Problem 1: Kernels

1.1 $K_1(x,x') = \phi_1(x)^T \phi_1(x')$; $K_2(x,x') = \phi_2(x)^T \phi_2(x')$

either ① design a feature map ϕ using ϕ_1, ϕ_2 such that $K(x,x') = \phi(x)^T \phi(x')$

or ② explain why it cannot exists

Let $K_1 : \underset{(\phi_1(x))}{\mathcal{X}} \times \underset{(\phi_1(x'))}{\mathcal{X}} \to \mathbb{R}$; $K_2 : \underset{(\phi_2(x))}{\mathcal{X}} \times \underset{(\phi_2(x'))}{\mathcal{X}} \to \mathbb{R}$.

K_1 & K_2 are positive semi-definite

a) $K(x,x') = cK_1(x,x')$ for any $c \geq 0$.

Since K_1 is kernel.

$\Rightarrow K_1 = \phi_1(x) \cdot \phi_1(x)^T$ would be corresponding kernel matrix for $k_1(x,x')$

where $(K_1)_{ij} = \phi_1(x_i)^T \phi_1(x_j)$

Similarly,

$K_2 = \phi_2(x) \phi_2(x)^T$ would be corresponding to kernel matrix for $k_2(x,x')$, where $(K_2)_{ij} = \phi_1(x_i)^T \phi_1(x_j)$.

$\Rightarrow K = cK_1$

For any vector $v \in \mathbb{R}^n$.

$\Rightarrow v^T K v = v^T (cK_1) v = c v^T K_1 v \geq 0$

Hence. we showed that K is positive semi-definite matrix

Hence K is a kernel matrix $\Rightarrow k(x,x')$ is kernel function.

b) $K(x,x') = K_1(x,x') + K_2(x,x')$

As the set up above,

$K_1 = \phi_1(x)\phi_1^T(x)$

$K_2 = \phi_2(x)\phi_2^T(x)$.

$K = K_1 + K_2$.

$V^T K V = V^T(K_1+K_2)V = (V^T K_1 + V^T K_2)V = V^T K_1 V + V^T K_2 V$

Since K_1 & K_2 are psd for any vector $V \in R^n$

$\Rightarrow V^T K_1 V \geq 0$ & $V^T K_2 V \geq 0$

$\Rightarrow V^T K V = V^T K_1 V + V^T K_2 V \geq 0$.

$\Rightarrow K'$ is kernel matrix

$\Rightarrow K'(x,x')$ is kernel function.

c). $K'(x,x') = K_1(x,x') - K_2(x,x')$

$K' = K_1 - K_2$ $V^T K_1 V \geq 0$ & $V^T K_2 V \geq 0$. for any vector $V \in R^n$.

$V^T K' V = \underline{V^T K_1 V - V^T K_2 V}$.

this is not guaranteed to be ≥ 0

Hence, K' might not be kernel but if $V^T K_1 V \geq V^T K_2 V$, then K' could be kernel.

(d). $k(x,x') = k_1(x,x') \cdot k_2(x,x') = \phi_1^T(x)\phi_1(x)\phi_2^T(x)\phi_2(x)$.

we can construct $\phi_K(x) = \phi_1(x) \otimes \phi_2(x)$.

$\Rightarrow k(x,x') = \phi_K(x)^T \phi_K(x') = (\phi_1(x) \otimes \phi_2(x))^T (\phi_1(x') \otimes \phi_2(x'))$

where $\otimes$ denotes the Tensor product

Hence, it is a valid kernel.

(e) $k(x,x') = k_1(f(x), f(x')) = \phi_1(f(x))^T \phi_1(f(x'))$

let $\phi_K(x) = \phi_1(f(x))$.

Hence $k(x,x') = \phi_K(x)^T \phi_K(x) = \phi_1(f(x))^T \phi_1(f(x))$

$\Rightarrow k(x,x')$ is a valid kernel.

1.2 Show that $k'(x,x') = \sum_{i=1}^{d} a_i k(x,x')$ is a valid kernel for any valid kernel k if $a_i \geq 0$ for all $i \in [d]$?

$$k'(x,x') = \sum_{i=1}^{d} a_i k(x,x')^i = a_1 k(x,x')^1 + a_2 k(x,x')^2 + \cdots$$

accordy to a) $k' = c \cdot k$ is still valid kernel

d) $k' = k_1 \cdot k_2$ is still valid, hence

$k(x,x')^2 = k(x,x') \cdot k(x,x') \rightarrow$ Kernel ✓

$k(x,x')^3 = k(x,x')^2 \cdot k(x,x') \rightarrow$ kernel

$\vdots$

$k(x,x')^d = k(x,x')^{d-1} \cdot k(x,x') \rightarrow$ kernel.

