CSE 512 – Homework II

By Yasha Singh - 112970310

April 20th 2020

1 Theory Questions

1.1 Gaussian distributions

Part 1. We have,

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

To prove: -

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

So,

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

$$E[X] = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx$$

Let us take, $t = \frac{x-\mu}{\sqrt{2}\sigma}$

$$= \frac{\sqrt{2}\sigma}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sqrt{2}\sigma t + \mu) \exp\left(-t^2\right) \mathrm{d}t$$

$$\begin{aligned}
&= \frac{1}{\sqrt{\pi}}\left(\sqrt{2}\sigma \int_{-\infty}^{\infty} t \exp\left(-t^2\right) \mathrm{d}t + \mu \int_{-\infty}^{\infty} \exp\left(-t^2\right) \mathrm{d}t\right) \\
&= \frac{1}{\sqrt{\pi}}\left(\sqrt{2}\sigma \left[-\frac{1}{2}\exp\left(-t^2\right)\right]_{-\infty}^{\infty} + \mu\sqrt{\pi}\right) \\
&= \frac{\mu\sqrt{\pi}}{\sqrt{\pi}} \\
&= \mu
\end{aligned}$$

Part 2.
Given to us,

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

To prove:-

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

Proof:-

$$\begin{aligned}
\frac{dN}{d\sigma^2} &= \frac{\frac{dN}{d\sigma}}{\frac{d\sigma^2}{d\sigma}} \\
\frac{dN}{d\sigma} &= \frac{1}{\sqrt{2\pi}}\left(e^{-\frac{1}{2\sigma 2}(x-\mu)^2}\left(\frac{-1}{\sigma^2}\right)+\frac{1}{\sigma}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}. - \frac{1}{2}\cdot\frac{-2}{\sigma^3}.(x-\mu)^2\right) \\
\frac{dN}{d\sigma} &= -\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\left(\frac{1}{\sigma^2}+\frac{-(x-\mu)^2}{\sigma\mu}\right) \\
\frac{d\sigma^2}{d\sigma} &= 2\sigma
\end{aligned}$$

$$\int_{-\infty}^{\infty} -N\left(\frac{1}{2\sigma^2}+\left(\frac{-(x-\mu)^2}{2\sigma^4}\right)\right)dx = 0$$

$$\begin{aligned}
\int_{-\infty}^{\infty} N\left(\frac{(x-\mu)^2}{2\sigma^4}-\frac{1}{2\sigma^2}\right)dx = 0 \\
\frac{1}{2\sigma^2}\int_{-\infty}^{\infty} N\left((x-\mu)^2-\sigma^2\right)dx = 0
\end{aligned}$$

$$\begin{aligned}
&\int_{-\infty}^{\infty} N\left(x^2+\mu^2-2x\mu-\sigma^2\right)dx = 0 \\
&\int_{-\infty}^{\infty} Nx^2dx + \int_{-\infty}^{\infty}\mu^2Ndx - \int_{-\infty}^{\infty}(2\pi x\mu)dx - \int_{-\infty}^{\infty}N\sigma^2dx \quad = 0
\end{aligned}$$

$$E\left[x^2\right]+\mu^2-E[x]\cdot 2\mu-\sigma^2=0$$

$$E\left[x^2\right]+\mu^2-2\mu^2-\sigma^2=0$$

$$E\left[x^2\right]=\mu^2+\sigma^2$$

Part 3.
From part 1, we know:-

$$E[x]^2=\mu^2$$

And from part 2, we know:-

$$E[x^2]=\mu^2+\sigma^2$$

And formula for variance is $\operatorname{var}(x)=E\left[x^2\right]-E[x]^2$
Substituting values, we get

$$\operatorname{var}(x)=\mu^2+\sigma^2-\mu^2$$

Therefore,

$$\operatorname{var}(x)=\sigma^2$$

1.2 Strongly convex function

Given f is a strongly convex function with parameter λ, then for every w, u, and $v \; \epsilon \; \delta f(w)$
To show:-

