$

The the loss function of kernel logistic regression becomes:

$$\min_w \frac{1}{n^+} \sum_{i:y_i=1} log(1 + exp(-y_i \sum_l \alpha_l K(x, x_i))) + \frac{1}{n^-} \sum_{j:y_j=-1} log(1 + exp(-y_j \sum_l \alpha_l K(x, x_j))) + \lambda w^t w$$

To calculate the Gradient, we do the following:

$$\nabla f(\alpha) = \frac{1}{n^+}\sum_{i:y_i=1} -p(x_i)(\frac{1}{p(x_i)}-1)y_iK(x,x_i)+\frac{1}{n^-}\sum_{i:y_j=-1} -p(x_j)(\frac{1}{p(x_j)}-1)y_jK(x,x_j)+2\lambda K\alpha$$

To calculate the Hessian, we do the following:

$$\text{let } s_{ks} = \frac{\partial}{\partial\alpha_s}g_k$$

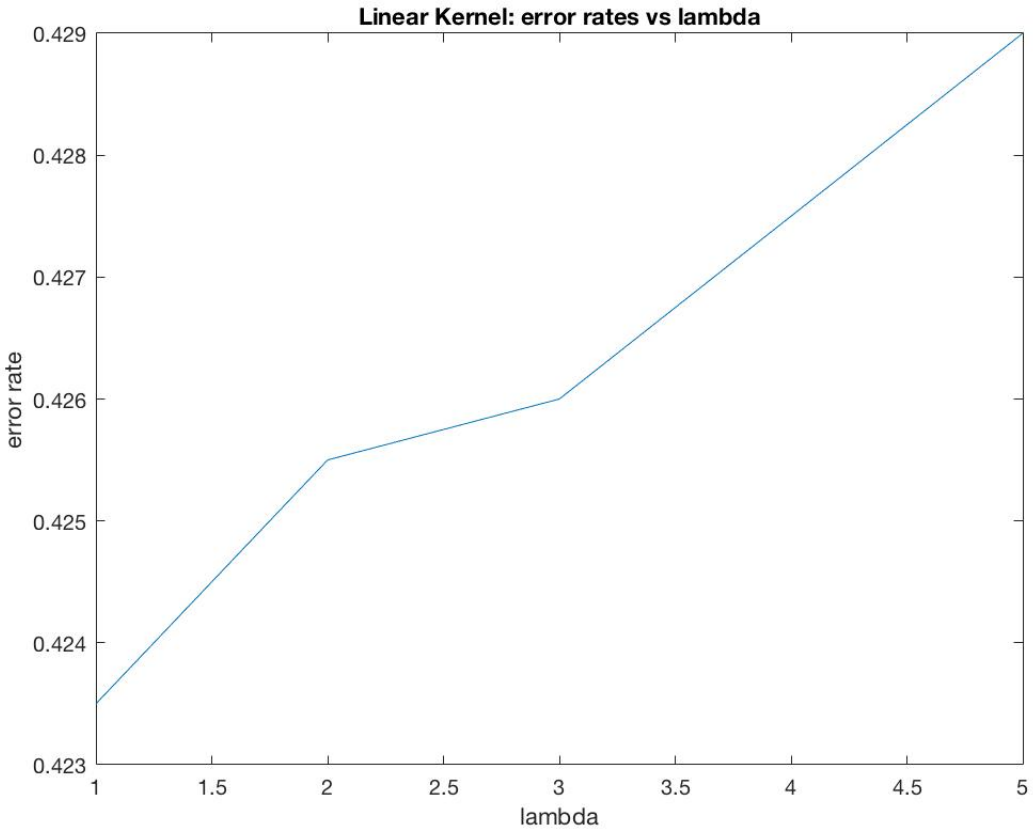
$$\text{Then } s_{ks}=\frac{1}{n^+}\sum_{i:y_i=1}\frac{e^{-y_i\sum_l \alpha_l k(x,x_i)}K(x_k,x_i)K(x_s,x_i)}{(1+e^{-y_i\sum_l \alpha_l K(x,x_i)})^2}+\frac{1}{n^-}\sum_{j:y_j=-1}\frac{e^{-y_i\sum_l \alpha_l k(x,x_j)}K(x_k,x_j)K(x_s,x_j)}{(1+e^{-y_j\sum_l \alpha_l K(x,x_j)})^2}+2\lambda K(x_k,x_s)$$

