Assignment 2

Manasa Navalgund(112871754)

April 20, 2020

1 Theory

Q1.

$$
\begin{aligned}
\textit{Given } N(x|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} exp\{\frac{-1}{2\sigma^2}(x-\mu)^2)\} \\
E[x] &= \int_{-\infty}^{\infty} N(x|\mu,\sigma^2)xdx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} exp\{\frac{-1}{2\sigma^2}(x-\mu)^2)\}xdx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} exp\{\frac{-1}{2\sigma^2}(x-\mu)^2)\}xdx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} exp\left(-\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)^2\right)xdx
\end{aligned}
$$

$$
\begin{aligned}
\textit{Substituting } t &= \frac{x-\mu}{\sqrt{2}\sigma} \\
= dt &= \frac{dx}{\sqrt{2}\sigma} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} exp\left(-t^2\right)(\sqrt{2}\sigma t+\mu)(\sqrt{2}\sigma dt) \\
&= \frac{\sqrt{2}\sigma}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} exp\left(-t^2\right)(\sqrt{2}\sigma t+\mu)dt \\
&= \frac{1}{\sqrt{\pi}}\left(\int_{-\infty}^{\infty} \sqrt{2}\sigma t \exp\left(-t^2\right)dt + \int_{-\infty}^{\infty} exp\left(-t^2\right)\mu dt\right)
\end{aligned}
$$

From Gaussian integral, [1]

$$\int_{-\infty}^{\infty} \exp\left(-t^2\right)dt = \sqrt{\pi}$$

Substituting this in above equation,

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

$$\frac{d}{dt} exp(-t^2) = -2texp(-t^2)$$

Susbtituting this in above,

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

Therefore $E[x] = \mu$
Hence proved.

We know that,

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

And since this is a valid probability distribution,

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

Differentiating this w.r.t σ^2

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

Using Lebiniz theorem, we can interchange integration and differentiation[3],

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

Substituting $\sigma^2 = t$

$$\int_{-\infty}^{\infty} \frac{d}{dt}\left(\frac{1}{\sqrt{2\pi t}} exp\{\frac{-1}{2t}(x-\mu)^2)\}\right) dx = \frac{d}{dt}(1)$$

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

Substituting t = σ^2

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\left(\frac{1}{\sqrt{\sigma^2}} exp\{\frac{-1}{2\sigma^2}(x-\mu)^2)\}\left(\frac{-(x-\mu)^2}{2}\left(\frac{-1}{\sigma^4}\right)\right) + exp\{\frac{1}{2\sigma^2}(x-\mu)^2)\}\left(\frac{-1}{2\sigma^3}\right)\right) dx = 0$$

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

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

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

$$\int_{-\infty}^{\infty} N(x^2)dx + \int_{-\infty}^{\infty} N(\mu^2)dx - \int_{-\infty}^{\infty} N(2x\mu)dx - \int_{-\infty}^{\infty} N(\sigma^2)dx = 0$$

Since μ and σ are constants w.r.t x and $\int_{-\infty}^{\infty} N dx = 1$ and above result,

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

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

Hence proved.

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

Q2.
Given: f is a strongly-convex function with parameter λ
For a strongly-convex function, we have, DEFINITION 13.4, UML
$f(\alpha u + (1-\alpha)w) \leq \alpha f(u) + (1-\alpha)f(w) - \frac{\lambda}{2}\alpha(1-\alpha)||u-w||^2$
Dividing the definition by α, we get,

$$\frac{f(w+\alpha(u-w)) - f(w)}{\alpha} \leq f(u) - f(w) - \frac{\lambda}{2}(1-\alpha)||w-u||^2 \qquad (3)$$

We know from LEMMA 14.3 and DEFINITION 14.4 of UML that a subgradient v, satisfies following equation,
$f(u) - f(w) \geq < u-w, v >$
Substituting $u = w + \alpha(u-w)$,
$f(w+\alpha(u-w)) - f(w) \geq < w+\alpha(u-w) - w, v >$
$f(w+\alpha(u-w)) - f(w) \geq \alpha < (u-w), v >$
Substituting this in (3) above,
$< (u-w), v > \leq f(u) - f(w) - \frac{\lambda}{2}(1-\alpha)||w-u||^2$
Multiplying both sides by -1, we get:
$\implies < (w-u), v > \geq f(w) - f(u) + \frac{\lambda}{2}(1-\alpha)||w-u||^2$
For $\alpha -> 0$, Since u is minimizer of f, it follows $\alpha = 0$ is minimizer
$\implies < (w-u), v > \geq f(w) - f(u) + \frac{\lambda}{2}||w-u||^2$

