HW - 3 EE - 660

Q1. (a) $\theta(s) = \dfrac{e^s}{1+e^s}$ logistic Function θ

$$\tanh(s) = \frac{e^s - e^{-s}}{e^s + e^{-s}}$$

$$= \frac{e^s - 1/e^s}{e^s + 1/e^s}$$

$$= \frac{e^{2s} - 1}{e^{2s} + 1}$$

Now,

$$\theta(s) + \theta(s)\, e^s = e^s$$

$$\therefore \quad e^s\,(1 - \theta(s)) = \theta(s)$$

$$\therefore \quad e^s = \frac{\theta(s)}{1 - \theta(s)}$$

Substituting,

$$\tanh(s) = \frac{(e^s)^2 - 1}{(e^s)^2 + 1} = \frac{\left[\frac{\theta(s)}{1-\theta(s)}\right]^2 - 1}{\left[\frac{\theta(s)}{1-\theta(s)}\right]^2 + 1}$$

$$= \frac{(\theta(s))^2 - (1-\theta(s))^2}{(\theta(s))^2 + (1-\theta(s))^2}$$

$$= \frac{2\theta(s) - 1}{2\theta(s)^2 - 2\theta(s) + 1}$$

So,

$\tanh(s)$ is a scaled and shifted version of the sigmoid logistic function $\theta(s)$.

(b) $s \to +\infty$

$$\theta(s) = \frac{e^s}{1+e^s} = \frac{\infty}{\infty} = \text{Not Defined}$$

$$\tanh(s) = \frac{e^{2s}-1}{e^{2s}+1} = \frac{\infty}{\infty} = \text{Not Defined}$$

$s = 0$

$$\theta(s) = \frac{e^s}{1+e^s} = \frac{1}{2}$$

$$\tanh(s) = \frac{e^{2s}-1}{e^{2s}+1} = 0$$

$s \to -\infty$

$$\theta(s) = \frac{e^s}{1+e^s} = 0$$

$$\tanh(s) = \frac{e^{2s}-1}{e^{2s}+1} = \frac{0-1}{0+1} = -1$$

(c) $$1-\theta(s) = 1 - \frac{e^s}{1+e^s} = \frac{1+e^s-e^s}{1+e^s}$$

$$= \frac{1}{1+e^s}$$

$$\theta(-s) = \frac{e^{-s}}{1+e^{-s}} = \frac{1/e^s}{1+1/e^s} = \frac{1}{1+e^s}$$

So, $\boxed{1-\theta(s) = \theta(-s)}$

Q2. (a)

$$\underset{w}{\operatorname{argmin}} \; -\frac{1}{N} \ln \left(\prod_{n=1}^{N} P(y_n \mid x_n) \right)$$

$$= \underset{w}{\operatorname{argmin}} \; +\frac{1}{N} \sum_{n=1}^{N} \ln \left(\frac{1}{P(y_n \mid x_n)} \right)$$

$$= \underset{w}{\operatorname{argmin}} \; \frac{1}{N} \sum_{n=1}^{N} \ln \left(\frac{1}{\theta(y_n w^T x_n)} \right)$$

$$= \underset{w}{\operatorname{argmin}} \; \frac{1}{N} \sum_{n=1}^{N} \ln \left(\frac{1 + e^{y_n w^T x_n}}{e^{y_n w^T x_n}} \right)$$

$$= \underset{w}{\operatorname{argmin}} \; \frac{1}{N} \sum_{n=1}^{N} \ln \left(1 + e^{-y_n w^T x_n} \right)$$

So,

$$E_{in}(w) = \frac{1}{N} \sum_{n=1}^{N} \ln \left(1 + e^{-y_n w^T x_n} \right)$$

(b)

