CSE 512 – Homework II

Hang Zhao ID: 112524698

Due **April 20 (Sunday), 2020 [11:59 pm]**.

1 Theory

I: Gaussian distributions
A Gaussian distribution is defined by two parameters: a mean μ and a variance σ^2. For a scalar x, it is defined as

$$\mathcal{N}(x|\mu,\, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

Using this definition, we can find expectations of various functions of x. Prove the following:

$$E[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\, \sigma^2) x dx = \mu$$

$$E[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\, \sigma^2) x dx$$

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

Now we use t to substitute $\frac{x-\mu}{\sqrt{2}\sigma}$ and we can get:

$$= \frac{\sqrt{2}\sigma}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sqrt{2}\sigma t + \mu) e^{-t^2} dt$$

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

Since $\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$, we can get:

$$= \frac{1}{\sqrt{\pi}} (\sqrt{2}\mu [-\frac{1}{2} e^{-t^2}]_{-\infty}^{\infty} + \mu\sqrt{\pi})$$

For $exp\ x$ will approach 0 if x is approaching $-\infty$.

$$= \frac{\mu\sqrt{\pi}}{\sqrt{\pi}}$$

$$= \mu$$

Also, this is a valid probability distribution, and therefore

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\, \sigma^2) dx = 1$$

Differentiate the above w.r.t. σ^2 to show that

$$E[x^2] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\, \sigma^2)x^2 dx = \mu^2 + \sigma^2$$

$$E[x^2] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\, \sigma^2)x^2 dx$$

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

Now we use t to substitute $\frac{x-\mu}{\sqrt{2}\sigma}$ and we can get:

$$= \frac{\sqrt{2}\sigma}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sqrt{2}\sigma t + \mu)^2 e^{-t^2} dt$$

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

Since $\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$, we can get:

$$= \frac{1}{\sqrt{\pi}}([2\sigma^2(-\frac{1}{2}te^{-t^2} + \frac{\sqrt{\pi}}{4}erf(t)]_{-\infty}^{\infty} + 2\sqrt{2}\sigma\mu[-\frac{1}{2}e^{-t^2}]_{-\infty}^{\infty} + \mu^2\sqrt{\pi})$$

For $exp(x)$ will approach 0 if x is approaching $-\infty$ and ∞. Also, $x\ exp(-x^2)$ will approach 0 if x is approaching $-\infty$ and ∞. For $erf(x)$, when x is approaching ∞, this function will approach 1. When x is approaching $-\infty$, this function will approach to negative 1. Thus the difference between $-\infty$ and ∞ will be 2. And the expression above will become:

$$= \frac{1}{\sqrt{\pi}}(\sigma^2\sqrt{\pi} + \mu^2\sqrt{\pi})$$

$$= \sigma^2 + \mu^2$$

Using the above results, finally show hat the variance of a Gaussian distribution, defined as $var(x) = E[x^2] - E[x]^2$, is indeed σ^2.

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

II: Strongly convex function

Show that if f is a strongly convex function with paramere λ, then for every w, u and $v \in \partial f(w)$, the following holds:

$$< w - u, v > \geq f(w) - f(u) + \frac{\lambda}{2}\|w - u\|^2$$

According to the lemma 13.5 from the book, we can get:

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

And since v is a subgradient of f(w) we can have:

$$f(w + \alpha(u - w)) - f(w) \geq \alpha < u - w, v >$$

Put the two formula together we can get:

$$< u - w, v > \leq f(u) - f(w) - \frac{\lambda}{2}(1 - \alpha)\|w - u\|^2$$

When α is approaching 0, we can get the following:

$$< w - u, v > \geq f(w) - f(u) + \frac{\lambda}{2}\|w - u\|^2$$

Hence the proof holds.

III: Kernel construction

Suppose K_1 and K_2 are valid kernels on a domain X, i.e., their Gram matrix is symmetric and positive-definite. Show that the following are also, then, valid kernel functions:
(a) $K(u, v) = \alpha K_1(u, v) + \beta K_2(u, v)$, for any scalars $\alpha, \beta \geq 0$.
(b) $K(u, v) = K_1(u, v)K_2(u, v)$.
(a) Let ϕ_1 and ϕ_2 be the feature vectors of K_1 and K_2. The vector ϕ is the combination of $\sqrt{\alpha}\phi_1$ and $\sqrt{\beta}\phi_2$.

$$\phi(x) = [\sqrt{\alpha}\phi_1^1(x), \sqrt{\alpha}\phi_1^2(x), ..., \sqrt{\alpha}\phi_1^m(x), \sqrt{\beta}\phi_2^1(x), \sqrt{\beta}\phi_2^2(x), ..., \sqrt{\beta}\phi_2^m(x)]$$

$$< \phi(u), \phi(v) >= \sum_{i=1}^{m} \phi^i(u) \times \phi^i(v)$$

$$= \alpha < \phi_1(u), \phi_1(v) > +\beta < \phi_2(u), \phi_2(v) >$$

$$= \alpha K_1(u, v) + \beta K_2(u, v) = K(u, v)$$

So K is also a valid kernel function.

(b) We know that positive definite matrix is also covariance matrix. Given positive definite matrix K_1 and K_2, we can get that they are also covariance matrices. In order to prove K is a valid kernel function, we need to show that K is covariance matrix. Since covariance matrix is positive definite matrix, we can get K is a valid kernel function.
Define matrix for K_1 $(i_1, i_2, ..., i_n)$ as I ,matrix for K_2 $(j_1, j_2, ...j_n)$ $as J with average 0. Then the matrix for K will be (i_1 j$ as W.

$$cov(W)_{k,m} = E[(W_k - \mu)(W_m - \mu)] = E[W_k W_m]$$

$$= E[(I_k J_k)(I_m J_m)] = E[I_k I_m]E[J_k J_m] = K(u, v)$$

So K is also a covariance matrix which is symmetric and positive definite.