All are valid kernel.

Then accordig to b), the sum of valid kernels are still kernel

$\Rightarrow K'(x,x') = \sum_{i=1}^{d} a_i K(x,x')^{\forall}$ is a valid kernel fuction. for all $a_i \geq 0$.

Bonus: Show that $k'(x,x') = \exp(\min(x,x'))$ is a valid kernel over input space $\mathcal{X} = [0,1] \in \mathbb{R}^2$.

$K_{ij} = k(x_i, x_j) = \exp(\min(x_i, x_j))$

For any vector $v \in \mathbb{R}^n$.

$$v^T K v = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}^T K \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = [v_1 \cdots v_n] \begin{bmatrix} K_{11} & K_{12} & \cdots & K_{1n} \\ \vdots & & & \vdots \\ K_{n1} & \cdots & & K_{nn} \end{bmatrix}_{n \times n} \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} v_i v_j \exp(\min(x_i, x_j))$$

Since the exponential fuction is alway positive

$\Rightarrow f(x) = e^x \geq 0$ for all $x \in \mathbb{R}$.

and Sine $x \in [0,1] \Rightarrow \min\{x_i, x_j\} \in [0,1]$ for all $x_i, x_j \in [0,1]$.

$\Rightarrow \exp(\min[x_i, x_j]) > 0 \Rightarrow$ the summation of positive terms is still positive.

$\Rightarrow v^T K v \geq 0$ for any $v \in \mathbb{R}^n$.

$\Rightarrow K'$ is Kernel.

Bonus continuing.

$K'(x,x') = \exp(\min(x,x'))$. K' = the kernel matrix for $K'(x,x')$.

We just need to show $\exp(K_1)$ is kernel and

$K_1 = \min(x,x')$ is kernel.

Let $c_1 \cdots c_n \in \mathbb{C}$ and $x_1, \cdots x_n \in \mathbb{R}$. Let $K(x,x') = \min(x,x')$.

Then $\sum_{i,j=1}^{n} c_i \bar{c}_j K(x_i,x_j) = \sum_{i,j=1}^{n} c_i c_j \mathbb{1}[s \le x_i]\mathbb{1}[s \le x_j]\,ds$.

$$= \int_0^\infty \left| \sum_{i=1}^{n} c_i \mathbb{1}[s \le x_i] \right|^2 ds \ge 0.$$

which means $c^T K c \ge 0$, is indeed kernel.

And K_1

Then we just need to show $K = \exp(K_1)$ is kernel.

$$\exp(K_1) = \exp(0) + \exp(0)K_1 + \frac{\exp(0)}{2!}K_1^2 + \cdots +$$

$$= 1 + K_1 + \frac{1}{2}K_1^2 + \cdots$$

Acordy to 1.4, $K' = \exp(K_1)$ is a kenel.

Hence, we showed that $K'(x,x') = \exp(\min(x,x'))$ is kernel.

Problem 2. Gradient Descent

$$\min_w \frac{1}{m}\sum_{i=1}^{m}\log(1+\exp(-y_i w^T x_i)) + R(w)$$

Show properties about when this objective is convex, L-smooth, or possibly μ-strongly convex

2-1 $R(w)=0$

$\Rightarrow F(w)=\frac{1}{m}\sum_{i=1}^{m}\log(1+\exp(-y_i w^T x_i))$

WTS: $\nabla^2 F(w) \geq \mu I$

Proof: $\nabla F(w)=\frac{1}{m}\sum_{i=1}^{m}\frac{\partial(\log(1+\exp(-y_i w^T x_i)))}{\partial w}$

$$\nabla^2 F(w)=\frac{1}{m}\sum_{i=1}^{m}\frac{-y_i x_i e^{(-y_i w^T x_i)}}{1+e^{(-y_i w^T x_i)}}$$

$$\nabla^2 F(w)=\frac{1}{m}\sum_{i=1}^{m}\frac{y_i^2 x_i^T x_i e^{-y_i w^T x_i}}{(1+e^{-y_i w^T x_i})^2}$$

Since $y_i \in \{+1,-1\} \Rightarrow y_i^2=1$

$$\Rightarrow \nabla^2 F(w)=\frac{1}{m}\sum_{i=1}^{m}\frac{x_i^T x_i e^{(y_i w^T x_i)}}{(1+e^{-y_i w^T x_i})^2}$$