(i) For n being one data point,

$$E_{in}(w) = \ln \left(1 + e^{-y_n w^T x_n} \right)$$

For $y_n = +1$ in above,

$$E_{in}(w) = \ln \left(1 + e^{-w^T x_n} \right)$$

Here, when, $w^T x_n = 0$ $\quad E_{in}(w) = \ln(2)$

when, $w^T x_n \to \infty$ $\quad E_{in}(w) \approx \ln(1) \approx 0$

when, $w^T x_n \to -\infty$ $\quad E_{in}(w) \approx \ln(\infty) \approx \infty$

So, when $w^T x_n \to -\infty$, $E_{in} \approx \infty$ (maximum)

(ii) For m being one data point, and $y_m = -1$,

$$E_{in}(\underline{w}) = \ln(1 + e^{\underline{w}^T \underline{x}_m})$$

Here, when,

$$\underline{w}^T \underline{x}_m = 0 \qquad E_{in}(w) = \ln 2$$
$$\underline{w}^T \underline{x}_m \longrightarrow \infty \qquad E_{in}(w) \approx \ln(\infty) \approx \infty$$
$$\underline{w}^T \underline{x}_m \longrightarrow -\infty \qquad E_{in}(w) \approx \ln(1) \approx 0$$

so, when $\underline{w}^T \underline{x}_m \longrightarrow \infty$, $E_{in} \approx \infty$ (maximum)

(iii)

y_n	$\underline{w}^T \underline{x}$	$E = \ln(1 + e^{-y_n \underline{w}^T \underline{x}_n})$	
+1	>0	$0 < E_i < \ln 2$	case I
+1	$>>0$	$E_i \gtrsim 0$	
+1	<0	$E_i > \ln 2$	case II
+1	$<<0$	$E_i >> \ln 2$	

Here, the two parts in case I are correctly classified whereas the two parts in case II are incorrectly classified. As we can see from the third column, the E^{inc} is much larger than E^c. This is because in incorrect classification, the value of the discriminant function is very low. This causes the value of the exponential term inside the error to go high ($\because -y_n \underline{w}^T \underline{x}_n$). This, at the same time makes error $>> \ln 2$. So the size of contribution to error is much more in case of incorrect classification.

Q3. (a) $f(\underline{w}) = (\underline{a}^T\underline{w} - b)^2$ $\underline{a}, \underline{w}$ - D dimensional vectors

b - constant

we can say,

$$f(\underline{w}) = [a_1w_1 + a_2w_2 + \cdots + a_Dw_D - b]^2$$

Now, $$\nabla_{\underline{w}}(f) = \frac{\partial}{\partial w}(\underline{a}^T\underline{w} - b)^2$$

$$= 2(\underline{a}^T\underline{w} - b)\cdot\underline{a}$$

$$= 2[\underline{a}^T\underline{w}\cdot\underline{a} - b\cdot\underline{a}]$$

and,

$$\nabla^2_{\underline{w}}(f) = \frac{\partial^2}{\partial w^2}(\underline{a}^T\underline{w} - b)^2$$

$$= \frac{\partial}{\partial w}[2(\underline{a}^T\underline{w}\ \underline{a} - \underline{b}\cdot\underline{a})]$$

$$= 2a\cdot a^T$$

$$= 2\|a\|_2^2$$

So, $$\nabla^2_{\underline{w}}(f) \geq 0$$

As Hessian of the function is always greater than or equal to 0, so, $f(\underline{w})$ is a convex problem.

(b) $J(\underline{w}) = \| \underline{\underline{X}} \cdot \underline{w} - \underline{y} \|_2^2 + \underline{c}^T \underline{w}$

$$= \sum_{i=1}^{N} (\underline{x}_i^T \underline{w} - y_i)^2 + \underline{c}^T \underline{w}$$

