1. Shown that in regularized linear regression, $\boldsymbol{w}^* \underset{①}{=} (\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I}_d)^{-1}\boldsymbol{X}^\top\boldsymbol{r} \underset{②}{=} \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top+\lambda\boldsymbol{I}_N)^{-1}\boldsymbol{r}$, a linear combination of the examples.

- Regularized linear regression can be written as follow :

  $\arg\min \frac{1}{2}||r - Xw||^2 + \frac{1}{2}\lambda||w||^2$.

- Applying first derivative (to w), we get $X^TXw - X^Tr + \lambda w$

- Minimum occurs when $X^TXw + \lambda \mathbf{I}_d w = (X^TX+\lambda \mathbf{I}_d)w = X^Tr$,

  and $w^* = (X^TX+\lambda \mathbf{I}_d)^{-1}X^Tr$

1. Shown that in regularized linear regression, $\boldsymbol{w}^* \overset{①}{=} (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_d)^{-1} \boldsymbol{X}^\top \boldsymbol{r} \overset{②}{=} \boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{X}^\top + \lambda \boldsymbol{I}_N)^{-1} \boldsymbol{r}$, a linear combination of the examples.

- $(A+BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1}+DA^{-1}B)^{-1}DA^{-1}$ $\quad A=\lambda I_d$, $B=X^T$, $C=I_N$, $D=X$

- $(X^TX+\lambda I_d)^{-1}X^T = (1/\lambda)I_d - (1/\lambda)I_d X^T(I_N+X\cancel{(1/\lambda)I_d}X^T)^{-1}X\cancel{(1/\lambda)I_d}X^T$

$= (1/\lambda)I_dX^T - (1/\lambda)I_dX^T(\lambda I_N+XX^T)^{-1}XX^T$

$= (1/\lambda)\, I_dX^T(I_N - \boxed{(\lambda I_N+XX^T)^{-1}XX^T - (\lambda I_N+XX^T)^{-1}(\lambda I_N)} + (\lambda I_N+XX^T)^{-1}(\lambda I_N))$

$= (1/\lambda)\, I_dX^T(I_N - \boxed{(\lambda I_N+XX^T)^{-1}(\lambda I_N+XX^T)} + (\lambda I_N+XX^T)^{-1}(\lambda I_N))$

$= (1/\lambda)\, I_dX^T(I_N - I_N + (\lambda I_N+XX^T)^{-1}(\lambda I_N))$

$= X^T(\lambda I_N+XX^T)^{-1}$

- 因此， $(X^TX+\lambda I_d)^{-1}X^Tr = X^T(\lambda I_N+XX^T)^{-1}r$

2. Show that in an RKHS, the inner product $\langle f, g\rangle := \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha^{(i)}\beta^{(j)}k(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(j)})$ for any $f = \sum_{i=1}^{n} \alpha^{(i)}k(\boldsymbol{x}^{(i)}, \cdot)$ and $g = \sum_{j=1}^{m} \beta^{(j)}k(\boldsymbol{y}^{(j)}, \cdot)$ is well-defined; i.e., it satisfies

(a) *symmetry*: $\langle f, g\rangle = \langle g, f\rangle$; 交換律

- $\sum_{i=1}^{n} \sum_{j=1}^{m}$ 是Symmetric，$\alpha^{(i)}\beta^{(j)}$ 也是Symmetric，因此只要 k 是symmetric，那麼 $\sum_{i=1}^{n} \sum_{j=1}^{m} \alpha^{(i)}\beta^{(j)}k(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(j)})$ 就是Symmetric。事實如此，因為對於原始資料**x**，它的Lifting function Φ 對於**x** 的每個維度的feature都是對稱的。

(b) *linearity*: $\langle af + bg, h\rangle = a\langle f, h\rangle + b\langle g, h\rangle$ for any $a, b \in \mathbb{R}$; and 加法結合律、乘法分配律