Therefore

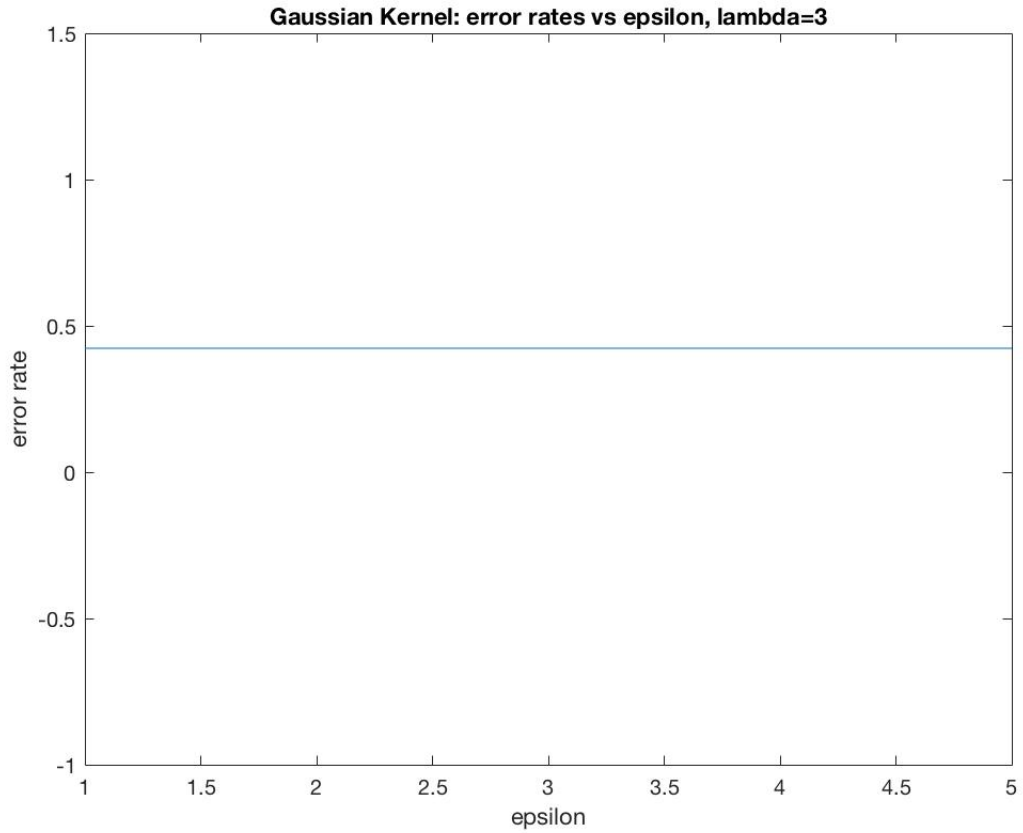
$$s_{ks} = \frac{1}{n^+}\sum_{i:y_i=1}(p(x_i)-p^2(x_i))K(x_k,x_i)K(x_s,x_i)+\frac{1}{n^-}\sum_{j:y_j=-1}(p(x_j)-p^2(x_j))K(x_k,x_j)K(x_s,x_j)+2\lambda K(x_k,x_s)$$

$$\nabla^2 f(\alpha) = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1d} \\ s_{21} & s_{22} & \dots & s_{2d} \\ \dots & & & \\ s_{d1} & s_{d2} & \dots & s_{dd} \end{bmatrix}$$

