Homework #2

(Due: Apr 19)

1 Theory

1.1 Gaussian Distributions

To prove :

$$E[x] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} x dx \tag{1}$$

Substituting

$$t = \frac{(x-\mu)}{\sqrt{2}\sigma}$$

we get

$$E[x] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (\sqrt{2}\sigma t + \mu)(\sqrt{2}\sigma) e^{-t^2} dt \tag{2}$$

$$\implies E[x] = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} (\sqrt{2}\sigma t + \mu) e^{-t^2} dt \tag{3}$$

$$\implies E[x] = \frac{1}{\sqrt{\pi}} \left(\sqrt{2}\sigma \int_{-\infty}^{\infty} t e^{-t^2} dt + \mu \int_{-\infty}^{\infty} e^{-t^2} dt \right) \tag{4}$$

$$\implies E[x] = \frac{1}{\sqrt{\pi}} \left(\sqrt{2}\sigma \left[\frac{-e^{-t^2}}{2} \right]_{-\infty}^{\infty} + \mu \int_{-\infty}^{\infty} e^{-t^2} dt \right) \tag{5}$$

$$\implies E[x] = \frac{\mu}{\sqrt{\pi}} \left(\int_{-\infty}^{\infty} e^{-t^2} dt \right) \tag{6}$$

Directly using the result of $\left(\int_{-\infty}^{\infty} e^{-t^2} dt \right) = \sqrt{\pi}$, since the function is non-integrable in one dimension.

$$\implies E[x] = \frac{\mu}{\sqrt{\pi}} (\sqrt{\pi}) = \mu \tag{7}$$

Since,

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx = 1 \tag{8}$$

Let $t = \sigma^2$, differentiating the above with respect to t we get,

$$\implies \int_{-\infty}^{\infty} \frac{-1}{2\sqrt{2\pi}t^{\frac{3}{2}}} e^{\frac{-(x-\mu)^2}{2t}} dx + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi t}} e^{\frac{-(x-\mu)^2}{2t}} \left(\frac{(x-\mu)^2}{2t^2} \right) dx = 0 \tag{9}$$

$$\implies \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}t^{\frac{3}{2}}} e^{\frac{-(x-\mu)^2}{2t}} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi t}} e^{\frac{-(x-\mu)^2}{2t}} \left(\frac{(x-\mu)^2}{t^2} \right) dx \tag{10}$$

$$\implies \frac{1}{t} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi t}} e^{\frac{-(x-\mu)^2}{2t}} dx = \frac{1}{t^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi t}} e^{\frac{-(x-\mu)^2}{2t}} (x^2 - 2\mu x + \mu^2) dx \tag{11}$$

$$\implies t = E[x^2] - 2\mu E[x] + \mu^2 \tag{12}$$

$$\begin{aligned} \implies E[x^2] &= t + 2\mu E[x] - \mu^2 \\ &= t + 2\mu^2 - \mu^2 \\ &= \sigma^2 + \mu^2 \end{aligned} \tag{13}$$

Therefore $var(x) = E[x^2] - (E[x])^2 = \sigma^2 + \mu^2 - \mu^2 = \sigma^2$.

1.2 Strongly convex function

For a strongly convex function with parameter λ we can write

$$f(\alpha u + (1-\alpha)w) \leq \alpha f(u) + (1-\alpha) f(w) - \frac{\lambda}{2}\alpha(1-\alpha)\|u-w\|^2 \tag{14}$$

$$\implies f(\alpha u + (1-\alpha)w) \leq \alpha f(u) + (1-\alpha) f(w) - \frac{\lambda}{2}\alpha(1-\alpha)\|w-u\|^2 \tag{15}$$

$$\implies f(w + \alpha(u-w)) \leq \alpha(f(u) - f(w)) + f(w) - \frac{\lambda}{2}\alpha(1-\alpha)\|w-u\|^2 \tag{16}$$

$$\implies \frac{f(w + \alpha(u-w)) - f(w)}{\alpha} \leq f(u) - f(w) - \frac{\lambda}{2}(1-\alpha)\|w-u\|^2 \tag{17}$$

$$\implies \frac{f(w + \alpha(u-w)) - f(w)}{\alpha(u-w)}(u-w) \leq f(u) - f(w) - \frac{\lambda}{2}(1-\alpha)\|w-u\|^2 \tag{18}$$

Taking the limit as α goes to 0, we have

$$\implies f^{'}(w)(u-w) \leq f(u) - f(w) - \frac{\lambda}{2}\|w-u\|^2 \tag{19}$$

$$\implies \langle v, (u-w)\rangle \leq f(u) - f(w) - \frac{\lambda}{2}\|w-u\|^2 \tag{20}$$

Multiplying both sides by -1.