- 關於純量積，a $\sum_{i=1}^{n} \alpha^{(i)}k(\boldsymbol{x}^{(i)}, \cdot)$ = $\sum_{i=1}^{n}$ a $\alpha^{(i)}k(\boldsymbol{x}^{(i)}, \cdot)$ 因此如此定義滿足乘法分配律。
- 而af+bg= $\sum_{i=1}^{n} a\alpha^{(i)} k(x^{(i)},\cdot) + \sum_{j=1}^{m} b\beta^{(j)} k(y^{(j)},\cdot)$，定義h= $\sum_{q=1}^{p} \gamma^{(q)} k(z^{(q)},\cdot)$
- $\langle af+bg,h\rangle = \sum_{q=1}^{p} \left(\sum_{i=1}^{n} a\alpha^{(i)} \gamma^{(q)} k(x^{(i)},z^{(j)}) + \sum_{j=1}^{m} b\,\beta^{(j)} \gamma^{(q)} k(y^{(j)}, z^{(j)}) \right)$

$= \sum_{i=1}^{n} \sum_{q=1}^{p} a\alpha^{(i)} \gamma^{(q)} k(x^{(i)},z^{(j)}) + \sum_{j=1}^{m} \sum_{q=1}^{p} b\,\beta^{(j)} \gamma^{(q)} k(y^{(j)}, z^{(j)}) = a\langle f,h\rangle + b\langle g,h\rangle$

2. Show that in an RKHS, the inner product $\langle f, g\rangle := \sum_{i=1}^{n}\sum_{j=1}^{m}\alpha^{(i)}\beta^{(j)}k(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(j)})$ for any $f = \sum_{i=1}^{n}\alpha^{(i)}k(\boldsymbol{x}^{(i)}, \cdot)$ and $g = \sum_{j=1}^{m}\beta^{(j)}k(\boldsymbol{y}^{(j)}, \cdot)$ is well-defined; i.e., it satisfies

(c) *positive definiteness:* $\langle f, f\rangle \geq 0$ with equality holds iff $f(\cdot) = 0(\cdot)$.

自己內積自己恆正

- $\langle f,f\rangle = \Sigma_{i=1}^{n} \Sigma_{j=1}^{n} \alpha^{(i)} \alpha^{(j)} k(x^{(i)}, x^{(j)})$ 根據定義

$= \Sigma_{i=1}^{n} \Sigma_{j=1}^{n} k(\alpha^{(i)}x^{(i)}, \alpha^{(j)}x^{(j)})$ 純量積(乘法分配律)

$= \Sigma_{i=1}^{n} \Sigma_{j=1}^{n} \langle \alpha^{(i)}\Phi(x^{(i)}), \alpha^{(j)}\Phi(x^{(j)}) \rangle$ 根據定義

$= \langle \Sigma_{i=1}^{n} \alpha^{(i)}\Phi(x^{(i)}), \Sigma_{j=1}^{n} \alpha^{(j)}\Phi(x^{(j)}) \rangle$ 加法結合律

$= \langle \Sigma_{i=1}^{n} \alpha^{(i)}\Phi(x^{(i)}), \Sigma_{i=1}^{n} \alpha^{(i)}\Phi(x^{(i)}) \rangle$ 就是這樣

$= \| \Sigma_{i=1}^{n} \alpha^{(i)}\Phi(x^{(i)}) \|^2$

≥ 0， $\alpha^{(i)}>0$，因此等號成立在 $f(\cdot) = 0(\cdot)$

3. Show that in a large-margin linear classifier, the margin between the hyperplanes $\{\boldsymbol{x} : \boldsymbol{w}^\top \boldsymbol{x} - b = 1\}$ and $\{\boldsymbol{x} : \boldsymbol{w}^\top \boldsymbol{x} - b = -1\}$ is $2/\|\boldsymbol{w}\|$.

- 推導兩平面間的距離公式的過程通常很冗長。因此假設我們已經有平行的觀念，亦即在N+1維空間中的2個N維超平面互相平行，即代表它們有相同的法向量，同時此兩個超平面的距離在任意處皆相等
- $w^Tx - b - 1 = 0$ 和 $w^Tx - b + 1 = 0$ 這兩個超平面互相平行，因為它們的法向量皆為w
- 對於超平面$w^Tx - 1 = b$上任一點x_0，x_0到超平面$w^Tx + 1 = b$的最短路徑，必然是沿著它們的法向量$w/\|w\|$走δ的距離，因此可以記為$w^Tx_0 - 1 = b$， $w^T(x_0+\delta w/\|w\|) + 1 = b$
- 因此， $w^Tx_0 - 2 = w^T(x_0+\delta/\|w\|)$ ，即$w^T(\delta/\|w\|)w = (\delta/\|w\|)w^Tw = \delta\|w\|^{(-1+2)} = \delta\|w\| = -2$
- 最後， $|\delta| = 2/\|w\|$ ，我們得到這兩個超平面距離為$2/\|w\|$