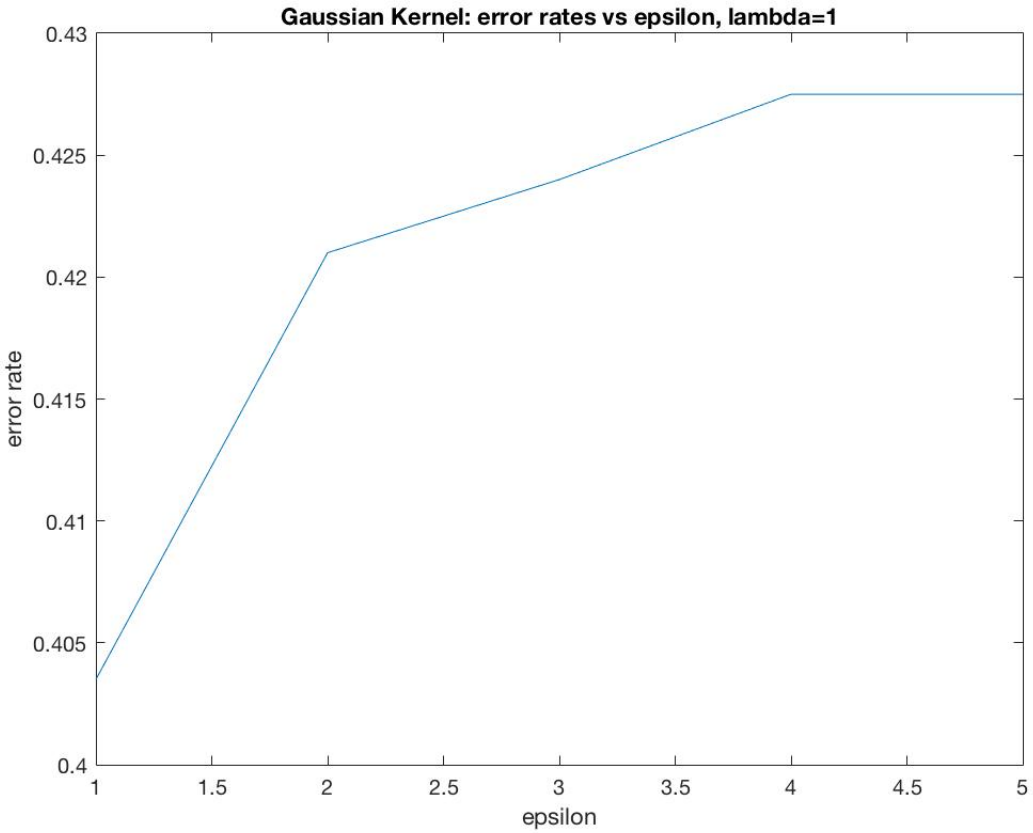
For linear kernel, we have:



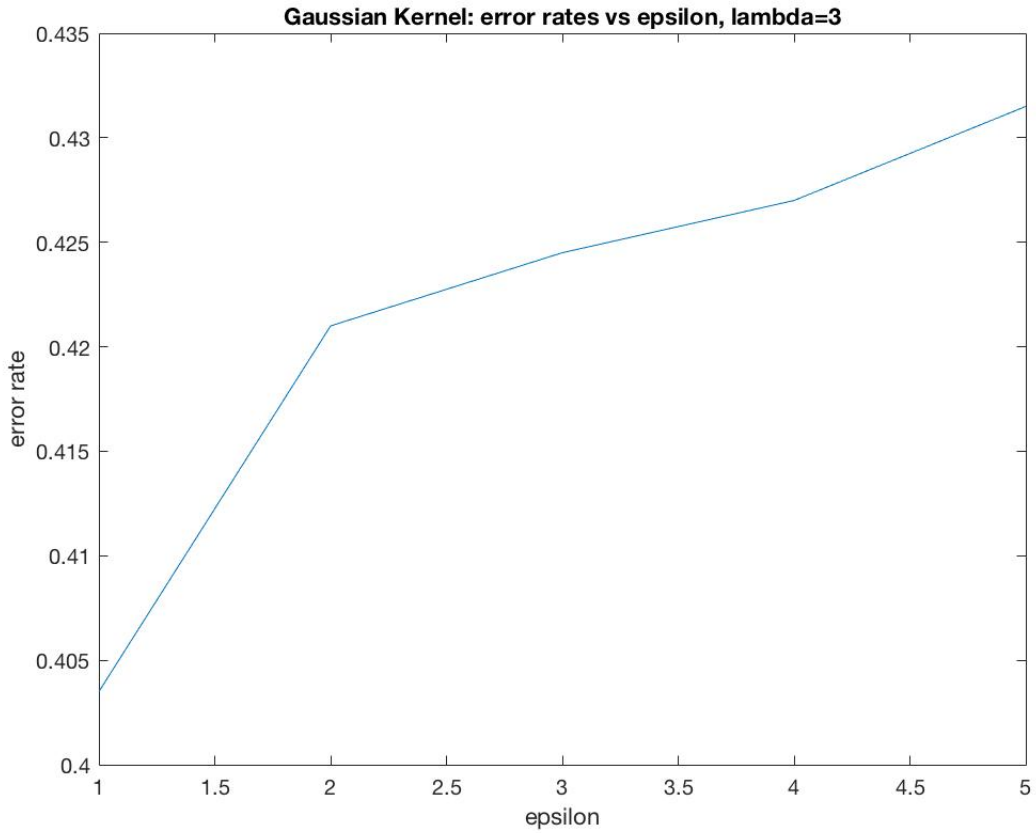
For polynomial kernel, we have(The error rate maintains 42.35% for all λ values from 1 to 5):