$x_i^T x_i = \|x_i\|_2^2 \leq 1$

$$0 \leq \nabla^2 F(w)=\frac{1}{m}\sum_{i=1}^{m}\frac{x_i^T x_i e^{(-y_i w^T x_i)}}{(1+e^{-y_i w^T x_i})^2} \leq \frac{1}{m}\sum_{i=1}^{m}\frac{e^{(-y_i w^T x_i)}}{(1+e^{-y_i w^T x_i})^2}$$

since sigmoid fnct $\sigma(x)=\frac{1}{1+e^{-x}}$ has codomain $(0,1)$.

$$\leq \frac{1}{m}\sum_{i=1}^{m}\frac{e^{-y_i w^T x_i}+1}{(1+e^{-y_i w^T x_i})^2}$$

This is sigmoid.

$$0=\frac{1}{m}\cdot m \cdot 0 < \left(=\frac{1}{m}\sum_{i=1}^{m}\boxed{\frac{1}{1+e^{-y_i w^T x_i}}}\right) \leq \frac{1}{m}\cdot mI = I.$$

However, we can't find its lower bound but only it is >0.

$\Rightarrow$ we cannot say it is strong convex but only it is convex.

i.e there doesn't exist a μ s.t $\nabla^2 F(w) \geq \mu I$.

2-2. In order to show it's 1-smooth.

We have gotten the $\nabla^2 F(w)$ previous part:

$$\nabla^2 F(w) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{1+e^{-y_i w^T x_i}}$$

As we see, $\frac{1}{1+e^{-y_i w^T x_i}}$ is a squashing (Sigmoid) Function.

$$\sigma(w) = \frac{1}{1+e^{-y_i w^T x_i}} \in (0,1)$$

Hence $\nabla^2 F(w) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{1+e^{-y_i w^T x_i}} \in \left(\frac{1}{m} \cdot m \cdot 0, \frac{1}{m} \cdot m \cdot I\right)$

$\Rightarrow (0, I)$.

$\Rightarrow \nabla^2 F(w) \leq \begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix} = LI = 1 \cdot I$.

which means $L = 1$.

$\Rightarrow F(w)$ is 1-smooth fuction.

2.3 convergence rate of GD when $R(w)=0$

$\|w_{T+1}-w^*\|_2 \le \varepsilon$

$F(w_{T+1})-F(w_*) \le \frac{L}{2T}$ > The LHS's meaning are same, both mean as T+1 close to the optimal value, one is for domain & the other is for codomain, but the round T needed should be same.

$\frac{L}{2T}=\varepsilon \Rightarrow \frac{1}{2T}=\varepsilon$.

$\boxed{T=\frac{1}{2\varepsilon}\sim O(\frac{1}{2\varepsilon})}$

$\Rightarrow$ convergence rate $\boxed{\varepsilon\sim O(\frac{1}{2T})}$.

$\rightarrow \frac{1}{2\varepsilon}\approx T \Rightarrow \varepsilon\approx\frac{1}{T}\Rightarrow \varepsilon\sim O(\frac{1}{T})$

2.4. $R(w)=\sum_{j=1}^{d}\lambda_j w_j^2$ (regularizer).

Show that the objective with $R(w)$ is μ-strongly convex and L-smooth for $\mu=2\min_{j\in[d]}\lambda_j$ and $L=1+2\max_{j\in[d]}\lambda_j$

$\Rightarrow F(w)=\underbrace{\frac{1}{m}\sum_{i=1}^{m}\log(1+\exp(-y_i w^T x_i))}_{F_0(w)}+\sum_{j=1}^{d}\lambda_j w_j^2$.

① $\nabla^2 F(w)=\nabla^2 F_0(w)+\nabla^2(\sum_{j=1}^{d}\lambda_j w_j^2)$

$\nabla(\sum_{j=1}^{d}\lambda_j w_j^2)=\nabla(\sum_{j=1}^{d}2\lambda_j w_j)=\sum_{i=1}^{d}2\lambda_j$

sigmoid $\in(0,1)$.

$\Rightarrow \nabla^2 F(w)=\frac{1}{m}\sum_{i=1}^{m}\frac{1}{1+e^{y_i w^T x_i}}+\sum_{i=1}^{d}2\lambda_j$.