$$\implies \langle v, (w-u)\rangle \geq f(w) - f(u) + \frac{\lambda}{2}\|w-u\|^2 \tag{21}$$

Since $\langle u, w\rangle = \langle w, u\rangle$ we have,

$$\langle (w-u), v\rangle \geq f(w) - f(u) + \frac{\lambda}{2}\|w-u\|^2 \tag{22}$$

1.3 Kernel construction

1.3.1 a

Since

$$\alpha K_1(u,v) = \langle \sqrt{\alpha}\phi_1(u), \sqrt{\alpha}\phi_1(v) \rangle \tag{23}$$

and

$$\beta K_2(u,v) = \langle \sqrt{\beta}\phi_2(u), \sqrt{\beta}\phi_2(v) \rangle \tag{24}$$

Therefore

$$\begin{aligned} K(u,v) &= \alpha K_1(u,v) + \beta K_2(u,v) \\ &= \langle \sqrt{\alpha}\phi_1(u), \sqrt{\alpha}\phi_1(v) \rangle + \langle \sqrt{\beta}\phi_2(u), \sqrt{\beta}\phi_2(v) \rangle \\ &= \langle [\sqrt{\alpha}\phi_1(u), \sqrt{\beta}\phi_2(u)], [\sqrt{\alpha}\phi_1(v), \sqrt{\beta}\phi_2(v)] \rangle \end{aligned} \tag{25}$$

As we can see it is represented as an inner product of 2 vectors which can be interpreted as

$$\langle \Phi_1(u), \Phi_2(v) \rangle \tag{26}$$

1.3.2 b

$$K_1(u,v) = \sum_i \phi_i(u), \phi_i(v) \tag{27}$$

$$K_2(u,v) = \sum_j \psi_i(u), \psi_j(v) \tag{28}$$

$$\begin{aligned} \implies K_1(u,v)K_2(u,v) &= \left(\sum_i \phi_i(u), \phi_i(v) \right) \left(\sum_j \psi_j(u), \psi_j(v) \right) \\ &= \sum_{i,j} \phi_i(u), \psi_j(u), \phi_i(v), \psi_j(v) \end{aligned} \tag{29}$$

We can write $\Phi_k = \phi_i(u), \psi_j(u)$.

$$\implies K_1(u,v)K_2(u,v) = \sum_k \Phi_k(u)\Phi_k(v) = K(u,v) \tag{30}$$

Therefore, product of valid kernels is a valid kernel.

1.4 Local minimum

Lets consider a sample point (x,y) such that $x = c$ and $y = 1$. By just using a point in the sample space we can show that 0-1 loss function suffers from local minima.

Let $w = -c$, $sign(\langle w, x\rangle) = -1 \neq y$, so $L_s(w) = 1$.
Let ϵ be a small scalar value and for every $w^{'}$, $\|w^{'} - w\| \leq \epsilon$.

$$\begin{aligned}\langle w^{'}, x\rangle &= \langle w, x\rangle + \langle w^{'} - w, x\rangle \\ &= -c^2 + \langle w^{'} - w, x\rangle \\ &\leq -c^2 + \|w^{'} - w\|\|x\| \text{ (Using Cauchy Schwartz Inequality)} \\ &= -c^2 + \epsilon c \\ &< 0\end{aligned} \tag{31}$$

Therefore $L_s(w^{'}) = L_s(w) = 1$, and hence w is a local minimum.
Let $w^* = c$, $\langle w^*, x\rangle = c^2$ and so $L_s(w^*) = 0$.
$L_s(w^*) < L_s(w)$ which shows that w is not a global minimum.

1.5 Learnability of logistic regression

Let $f(x) = \log(1 + e^x)$. The gradient of $f(x)$ is

$$f^{'}(x) = \frac{e^x}{1 + e^x} = \frac{1}{1 + e^{-x}} < 1$$

for $x \in R$, so the function f(x) is 1-Lipschitz.
$g(-y\langle w, x\rangle)$ is B-Lipschitz since the norm is bounded by B.
Using claim 12.7 in the book, which says that if $f(x) = g_1(g_2(x))$ and g_1 is ρ_1-Lipschitz and g_2 is ρ_2-Lipschitz, then $f(x)$ is $\rho_1\rho_2$-Lipschitz.
Therefore $l(w, \langle x, y\rangle))$ is (1*B)-Lipschitz = B-Lipschitz.

$$f^{''}(x) = f^{'}\left(\frac{1}{1 + e^{-x}}\right) = \frac{e^{-x}}{(1 + e^{-x})^2} = \frac{1}{(1 + e^{x})(1 + e^{-x})} = \frac{1}{(2 + e^{x} + e^{-x})} \leq \frac{1}{4} \tag{32}$$

Since $f^{''}(x) \leq \frac{1}{4}$, we can conclude that $f^{'}(x)$ is $\frac{1}{4}$-Lipschitz. Since $f^{''}(x)$ is nonnegative, $f(x)$ is also convex. Therefore from the definition of smooth function we have that $f(x)$ is a $\frac{1}{4}$ smooth function. We know that composition of a smooth scalar function over a linear function preserves smoothness. Since $l(w, \langle x, y\rangle)) = f(-y\langle w, x\rangle)$, l is $\frac{1}{4}\|x\|^2 = \frac{1}{4}B^2$ smooth.