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

We start with the definition of strong convexity which states:-

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

Now, dividing the above equation by alpha we get:-

$$\begin{aligned}\frac{f(\mathbf{w}+\alpha(\mathbf{u}-\mathbf{w}))-f(\mathbf{w})}{\alpha} &\leq f(\mathbf{u})-(\frac{1-\alpha}{\alpha})f(\mathbf{w})-\frac{\lambda}{2}(1-\alpha)\|\mathbf{w}-\mathbf{u}\|^2\\ \frac{f(\mathbf{w}+\alpha(\mathbf{u}-\mathbf{w}))-f(\mathbf{w})}{\alpha} &\leq f(\mathbf{u})+(\frac{1}{\alpha})f(\mathbf{w})-f(\mathbf{w})-\frac{\lambda}{2}(1-\alpha)\|\mathbf{w}-\mathbf{u}\|^2\\ \frac{f(\mathbf{w}+\alpha(\mathbf{u}-\mathbf{w}))-f(\mathbf{w})}{\alpha}-(\frac{1}{\alpha})f(\mathbf{w}) &\leq f(\mathbf{u})-f(\mathbf{w})-\frac{\lambda}{2}(1-\alpha)\|\mathbf{w}-\mathbf{u}\|^2\\ \frac{f(\alpha\mathbf{u}+(1-\alpha)\mathbf{w})-f(\mathbf{w})}{\alpha} &\leq f(\mathbf{u})-f(\mathbf{w})-\frac{\lambda}{2}(1-\alpha)\|\mathbf{w}-\mathbf{u}\|^2\\ \frac{f(\mathbf{w}+\alpha(\mathbf{u}-\mathbf{w}))-f(\mathbf{w})}{\alpha} &\leq f(\mathbf{u})-f(\mathbf{w})-\frac{\lambda}{2}(1-\alpha)\|\mathbf{w}-\mathbf{u}\|^2\end{aligned}$$

We know from problem statement that $v \ \epsilon \ \delta f(w)$ is basically the a subgradient of f at w, so,

$$f(\mathbf{w}+\alpha(\mathbf{u}-\mathbf{w}))-f(\mathbf{w}) \geq \alpha\langle\mathbf{u}-\mathbf{w},\mathbf{v}\rangle$$

From the results of above two equations we have that,

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

When the value of α will tend to 0, $(1-\alpha) \rightarrow 1$ so RHS will reduce to

$$f(\mathbf{w})-f(\mathbf{u})-\frac{\lambda}{2}\|\mathbf{w}-\mathbf{u}\|^2$$

. This is our required statement to be proved.
The LHS will be equivalent to derivative of $g(\alpha)$ when $\alpha = 0$, so $g(\alpha) = f(u+\alpha(w-u))$. Now if u is the minima for f then α will be the point where minimum for g is obtained.

1.3 Kernel construction

a.

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

for any scalars $\alpha, \beta \geq 0$
We know , $K(u,v) = \langle\Phi(u),\Phi(v)\rangle_{\mathcal{H}}$ where, $\Phi : \mathcal{X} \rightarrow \mathcal{H}$ in the Hilbert space H. So We can write, $\alpha K_1(u,v) = \langle\sqrt{\alpha}\Phi_1(u), \sqrt{\alpha}\Phi_1(v)\rangle$
and $\beta k_2(u,v) = \left\langle\sqrt{\beta}\Phi_2(u), \sqrt{\beta}\Phi_2(v)\right\rangle$
Substituting these values in the above equation,

$$\begin{aligned}K(\mathbf{u},\mathbf{v}) &= \alpha K_1(\mathbf{u},\mathbf{v}) + \beta K_2(\mathbf{u},\mathbf{v})\\ &= \left\langle\sqrt{\alpha}\Phi_1(u), \sqrt{\alpha}\Phi_1(v)\right\rangle + \left\langle\sqrt{\beta}\Phi_2(u), \sqrt{\beta}\Phi_2(v)\right\rangle\\ &= \left\langle\left[\sqrt{\alpha}\Phi_1(u)\sqrt{\beta}\Phi_2(u)\right], \left[\sqrt{\alpha}\Phi_1(v)\sqrt{\beta}\Phi_2(v)\right]\right\rangle\end{aligned}$$