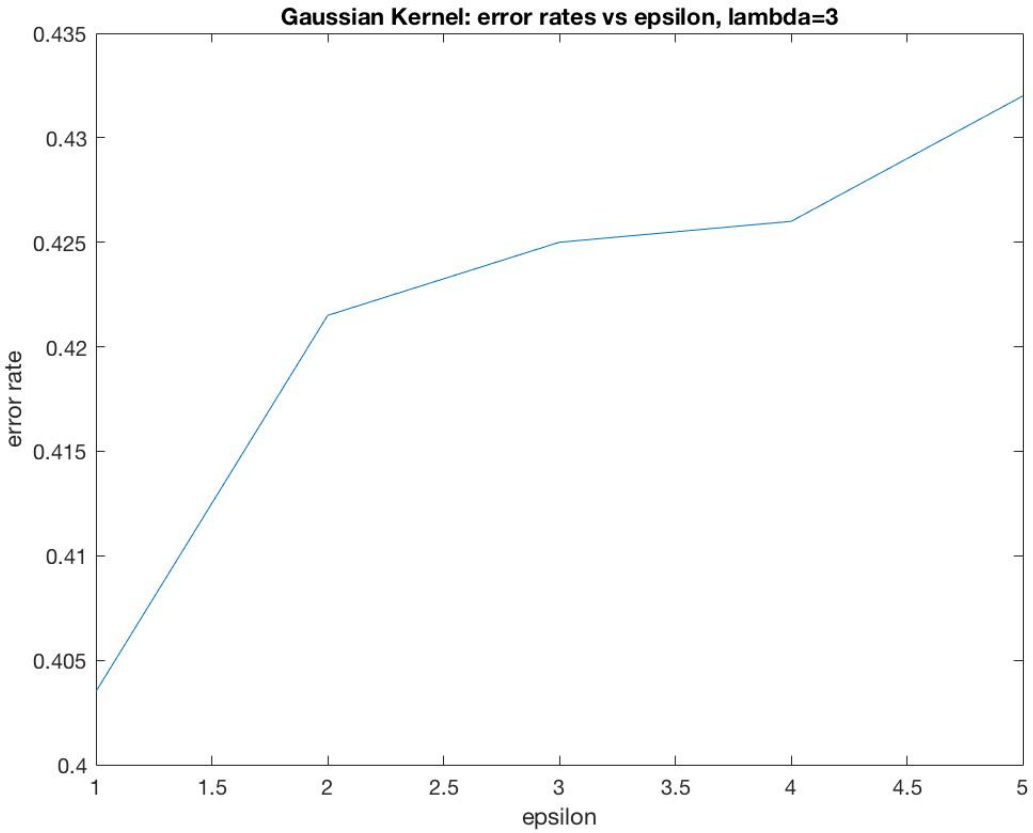
For Gaussian kernel, when $\lambda = 1$, we have:



For Gaussian kernel, when $\lambda = 3$, we have:



For Gaussian kernel, when $\lambda = 5$, we have:


Gaussian Kernel: error rates vs epsilon, lambda=3
error rate
epsilon
0.435
0.43
0.425
0.42
0.415
0.41
0.405
0.4
1
1.5
2
2.5
3
3.5
4
4.5
5

2 Exercise 2

2.1 Question 2.1

2.1

$$L(w, z, \alpha) = \frac{1}{2}\|z\|_2^2 + \frac{\lambda}{2}\|w\|_2^2 + \alpha^T(z - Xw + y)$$

$$\frac{\partial L}{\partial w}(w, z, \alpha) = \lambda w - X\alpha = 0 \Rightarrow w = \frac{1}{\lambda} X\alpha$$

$$\frac{\partial L}{\partial z}(w, z, \alpha) = z + \alpha = 0 \Rightarrow r = -\alpha$$

$$\therefore g(\alpha) = L(w, z, \alpha)$$

$$= \frac{1}{2}\|\alpha\|_2^2 - \frac{1}{2\lambda}\|X\alpha\|_2^2 + \alpha^T y$$

So the dual problem becomes

$$\max_{\alpha} -\frac{1}{2}\|\alpha\|_2^2 - \frac{1}{2\lambda}\|X\alpha\|_2^2 + \alpha^T y$$

$$= \min_{\alpha} \frac{1}{2}\|\alpha\|_2^2 + \frac{1}{2\lambda}\|X\alpha\|_2^2 - \alpha^T y$$

$$= \min_{\alpha} \frac{\lambda}{2}\|\alpha\|_2^2 + \frac{1}{2}\|X\alpha\|_2^2 - \lambda\alpha^T y$$

$$= \min_{\alpha} \frac{\lambda}{2}\alpha^T\alpha + \frac{1}{2}\alpha^T X^T X\alpha - \lambda\alpha^T y$$