The norm of the hypothesis class H is bounded above by B. Therefore it satisfies the first requirement of the Convex-Lipschitz Bounded and Convex-Smooth Bounded problem.

According to claim 12.4 in the book, the loss function $l(w, \langle x, y\rangle))$ is convex since its a composition of the convex function f onto a linear function. The function is also B-Lipschitz as proved earlier. Therefore the problem is a Convex-Lipschitz Bounded problem with parameters B, B.

The loss function $l(w, \langle x, y\rangle)) = \log(1 + e^{-y\langle w, x\rangle})$ is non-negative since logarithm function is non-negative for values greater than 1. It is also convex and $\frac{1}{4}B^2$ smooth as proved earlier. Therefore the problem is a Convex-Smooth Bounded problem with parameters $\frac{1}{4}B^2, B$.

1.6 Learnability of Halfspaces with hinge loss

$$l = max\{0, 1 - y\langle w, x\rangle\}$$

For some $x \in R^d$ and $y \in \{-1, +1\}$,

$$g(w) = 1 - y\langle w, x\rangle \tag{33}$$

is a convex function since $g^{'}(w)$ is constant(monotonically non-decreasing) and $g^{''}(w) = 0$ (which is non-negative).
Using the property that the maximum of convex functions is also convex(from the Claim 12.5 in the book), we get that $l = max\{0, 1 - y\langle w, x\rangle\}$ is also a convex function.
Let w_1, w_2 be two vectors such that $w_1, w_2 \in R^d$, then if we show that $\|l_1 - l_2\| \leq R\|w_1 - w_2\|$ then we can say that l is R-lipschitz.
Case 1:($y\langle w_1, x\rangle \geq 1, y\langle w_2, x\rangle \geq 1$)

$$\begin{aligned} \|l_1 - l_2\| &= 0 - 0 \\ &\leq R\|w_1 - w_2\| \end{aligned} \tag{34}$$

Case 2:($y\langle w_1, x\rangle < 1, y\langle w_2, x\rangle \geq 1$)

$$\begin{aligned} \|l_1 - l_2\| &= l_1 - l_2 \\ &= (1 - y\langle w_1, x\rangle) - 0 \\ &< (1 - y\langle w_1, x\rangle) - (1 - y\langle w_2, x\rangle) \\ &= y\langle w_2 - w_1, x\rangle \\ &\leq \|w_2 - w_1\|\|x\| \text{ (Using Cauchy-Schwartz Inequality)} \\ &\leq R\|w_2 - w_1\| \end{aligned} \tag{35}$$

Case 3:($y\langle w_1, x\rangle \geq 1, y\langle w_2, x\rangle < 1$)

$$\begin{aligned} \|l_1 - l_2\| &= -(l_1 - l_2) \\ &= -(0 - (1 - y\langle w_2, x\rangle)) \\ &= (1 - y\langle w_2, x\rangle) \\ &< (1 - y\langle w_2, x\rangle) - (1 - y\langle w_1, x\rangle) \\ &= y\langle w_1 - w_2, x\rangle \\ &\leq \|w_2 - w_1\|\|x\| \text{ (Using Cauchy-Schwartz Inequality)} \\ &\leq R\|w_2 - w_1\| \end{aligned} \tag{36}$$

Case 4:($y\langle w_1, x\rangle < 1, y\langle w_2, x\rangle < 1$)
Lets suppose $(1 - y\langle w_1, x\rangle) \geq (1 - y\langle w_2, x\rangle)$.

$$\begin{aligned} \|l_1 - l_2\| &= (l_1 - l_2) \\ &= (1 - y\langle w_1, x\rangle) - (1 - y\langle w_2, x\rangle) \\ &= y\langle w_2 - w_1, x\rangle \\ &\leq \|w_2 - w_1\|\|x\| \text{ (Using Cauchy-Schwartz Inequality)} \\ &\leq R\|w_2 - w_1\| \end{aligned} \tag{37}$$

The same is applicable when $(1 - y\langle w_1, x\rangle) \leq (1 - y\langle w_2, x\rangle)$. Therefore the function l is R-lipschitz.