From the above we can see that the linear combination of two positive-definite kernels K1 and K2 can be expressed as an inner product when their Gram matrix are symmetric.

b. The product of the two kernels K1 and K2 which are positive semi-definite matrices is essentially the Cartesian products of their features. It is because the eigenvalues of the product are all equal to the pairs of products of the eigenvalues of the two components. This is shown as below :-

$$K(\mathbf{u},\mathbf{v}) = K_1(\mathbf{u},\mathbf{v})K_2(\mathbf{u},\mathbf{v})$$

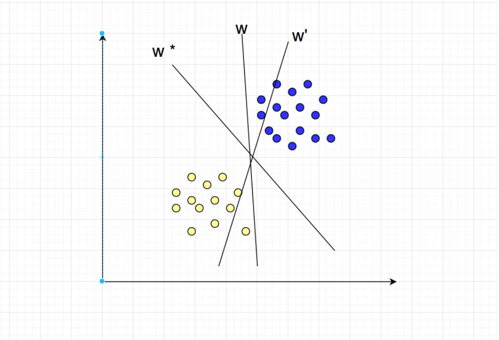
Upon substituting values for $K_1(\mathbf{u},\mathbf{v})$ and $K_2(\mathbf{u},\mathbf{v})$, we will get the equation for $K(\mathbf{u},\mathbf{v})$ as below :-

$$
\begin{aligned}
&= \sum_{p=1}^{n} \Phi_{1p}(u)\Phi_{1p}\left(v\right) \sum_{q=1}^{m} \Phi_{2q}(u)\Phi_{2q}\left(v\right) \\
&= \sum_{p=1}^{n}\sum_{q=1}^{m} \left(\Phi_{1p}(u)\Phi_{2q}(u)\right)\left(\Phi_{1p}\left(v\right)\Phi_{2q}\left(v\right)\right) \\
&= \sum_{x=1}^{nm} \Phi_{12x}(u)\Phi_{12x}\left(v\right) \\
&= \Phi_{12}(u)^T\Phi_{12}\left(v\right)
\end{aligned}
$$

So, K is essentially the covariance matrix which is therefore symmetric and positive definite.

1.4 Local minimum

Let us assume that our data is present in 2-D space which is linearly seperable as shown in figure below such the training sample $S \in (\mathcal{X} \times \{\pm 1\})^m$ and $X\mathcal{X} = R^2$.



Suppose that our hypothesis H generates a decision boundary $(\mathbf{w})$ which divides this data sample into two regions. We stop training when the loss becomes zero on training data.

While using 0-1 Loss function we may run into 2 scenarios because 0-1 Loss is not differentiable :

1. Once we get a decision boundary$(\mathbf{w})$ which is able to separate the training data, then in case 0-1 Loss function we stop updating the weights and we have obtained a local minimum. It is definitely better than a decision boundary$(\mathbf{w}')$ which is misclassifying few of the data points. This totally depends upon which data points our learning algorithm has seen so far. Thus satisfying the condition, $\|\mathbf{w} - \mathbf{w}'\| \leq \epsilon,$ we have $L_S^{(01)}(\mathbf{w}) \leq L_S^{(01)}\left(\mathbf{w}'\right)$.

2. However, this decision boundary which we obtained in case 1 might not be the most optimum one as shown in the figure. That is, when subjected to test data points it might not be able to separate them. So, it is not the global minima$(\mathbf{w}^*)$. In case of 0-1 Loss function weight updation stops as soon as we get a decision boundary that gave 0 error on the training data. Therefore, it is the condition of $L_S^{(01)}(\mathbf{w}^*) \leq L_S^{(01)}\left(\mathbf{w}\right)$.