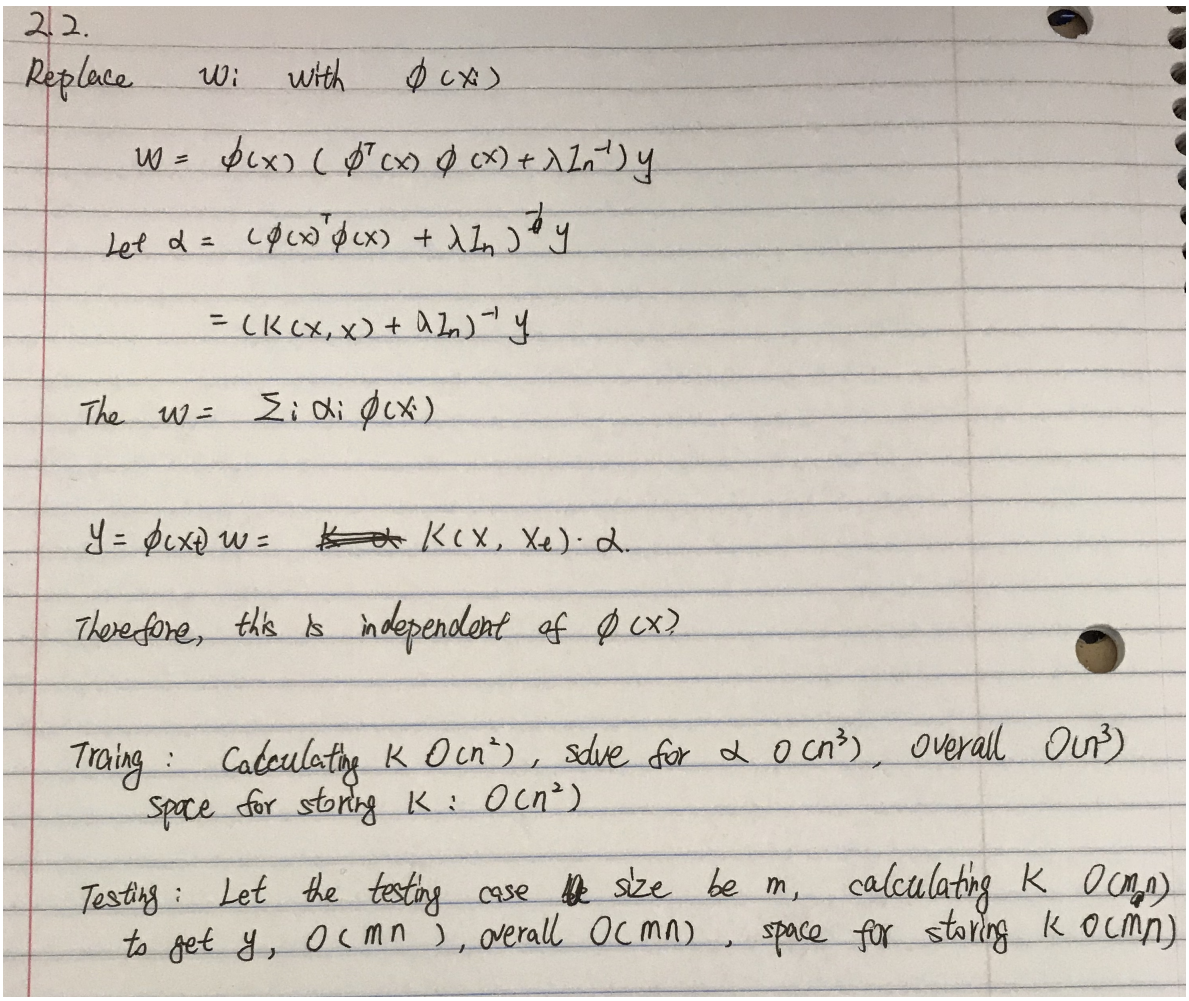
$$= \min_{\alpha} \frac{1}{2}\alpha^T(\lambda I + X^T X)\alpha - \lambda\alpha^T y$$