$=\frac{1}{m}\underbrace{\sum_{i=1}^{m}\frac{1}{1+e^{-y_i w^T x_i}}}_{\text{Sigmoid}}+\underbrace{2\sum_{i=1}^{d}\lambda_j}_{}\in(0+2\sum_{j=1}^{d}\lambda_j,\ 1+2\sum_{j=1}^{d}\lambda_j)$

Let's consider the Hessian matrix for $\sum_{i=1}^{d}\lambda_i w_i$ $\Rightarrow(0+\Pi,\ 1+\Pi)$.

upperbd of $\nabla^2 F(w)$ $=2\Lambda+I=\begin{bmatrix}2\lambda_1+1 & & \\ & \ddots & \\ & & 2\lambda_d+1\end{bmatrix}\le\begin{bmatrix}2\lambda_{max}+1 & & \\ & \ddots & \\ & & 2\lambda_{max}+1\end{bmatrix}=L\cdot I$

$\Rightarrow L=1+2\max_{j\in[d]}\lambda_j$

$2\min_{j\in[d]}\lambda_j \le \nabla^2 F(w) \le 1+2\sum_{i=1}^{d}\lambda_j \le 1+2\max_{j\in[d]}\lambda_j \Rightarrow$ It's $L=(2\max_{j\in[d]}\lambda_j+1)$-Smooth function.

and $\nabla^2 F(w) \geq (2\min_{j\in[d]} \lambda_j) I = \mu I$

Lower bd of $\nabla^2 F(w) = A + 0 = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{bmatrix} \geq \begin{bmatrix} \lambda_{min} & & \\ & \ddots & \\ & & \lambda_{min} \end{bmatrix}$

where $\mu = 2\min_{j\in[d]} \lambda_j \Rightarrow$ it's $(2\min\lambda_j)$-Strong convex function.

Hence we proved $F(w)$ with $R(w)$ is both L-smooth & with $\mu = 2\min_{j\in[d]} \lambda_j$ & $L = 2\max_{j\in[d]} \lambda_j + 1$ μ-strong convex.

Bonus.

$$\Rightarrow \|w_{T+1} - w^*\|_2^2 \leq (1 - \frac{\mu}{L})^T \|w_1 - w^*\|_2^2$$

$$F(w) = \frac{1}{m}\sum_{i=1}^m \log(1 + \exp(-y_i w^T x_i)) + \sum_{j=1}^d \lambda_j w_j^2$$

$(1 - \frac{\mu}{L})^T \|w_1 - w^*\|_2^2$

Assume $\|w^*\| = 1$ & $w_1 = 0$.

$= (1 - \frac{\mu}{L})^T \cdot 1 \leq \varepsilon$.

$1 - \frac{\mu}{L} = \log \varepsilon$.

$\Rightarrow T = \frac{L}{\mu} \log(\frac{1}{\varepsilon})$

$= \frac{2\max\lambda_j + 1}{2\min\lambda_j} \log(\frac{1}{\varepsilon})$

$= \frac{\max\lambda_j + \frac{1}{2}}{\min\lambda_j} \log(\frac{1}{\varepsilon})$.

$\mu = 2\min_{j\in[d]} \lambda_j$

$L = 2\max_{j\in[d]} \lambda_j + 1$

$\frac{\mu}{L} = \frac{\min\lambda_j}{\max\lambda_j + 1}, j \in [d]$

$1 - \frac{\mu}{L} = \frac{\frac{1}{2}\max\lambda_j - \min\lambda_j}{\max\lambda_j + 1}$

$\Rightarrow T \sim O(\frac{\max\lambda_j + \frac{1}{2}}{\min\lambda_j} \log(\frac{1}{\varepsilon}))$

$\sim O(\log\frac{1}{\varepsilon})$.

$\log \varepsilon$

$\Rightarrow$ Converge fast.

$O(e^{-T})$

T)

another Bound for T:

$$T = \frac{\frac{1}{2}+\max\lambda_j}{\min\lambda_j}\log\left(\frac{1}{\varepsilon}\right)$$

$$\log\varepsilon = -T\frac{\min\lambda_j}{\frac{1}{2}+\max\lambda_j}.$$

$$\varepsilon = e^{-T\left(\frac{\min\lambda_j}{\frac{1}{2}+\max\lambda_j}\right)} = e^{-\left(\frac{\min\lambda_j}{\frac{1}{2}+\max\lambda_j}\right)T} \sim O\left(e^{-\left(\frac{\min\lambda_j}{\frac{1}{2}+\max\lambda_j}\right)T}\right)$$

$$\sim O\left(e^{-T}\right).$$

$$\Rightarrow \varepsilon \sim O\left(e^{-\left[\left(\frac{\min\lambda_j}{\frac{1}{2}+\max\lambda_j}\right)T\right]}\right):$$

$$T \sim O\left(\left(\frac{\max\lambda_j+0.5}{\min\lambda_j}\right)\cdot\log\left(\frac{1}{\varepsilon}\right)\right).$$