1.5 Learnability of logistic regression

Given to us, is a hypothesis H, such that $\mathcal{H} = \mathcal{X} = \left\{\mathbf{x} \in R^d : \|\mathbf{x}\| \leq B\right\}$ and labels $\mathcal{Y} = \{\pm 1\}$ and loss function is $\ell(\mathbf{w},(\mathbf{x},y)) = \log[1 + \exp(-y\langle\mathbf{w},\mathbf{x}\rangle)$
For Convexity :-

We know, for a function g which has a $g : R \to R$ and function f which is a $f : R^d \to R$ and f can be expressed in terms g as $f(\mathbf{w}) = g(\langle\mathbf{w},\mathbf{x}\rangle + y)$ when $\mathbf{x} \in R^d, y \in R$. Then convexity of g will imply convexity of f. In our case, f here will be the loss function $\ell(\mathbf{w},(\mathbf{x},y)) = \log[1+\exp(-y\langle\mathbf{w},\mathbf{x}\rangle)$ and g is $g(a) = log(1+exp(a))$. Here g is is convex because double differentiation of $g(a)$ is non negative.

For Lipschitzness :-

For $g(a) = log(1+exp(a))$,
$|g'(a)| = \frac{\exp(a)}{1+\exp(a)}$
$|g'(a)| = \frac{1}{\exp(-a)+1} \leq 1$
$|g'(a)| \leq 1$
If $f(\mathbf{x}) = g_1(g_2(\mathbf{x}))$, and g_1 is ρ_1 Lipschitz and g_2 is ρ_2 Lipschitz then f will be $(\rho_1\rho_2)-$Lipschitz. Following this reference, l is B-Lipschitz.

For smoothness :-

We know that if $f(\mathbf{w}) = g(\langle\mathbf{w},\mathbf{x}\rangle + b)$ and g is $\beta-$ smooth then f will be $\left(\beta\|\mathbf{x}\|^2\right) -$ smooth
We find smoothness of g as below :-

$$\begin{aligned}
g''(a) &= \frac{\exp(-a)}{(1+\exp(-a))^2} \\
&= \left(\exp(a)(1+\exp(-a))^2\right)^{-1} \\
&= \frac{1}{\exp(a)+\exp(-a)+2} \\
g''(a) &\leq 1/4
\end{aligned}$$

So g' is 1/4 Lipschitz. So following this, our loss function l is $B^2/4$ smooth.

For Boundness,

According to the assumptions $\mathcal{H} = \mathcal{X} = \left\{\mathbf{x} \in R^d : \|\mathbf{x}\| \leq B\right\}$ given in the question, norm of every hypothesis is bounded by B.

Thus, we can state that the learning problem of logistic regression is Convex- Lipschitz-Bounded with parameters B, B and Convex-Smooth-Bounded with parameters $B_2/4$, B.

1.6 Learnability of Halfspaces with hinge loss

For Convexity :-

Hinge loss is given by $\ell(\mathbf{w},(\mathbf{x},y)) = \max\{0, 1-y\langle\mathbf{w},\mathbf{x}\rangle\}$. In order to tell that it is convex, firstly, we know that $\ell^{0-1}(\mathbf{w},(\mathbf{x},y)) \leq \ell^{\text{hinge}}(\mathbf{w},(\mathbf{x},y))$ that is, Hinge loss replaces the real loss by its upper bound. Secondly, linear function is convex and also the maximum of two convex functions is convex. Therefore, we can say that Hinge Loss is convex.

For R-Lipschitzness:-

Given to us is a bounded domain $\mathcal{X} = \{\mathbf{x} : \|\mathbf{w}\| \leq R\}$. To determine the Lipschitzness of our loss function, let us consider two hinge loss functions l_1 and l_2 given by $\ell_{1,2}(\mathbf{w},(\mathbf{x},y)) = \max\{0, 1-y\langle\mathbf{w},\mathbf{x}\rangle\}$.