Let $X^T X = K$

$\therefore$ Dual is $\min_{\alpha} \frac{1}{2}\alpha^T(K + \lambda I)\alpha - \lambda\alpha^T y$, ~~$w = X(K+\lambda I)^{-1}y$~~

$\alpha = \lambda(K + \lambda I)^{-1} y$, then $w = X(K + \lambda I)^{-1} y$

2.2 Question 2.2



2.3 Question 2.3

To prove $(X^TX+\lambda I_d)^{-1} = \frac{1}{\lambda}I_d - \frac{1}{\lambda}X^T(XX^T+\lambda I_n)^{-1}X$

We can prove

$$(X^TX+\lambda I_d)^{-1} + \frac{1}{\lambda}X^T(XX^T+\lambda I_n)^{-1}X = \frac{1}{\lambda}I_d$$

$$\lambda(X^TX+\lambda I_d)^{-1} + X^T(XX^T+\lambda I_n)^{-1}X = I_d$$

Multiply both sides by X^T

$$\lambda(X^TX+\lambda I_d)^{-1}\cdot X^T + X^T(XX^T+\lambda I_n)^{-1}XX^T = X^T$$

Multiply both sides by y.

$$\lambda(X^TX+\lambda I_d)^{-1}X^Ty + X^T(XX^T+\lambda I_n)^{-1}XX^Ty = X^Ty$$

$\because (X^TX+\lambda I_d)^{-1}X^Ty = X^T(XX^T+\lambda I_n)^{-1}y$

$\therefore \lambda X^T(XX^T+\lambda I_n)^{-1}y + X^T(XX^T+\lambda I_n)^{-1}XX^Ty = X^Ty$

$\therefore X^T(XX^T+\lambda I_n)^{-1}\cdot(\lambda I_n + XX^T)\cdot y = X^Ty$

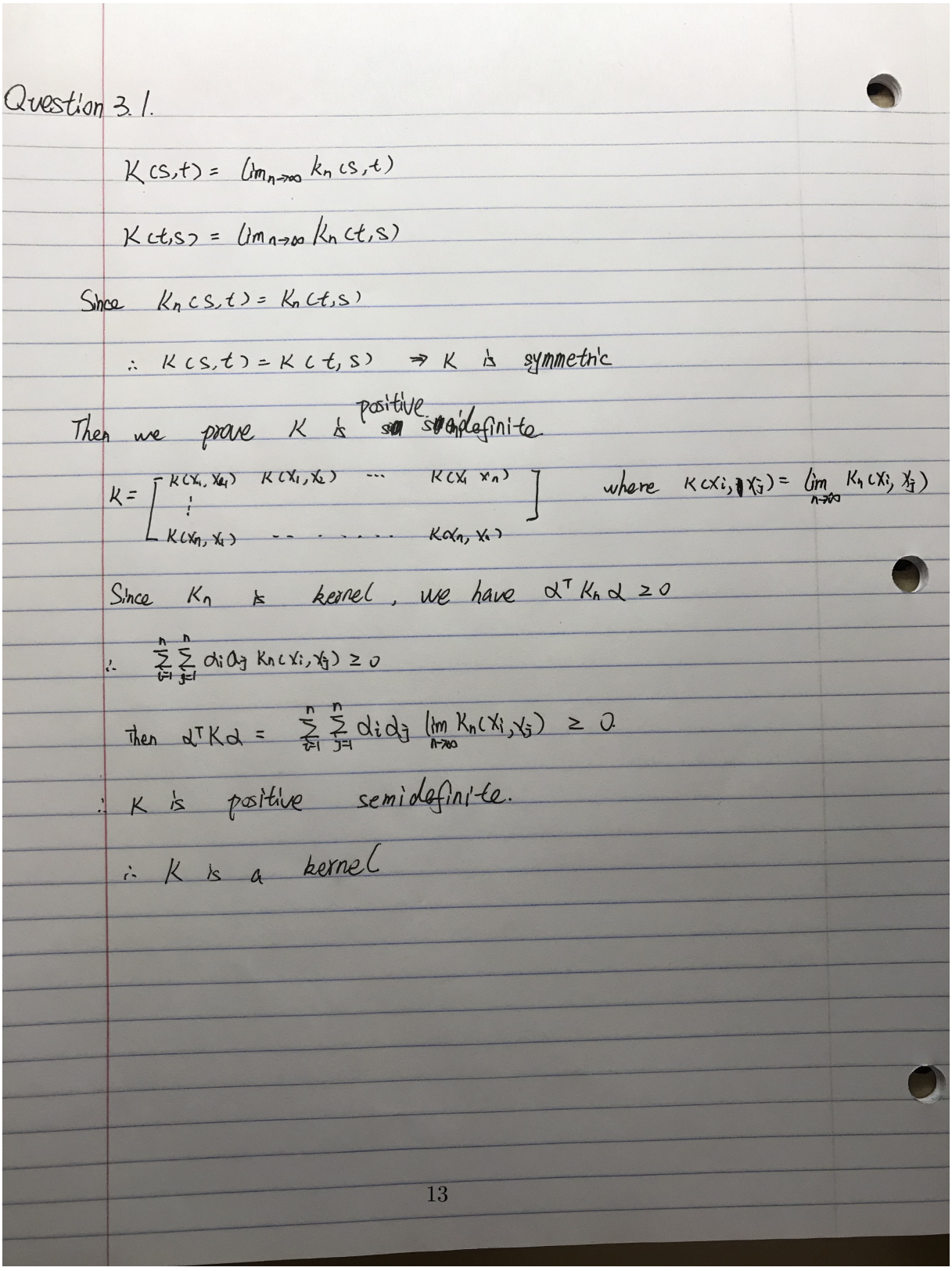
$\because (XX^T+\lambda I_n)^{-1}\cdot(\lambda I_n - XX^T) = I_n$