Q3.
Given: K1 and K2 are valid kernels over the domain of X.
a. $K(u,v) = \alpha K1(u,v) + \beta K2(u,v)$
αK1(u,v) can be representd as : $< \sqrt{\alpha}\Psi_1(u), \sqrt{\beta}\Psi_1(v) >$
$\beta(u,v)$ can be representd as : $< \sqrt{\beta}\Psi_2(u), \sqrt{\beta}\Psi_2(v) >$

$$\begin{aligned} Now, K(u,v) &= \alpha K1(u,v) + \beta K2(u,v) \\ &=< \sqrt{\alpha}\Psi_1(u), \sqrt{\alpha}\Psi_1(v) > + < \sqrt{\beta}\Psi_2(u), \sqrt{\beta}\Psi_2(v) > \\ &=< \sqrt{\alpha}\Psi_1(u)\sqrt{\beta}\Psi_2(u), \sqrt{\alpha}\Psi_1(v)\sqrt{\beta}\Psi_2(v) > \end{aligned}$$

Since, α, β are positive, $\sqrt{\alpha} and \sqrt{\beta}$ are positive scalars and we know Ψ_1, Ψ_2 are valid kernels,
Therefore this shows above inner product is a valid kernel

b. K(u,v) = K1(u,v)K2(u,v)
The gram matrix of this product is given as $K = k1 \cdot K2$
where $\cdot$ represents Hadamard product.
Since K1 and K2 are positive and symmetric positive definite matrices. When we perform hadamard product on eigen decompositions, it gives a symmetric positive definite matrix as result. Since, this matrix suffices conditions for gram matrix of some two arguments, it is valid kernel function.

Q4.
Consider an example Halfspaces with loss function as $l = max\{0, 1 - < w, x >\}$
Consider the following training data set,

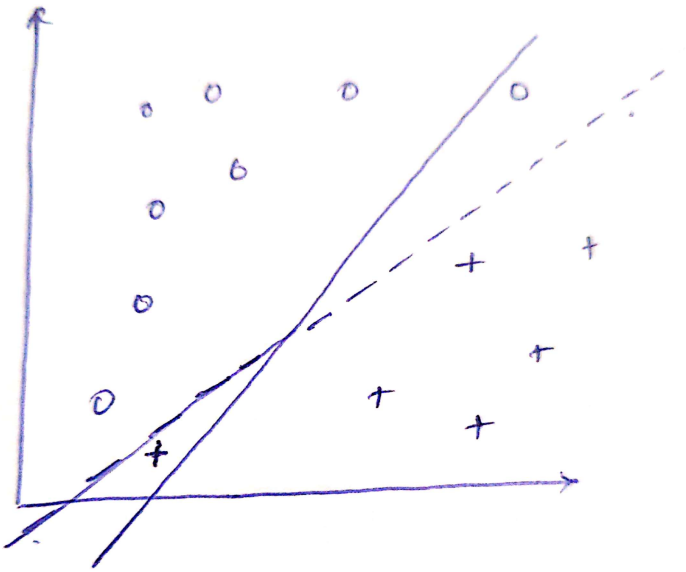

Figure 1: Training data

Here, with very small step size, the algoritm reaches till solid line and number of epochs end before line converges to dashed line. Since this has minimum loss till that point, algorithm returns dashed line as best hyperplane seperating the data.
However, with more epochs or larger step size, hyperplane could have converged to dashed line which classifies the data perfectly. This means solid line is global minima but algorithm converged at solid line which is local minima.

Therefore, halfspace class of functions can suffer from local minima

Q5

Convex-Lipschitz bounds

Given function $l = g = log(1 + exp(-y < w, x >))$

$g^{'}(x) = \left| \frac{exp(-y<w,x>)}{1+exp(-y<w,x>)} \right|$

$= \left| \frac{1}{1+exp(y<w,x>)} \right|$

Here, we can see that $g^{'}(x)$ is monotonically increasing function i.e., $g^{''}$ is non negative. Therefore, l is a convex function and given that $||x|| \leq$ B

We know from DEFINITION 12.6 UML, that Lipschitzness is defined as:
$||f(w_1) - f(w_2)|| \leq \rho||w_1 - w_2||$

By mean value theorem,
$||f(w_1) - f(w_2)|| \leq f^{'}(u)||w_1 - w_2||$

Given loss function is $g = log(1 + exp(-y < w, x >))$

$g^{'}(x) = \left| \frac{exp(-y<w,x>)}{1+exp(-y<w,x>)} \right|$

$= \left| \frac{1}{1+exp(y<w,x>)} \right| \leq 1$

Therefore, l is 1-Lipschitz.
But from claim 12.7, Lipschitzness for l is $\rho_1||w|| = B$
l is B-Lipschitz
Therefore, l is convex-Lipschitz bounded with parameters B, ρ=B

Convex-Smooth-bounded

Given function $l = g = log(1 + exp(-y < w, x >))$
From the definition of convex-smooth-bounded, we know that a learning problem is convex-smooth-bounded if,
1. For all w $\in$ H, $||w|| \leq B$
2. loss function is convex, non-negative and β-smooth

$g^{'}(x) = \left| \frac{exp(-y<w,x>)}{1+exp(-y<w,x>)} \right|$

$= \left| \frac{1}{1+exp(y<w,x>)} \right|$

Here, we can see that $g^{'}(x)$ is monotonically increasing function i.e., $g^{''}$ is non negative. Therefore, l is a convex function and given that $||x|| \leq$ B

loss function = $g = log(1 + exp(-y < w, x >))$
For a function to be β-smooth, its gradient has to be β-Lipschitz (From DEFINITION 12.8 UML),