$$|\ell_1 - \ell_2| = \ell_1 - \ell_2$$

Let us assume that $1 - y\langle\mathbf{w}1,\mathbf{x}\rangle \geq 1 - y\langle\mathbf{w}2,\mathbf{x}\rangle$ is true, then upon substituting the values we get,

$$
\begin{aligned}
|\ell_1 - \ell_2| &= 1 - y\,\langle \mathbf{w}_1, \mathbf{x} \rangle - \max\{0, 1 - y\,\langle \mathbf{w}_2, \mathbf{x} \rangle\} \\
&\leq 1 - y\,\langle \mathbf{w}_1, \mathbf{x} \rangle - (1 - y\,\langle \mathbf{w}_2, \mathbf{x} \rangle) \\
|\ell_1 - \ell_2| &= y\,\langle \mathbf{w}_2 - \mathbf{w}_1, \mathbf{x} \rangle \\
|\ell_1 - \ell_2| &\leq \|\mathbf{w}_1 - \mathbf{w}_2\|\,\|\mathbf{x}\|
\end{aligned}
$$

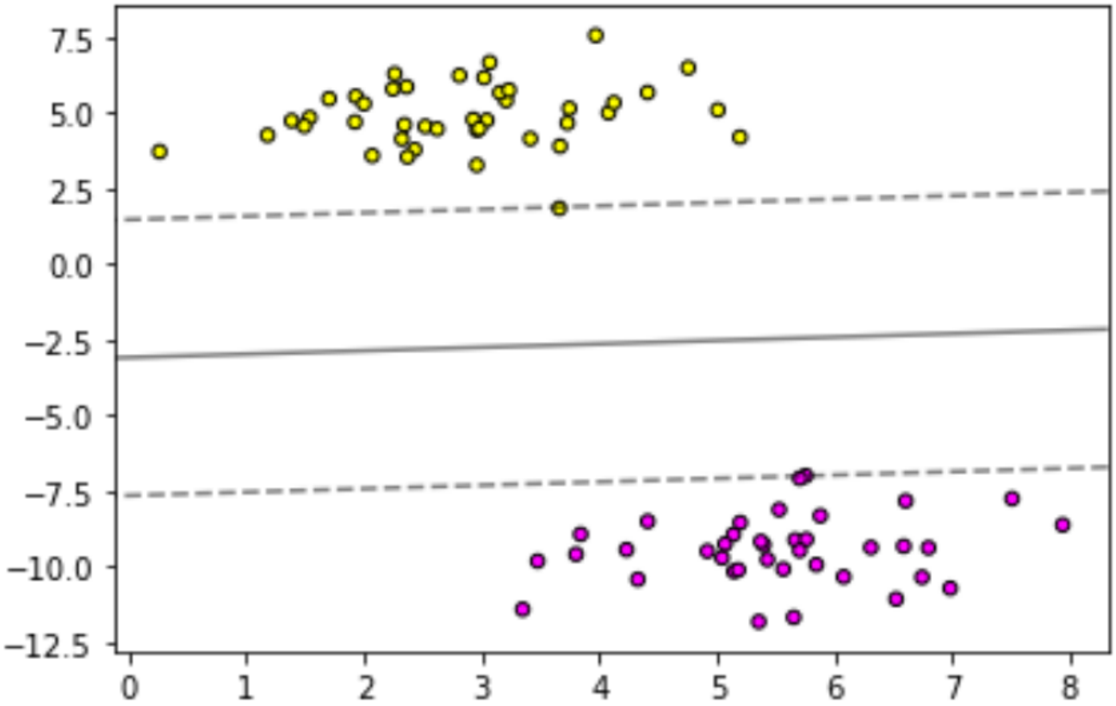
From the question statement, as $||\mathbf{x}\| \leq R$ so we replace it by R

$$|\ell_1 - \ell_2| \leq R\,\|\mathbf{w}_1 - \mathbf{w}_2\|$$

In the case when values are classified correctly w.r.t the labels i.e. $y * \langle \mathbf{w1}, \mathbf{x} \rangle$ and $y * \langle \mathbf{w2}, \mathbf{x} \rangle \geq 1$, then the difference of loss function will be 0, and as $0 \leq R$, thus hinge loss is R-Lipschitz.

1.7 Programming

Visualization of maximum-margin hyperplane:-



7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−12.5
0
1
2
3
4
5
6
7
8