4. Prove the Semiparametric Representer theorem.

- 改寫Representer Theorem的證明

1. 將$\tilde{g}$中的g分解為平行於和垂直於空間span$(k(x^{(1)},\cdot),\ldots, k(x^{(N)},\cdot))$的部分，得到$\tilde{g} = g_{//} + g_{\perp} + b\psi$，其中 $g_{//} = \sum_{t=1}^{N} c_t k(\boldsymbol{x}^{(t)}, \boldsymbol{x})$
2. Loss function可以寫成 $L((\mathbf{x}^{(i)},\mathbf{r}^{(i)}, g_{//}(\mathbf{x}^{(i)})+b\psi(\mathbf{x}^{(i)}))_{i=1,\ldots,N})$，因為$g_{\perp}(x^{(i)})=0$
3. $\|g_{//}\| \le \|g\|$，因此$\Omega(\|g_{//}\|_{RKHS}) \le \Omega(\|g\|_{RKHS})$，而將g代換成$g_{//}$不會對loss function和constraints造成任何影響，所以我們應該將g代換成$g_{//}$
4. 因此，若得以最小化loss function，必須有$g_{\perp}=0$，此時$\tilde{g} = g_{//} + b\psi$

4. Prove the Semiparametric Representer theorem.

5. 由於 $g_{//} \in \mathrm{span}(k(\mathbf{x}^{(1)},\cdot),\ldots, k(\mathbf{x}^{(N)},\cdot))$，我們可以將 $\tilde{g}^{(i)}{}_{i=1,\ldots,N}$改寫成

$$\tilde{g}^{(i)} = \Sigma_{t=1}^{N} c_t\, k(\mathbf{x}^{(t)},\mathbf{x}^{(i)}) + b\psi(\mathbf{x}^{(i)})$$

6. 因此，$\tilde{h}(\mathbf{x})$便有 $\Sigma_{t=1}^{N} c_t\, k(\mathbf{x}^{(t)},\mathbf{x}) + b\psi(\mathbf{x})$ 的形式

6. Show that the *Area Under the ROC Curve* (AUC) is equal to the probability that a classifier ranks a randomly chosen positive instance higher than a randomly chosen negative one.

Label	Rank
+1	1
+1	2
+1	3
-1	4
+1	5
-1	6
-1	7
+1	8
-1	9
+1	10

table (6-a)

X^+_i \ X^-_j	4	6	7	9
10	*0*	*0*	*0*	*0*
8	*0*	*0*	*0*	*1*
5	*0*	*1*	*1*	*1*
3	*1*	*1*	*1*	*1*
2	*1*	*1*	*1*	*1*
1	*1*	*1*	*1*	*1*

Area occupied by *1*

$$= \sum_{i=0}^{N^+}\sum_{j=0}^{N^-} C(i,j)$$

table (6-b)

- 在任何一個資料集，對於這個Classifier而言，隨機選擇一個Positive instance(X^+_i),$1 \le i \le N^+$和一個Negative instance(X^-_j),$1 \le j \le N^-$，那麼Rank(X^+_i)<Rank(X^-_j)的機率，即為：

$$\sum_{i=0}^{N^+}\sum_{j=0}^{N^-} C(i,j) / (N^+N^-),\ C(i,j) \overset{\text{如table (6-b)}}{=} \begin{cases} 1, \text{ if } Rank(X^+_i) < Rank(X^-_j) \\ 0, \text{ if } Rank(X^+_i) > Rank(X^-_j) \end{cases},$$

而這恰好就是「$\overset{\text{table (6-b)}}{\text{左方表格中}}$，被*1*所佔據的面積」

(表格中*0*和*1*所在的區域其長寬皆為1，而每小格面積均等)，

而表格中紅色折線即為ROC Curve