$\therefore X^T(XX^T+\lambda I_n)^{-1}\cdot(\lambda I_n + XX^T)\,y = X^TI_ny = X^Ty$

Proof Done

3 Exercise 3

3.1 Question 3.1



Question 3.1.

$K(s,t) = \lim_{n\to\infty} k_n(s,t)$

$K(t,s) = \lim_{n\to\infty} k_n(t,s)$

Since $k_n(s,t) = k_n(t,s)$

$\therefore K(s,t) = K(t,s) \Rightarrow K$ is symmetric

Then we prove K is positive semidefinite

$$K = \begin{bmatrix} K(x_1,x_1) & K(x_1,x_2) & \cdots & K(x_1,x_n) \\ \vdots & & & \\ K(x_n,x_1) & \cdots & \cdots & K(x_n,x_n) \end{bmatrix}$$ where $K(x_i,x_j) = \lim_{n\to\infty} k_n(x_i,x_j)$

Since k_n is kernel, we have $\alpha^T K_n \alpha \geq 0$

$\therefore \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j k_n(x_i,x_j) \geq 0$

Then $\alpha^T K \alpha = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j \lim_{n\to\infty} k_n(x_i,x_j) \geq 0$

$\therefore K$ is positive semidefinite.

$\therefore K$ is a kernel

3.2 Question 3.2 & 3.3

3.2. $e^k = \sum_{n=0}^{\infty} \frac{k^n}{n!} = 1 + k + \frac{k^2}{2!} + \cdots \frac{k^i}{i!} + \cdots$

Since if k_1, k_2 are kernels, $k_1 k_2$ is also a kernel ①

and if k is a kernel, so is λk, for any $\lambda \geq 0$ ②

and if k_1, k_2 are kernels, so is $k_1 + k_2$ ③

① implies all k^i's are kernels

② implies all $\frac{k^i}{i!}$'s are kernels

③ + ② + ① implies e^k is a kernel

3.3. $\exp(-\|x - x'\|_2^2 / 6) = \exp\left(-\frac{\|x\|_2^2 - 2x^T x' + \|x'\|_2^2}{6}\right)$

$= \sum_{i=0}^{\infty} \frac{(x^T x')^i}{i!} \cdot \left(\frac{2}{6}\right)^i \cdot \exp\left(-\frac{\|x\|_2^2}{6}\right) \cdot \exp\left(-\frac{\|x'\|_2^2}{6}\right)$

As we have proved in 3.2

$\sum_{i=0}^{\infty} \frac{(x^T x')^i}{i!}$ is a kernel.

Therefore, $\exp(-\|x - x'\|_2^2 / 6)$ is a kernel for any $6 > 0$