Now,

$$\nabla_{\underline{w}} (J) = \frac{\partial}{\partial \underline{w}} J(\underline{w})$$

so, $\underline{x}_i$ and $\underline{w}$ are D dimensional

$\underline{x}_i$ and $y_i = 1, 2, \cdots N$

$\underline{c}$ are constants

In matrix form, we can write,

$$\sum_{i=1}^{N} (\underline{x}_i^T \underline{w} - y_i)^2$$

$$= \frac{1}{N} (y - \underline{\underline{X}} \underline{w})^T (y - \underline{\underline{X}} \underline{w})$$

$$= \frac{1}{N} \left[y^T y - (\underline{\underline{X}} \underline{w})^T y - y^T \underline{\underline{X}} \underline{w} + (\underline{\underline{X}} \underline{w})^T \underline{\underline{X}} \underline{w} \right)$$

$$= \frac{1}{N} (y^T y - y^T \underline{\underline{X}} \underline{w} - y^T \underline{\underline{X}} \underline{w} + \underline{w}^T \underline{\underline{X}}^T \underline{\underline{X}} \underline{w})$$

So, $J(w) = \frac{1}{N} (y^T y - 2 y^T \underline{\underline{X}} \underline{w} + \underline{w}^T \underline{\underline{X}}^T \underline{\underline{X}} \underline{w}) + \underline{c}^T \underline{w}$

Now, $\nabla_w (\underline{a}^T \underline{w}) = \underline{a}$

$\nabla_w (\underline{w}^T Q \underline{w}) = 2 Q \underline{w}$

Thus,

$$\nabla_{\underline{w}} (\underline{y}^T \underline{y}) = 0$$
$$\nabla_w \underline{y}^T \underline{\underline{X}} \underline{w} = \underline{y}^T \underline{\underline{X}}$$
$$\nabla_w (\underline{w}^T \underline{\underline{X}}^T \underline{\underline{X}} \underline{w}) = 2 \underline{\underline{X}}^T \underline{\underline{X}} \underline{w}$$
$$\nabla_w (\underline{c}^T \underline{w}) = \underline{c}$$

Thus,

$$\nabla_{\underline{w}} J(\underline{w}) = \frac{1}{N} \left(-2 \underline{y}^T \underline{\underline{X}} + 2 \underline{\underline{X}}^T \underline{\underline{X}} \underline{w} \right) + \underline{c}$$

$$\therefore \nabla^2_{\underline{w}} J(\underline{w}) = \frac{2 \underline{\underline{X}}^T \underline{\underline{X}}}{N} \qquad \because \nabla_{\underline{w}} \underline{y}^T \underline{\underline{X}} = 0$$
$$\nabla_{\underline{w}} \underline{\underline{X}}^T \underline{\underline{X}} \underline{w} = \underline{\underline{X}}^T \underline{\underline{X}}$$

Now we know,

$$z^T X^T X z = (Xz)^T (Xz)$$
$$= \| Xz \|_2^2 \geq 0$$

So, $x_i^T x_i$ is semi positive definite matrix.

$$\therefore \nabla^2_{\underline{w}} J(\underline{w}) > 0$$
$$\therefore J(\underline{w}) \text{ is convex}$$

Q4. $\hat{f}(\underline{x}) = \text{sgn}(\underline{w}^T \underline{x})$

D training data — No of input variables : $\mathcal{D}$
Objective function : $J(\underline{w}, \mathbf{D})$
$w_0 = 1 \quad w_j \in \{1, 2\} \quad \forall \; j \in \{1, 2, \dots \mathcal{D}\}$

(a) Number of elements in hypothesis set in case of a binary target function f is $= 2^{\mathcal{D}}$

(b) Hoeffding Inequality states that for any sample size N,

$$P[|\nu - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N} \quad \forall \; \epsilon > 0$$

So here,

$$P[|E_{in}(\hat{h}) - E_{out}(\hat{h})| > \epsilon]$$

$$= 2 \cdot 2^{\mathcal{D}} \cdot e^{-2\epsilon^2 \mathcal{D}}$$

$$= 2^{\mathcal{D}+1} \cdot e^{-2\epsilon^2 \mathcal{D}}$$