$g^{'}(x) = \left| \frac{exp(-y<w,x>)}{1+exp(-y<w,x>)} \right|$

$g^{''}(x) = \left| \frac{exp(-y<w,x>)}{(1+exp(-y<w,x>))^2} \right|$

$g^{''}(x) = \left| \frac{1}{(1+exp(-y<w,x>))(1+exp(y<w,x>))} \right| \leq \frac{1}{4}$

Therefore, gradient is $\frac{1}{4}$-Lipschitz and loss function is $\frac{1}{4}$-smooth.
But from claim 12.9, smoothness for l is $\beta||x||^2 = \frac{B^2}{4}$.
l is $\frac{B^2}{4}$-smooth There we have proven convex-smooth-bounded with parameters B, $\frac{B^2}{4}$.

Q6. Given $l(w,(x,y)) = max0, 1 - y < w, x >$
Proof of convexity:
Proving l to be R-Lipschitz:
We know that Lipschitzness is given by :
$||l(w_1) - l(w_2)|| \leq \rho||w_1 - w_2||$
Let's consider a point (x,y) such that:
$l_{w1} = max\{0, 1 - y < w1, x >\}$
$l_{w2} = max\{0, 1 - y < w2, x >\}$
There can be multiple cases:

Case 1: $l_{w1} = l_{w2} = 0$ In this case, from definition of Lipschitzness, we have,
$|l_{w1} - l_{w2}| = 0 \leq R|w1 - w2|$ Since $R \geq 0$ and $|w1 - w2| \geq 0$
Case 2: When atleast one of $l_{w1} or l_{w2}$ is not 0
$|l_{w1} - l_{w2}| = |(1 - max\{0, 1 - y < w1, x >\}) - (1 - max\{0, 1 - y < w2, x >\})|$
Lets assume that $l_{w1} \geq l_{w2}$
$|l_{w1} - l_{w2}| = l_{w1} - l_{w2}$
$= (1 - y < w1, x >) - (max\{0, 1 - y < w2, x >\})$
$\leq (1 - y < w1, x >) - (1 - y < w2, x >) = y < w2 - w1, x >$
$\leq ||w2 - w1||||x||$
$||w1 - w2||$
Similarly if we assume $l_{w2} \geq l_{w1}$,
$|l_{w1} - l_{w2}| = l_{w2} - l_{w1}$
$= (1 - y < w2, x >) - (max\{0, 1 - y < w1, x >\})$
$\leq (1 - y < w2, x >) - (1 - y < w1, x >) = y < w1 - w2, x >$
$\leq ||w1 - w2||||x||$
$||w1 - w2||$
Therefore, $l(w,(x,y)) = max0, 1 - y < w, x >$ is R-Lipschitz

2 Programming

I get the following plot for the data:

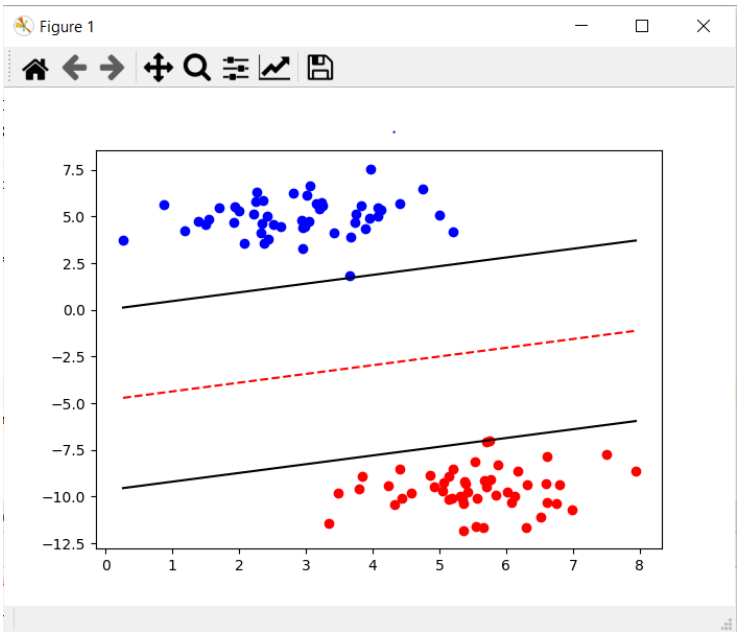

Figure 2: Training data

Output is redirected to file 'ReadMe.txt'
Output of above run:
the total number of data points on which your test method was run, : 20
the total number of data points on which your test method was misclassified, : 0

References

1. https://en.wikipedia.org/wiki/Gaussian_integral
2. Understanding Machine Learning: From Theory to Algorithms
3. https://math.stackexchange.com/questions/2530213/when-can-we-interchange-integration-and-differentiation