IV: Local minimum

Construct an example showing that the 0-1 loss function may suffer local minima. That is, construct a training sample $S \in (X \times \{\pm 1\})^m$ (for simplicity, you may assume that $XX = R^2$) $for which there exists a vector w and$ 0 such that

(1) For any w' such that $\left\|w - w'\right\| \leq \epsilon$, we have $L_S^{(01)}(w) \leq L_S^{(01)}(w')$, and
(2) There exists some w^* such that $L_S^{(01)}(w^*) < L_S^{(01)}(w)$.
The first condition formally shows that w is a local minimum, while the second condition shows that it is not a global minimum.
(1) We start to construct a homogenous halfspaces in R^d. Let x = e_1 and y = 1 to be a sample in S. Let $w = -e_1$. Then we can have $< w, x >= -1$ which means the loss of w is 1. Let $\epsilon \in (0, 1)$. For any w' such that $\left\|w - w'\right\| \leq \epsilon$, according to Cauchy-Schwartz inequality $| < u, v > | \leq \|u\|\|v\|$.

$$< w', x >=< w, x > + < w' - w, x >$$

$$= -1 + < w' - w, x >$$

$$\leq -1 + \left\|w' - w\right\|\|x\|$$

Since the difference of w' and w is at most 1 and x is an unit vector.

$$< -1 + 1 = 0$$

Such that we can get $< w', x >$ is less than 0 and the loss of w' is also 1. We can know that w is a local minima.
(2) There exists some w^*. The difference between w^* and w is bigger than 1 which might make the inner product of w^* and x bigger than 1 and thus the loss of w^* is 0. Thus total loss of w^* is less than the total less of w. And we can know that w is not a global minimum.

V: Learnability of logistic regression

Let $H = X = x \in R^d : \|x\| \leq B$ for some positive scalar B. Let the label set $Y = \{\pm 1\}$, and the loss function be defined as $l(w, (x, y)) = \log[1 + exp(-y < w, x >)]$. Note that this is the formal definition of the logistic regression problem. Show that this problem is both convex-Lipschitz-bounded, and convex-smooth-bounded. Specify the parameters for Lipschitzness and smoothness.

We can rewrite this loss function $l(w, (x, y)) = \log[1 + exp(-y < w, x >)]$ as :

$$f(x) = \log(1 + exp(x))$$

The derivative is:

$$f'(x) = \frac{exp(x)}{1 + exp(x)} = \frac{1}{exp(-x) + 1}$$

And second derivative is:

$$f(x)'' = \frac{exp(-x)}{(exp(-x) + 1)^2}$$

$$= (exp(x)(exp(-x) + 1)^2)^{-1}$$

$$= \frac{1}{2 + exp(x) + exp(-x)}$$

Notice this value is non-negative and combined with Claim 12.4 from the book we can conclude that function $l(w, (x, y)) = \log[1 + exp(-y < w, x >)]$ is convex.
Also, because $\frac{1}{exp(-x)+1}$ is less or equal to 1. We can know that function $l(w, (x, y)) = \log[1 + exp(-y < w, x >)]$ is 1-Lipschitz. According to Claim 12.7, we can conclude l is B-Lipschitz.
Next, we take a look at second derivative and we find out that this value $\frac{1}{2+exp(x)+exp(-x)}$ is less or

equal to $\frac{1}{4}$ when both $exp(-x)$ and $exp(x)$ are equal to 1. We can say that $f(x)'$ is 1/4-LIpschitz. Combined with Claim 12.9, we can conclude that l is $B^2/4$-smooth.
Since each function is bounded by B, we can conclude that this problem is both convex-Lipschitz-bounded, and convex-smooth-bounded.

VI: Learnability of Halfspaces with hinge loss

Consider the bounded domain $X = x : \|w\| \leq R$ with the label set $Y = \{\pm 1\}$, and the loss function defined as $l(w, (x, y)) = max\{0, 1 - y < w, x >\}$. Note that this is the formal definition of halfspace learning with hinge loss. The formulation already covers boundedness. Show that this learning problem is also convex and R-Lipschitz.

Since $1 - y < w, x >$ and 0 are both linear function, we know that all the linear functions are convex and the maximum of two convex functions is convex. So we can say that this hinge loss function is also convex. Consider some $(x, y) \in x \in R^d : \|x\| \leq R \times -1, 1$. Let w_1, $w_2 \in R^d$. For $l = max\{0, 1 - y < w_i, x >\}$. We want to show $|l_1 - l_2| \leq R\|w_1 - w_2\|$. If both $y < w_1, x >$ and $y < w_2, x >$ are greater or equal to 1. Then the loss of l_1 and l_2 are equal to 0. It must be less than $R\|w_1 - w_2\|$. If we can find out at least one w such that $y < w_1, x >$ less than 1. So there must exist w_2 such that $1 - y < w_1, x > \geq 1 - y < w_2, x >$.

$$
\begin{aligned}
|l_1 - l_2| &= l_1 - l_2 \\
&= 1 - y < w_1, x > - max\{0, 1 - y < w_2, x >\} \\
&\leq 1 - y < w_1, x > - (1 - y < w_2, x >) \\
&= y < w_2 - w_1, x > \\
&\leq \|w_1 - w_2\|\|x\| \\
&\leq R\|w_1 - w_2\|
\end{aligned}
$$