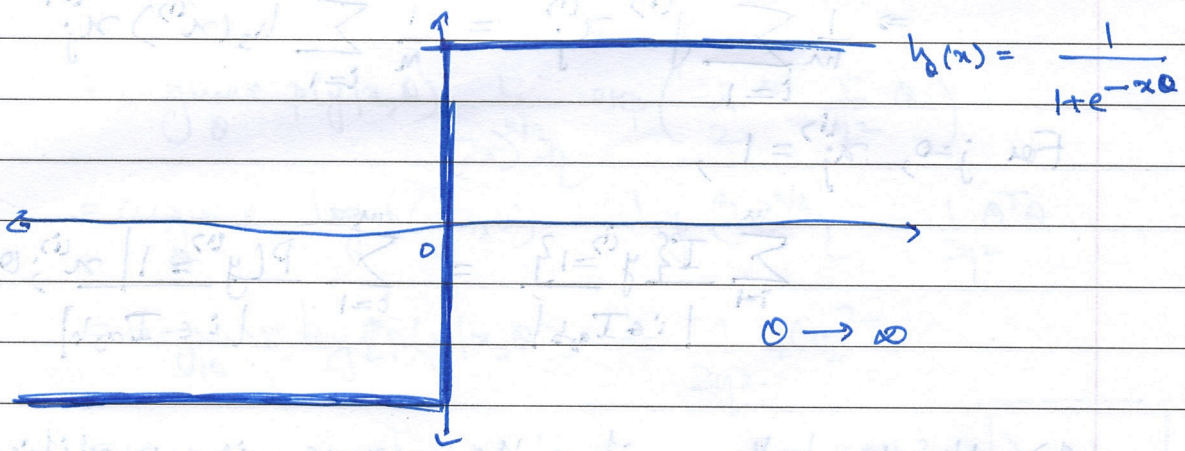
PS-2

M T W T F S S
Page No.:
Date:
YOUVA

1. (a) Convergence is achieved on dataset A but not on dataset B.

(b) Dataset B is more separated than dataset A. This effectively means that in the loss equation ($J(\theta)$), some of the parameters are going towards infinity. This means that in $h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$, θ can be increased to an infinitely large value and yet the separation may be retained. In a setting where θ, x are 1-D, for example, θ can be set to an infinite value to get a curve like the one below:



Hence, the θ_j's have the incentive to grow larger.

(c) (i) (iii) won't fix the issue as they don't modify the loss equation themselves.
(ii) (iv) (v) will work: (iv) penalizes the MLE params going towards infinity; (v) will make the dataset not readily separable.
(i) will ensure convergence eventually.

(d) SVMs do not fall prey to B-esque datasets as they tend to work with geometric margins, unlike classic logistic regression where functional margin is used. FM has the risk of being manipulated by the inflation of parameters; GM uses normalization as a safeguard.

2. (a) ~~Update~~ Gradient calculation:

$$\frac{\partial \ell(\theta)}{\partial \theta_j} = \frac{1}{m} \sum_{i=1}^{m} \left(y^{(i)} - h_\theta(x^{(i)}) \right) x_j^{(i)}$$

When model is done training:

$$\frac{\partial \ell(\theta)}{\partial \theta_j} = 0$$

$$\Rightarrow \frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} - h_\theta(x^{(i)}) \right] x_j^{(i)} = 0$$

$$\Rightarrow \frac{1}{m} \sum_{i=1}^{m} y^{(i)} x_j^{(i)} = \frac{1}{m} \sum_{i=1}^{m} h_\theta(x^{(i)}) x_j^{(i)}$$

For $j=0$, $x_j^{(i)} = 1$,

$$\sum_{i=1}^{m} \frac{I\{y^{(i)}=1\}}{|i \in I_{a,b}|} = \sum_{i=1}^{m} \frac{P(y^{(i)}=1 \mid x^{(i)}; \theta)}{|i \in I_{a,b}|}$$

(b) No to both, if the range is modified to say $(0.2, 1)$ then $0.2 < P(y^{(i)}=1 \mid x^{(i)}; \theta) < 1$

$$\Rightarrow \sum_i \frac{P(y \mid x; \theta)}{m} < \sum_i \frac{I\{y^{(i)}=1\}}{m}$$

Hence, perfect acc violated.

Similarly, it can be shown that perfectly accurate models needn't be perfectly calibrated.

(c)

$$\text{New } J = -\sum_i y^{(i)} \log h(x^{(i)}) + (1-y^{(i)})(1-\log h(x^{(i)})) + \frac{1}{2}\lambda \|\theta\|_2^2$$

$$\frac{\partial J}{\partial \theta_j} = \sum_{i=1}^{m} -y_j^{(i)} + h(x^{(i)}) x_j^{(i)} + \lambda \theta_j = 0 \text{ (full training)}$$

$$\Rightarrow \sum_i \frac{I\{y^{(i)}=1\}}{m} = \lambda\theta + \sum_i \frac{P(y^{(i)} \mid x^{(i)}; \theta)}{m}$$

Not well-calibrated

~~3. $\theta_{MAP} = \underset{\theta}{\text{argmax}}\ p(\theta|x,y)$~~

~~$= \underset{\theta}{\text{argmax}}\ p(y|x,\theta)$~~

3. (a) $\theta_{MAP} = \underset{\theta}{\text{argmax}}\ p(\theta|x,y) = \underset{\theta}{\text{argmax}}\ \frac{p(\theta,x,y)}{p(x,y)}$

$= \underset{\theta}{\text{argmax}}\ \frac{p(y|x,\theta)\,p(\theta|x)\,p(x)}{p(y,x)}$

$= \underset{\theta}{\text{argmax}}\ p(y|x,\theta)\,p(\theta|x) = \underset{\theta}{\text{argmax}}\ p(y|x,\theta)\,p(\theta)$

(b) $\theta_{MAP} = \underset{\theta}{\text{argmax}}\ p(y|x,\theta)\,p(\theta)$, $\theta \sim N(0, \eta^2 I)$

$= \underset{\theta}{\text{argmax}}\ p(y|x,\theta)\,\frac{1}{(2\pi)^{n/2}\eta}\exp\left(\frac{-1}{2}\theta^T\frac{I}{\eta^2}\theta\right)$

$= \underset{\theta}{\text{argmax}}\ \log p(y|x,\theta) - \log(2\pi)^{n/2}\eta - \frac{1}{2\eta^2}\theta^T\theta$

$= \underset{\theta}{\text{argmax}}\ \log p(y|x,\theta) - \frac{1}{2\eta^2}\|\theta\|_2^2$

$= \underset{\theta}{\text{argmin}} -\log p(y|x,\theta) + \frac{1}{2\eta^2}\|\theta\|_2^2 \Rightarrow \boxed{\lambda = \frac{1}{2\eta^2}}$

(c) $y = \theta^T x + \epsilon$, $\epsilon = y - \theta^T x$, $\epsilon \sim N(0,\sigma^2)$

$\Rightarrow P(\epsilon) = P(y|x,\theta) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-1}{2\sigma^2}(y-\theta^T x)^2\right)$

$\Rightarrow \theta_{MAP} = \underset{\theta}{\text{argmin}} -\log p(y|x,\theta) + \frac{1}{2\eta^2}\|\theta\|_2^2$

$= \underset{\theta}{\text{argmin}} -\log \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left(\frac{-1}{2}\frac{\|\vec{y} - X\theta\|_2^2}{\sigma^2}\right) + \frac{1}{2\eta^2}\|\theta\|_2^2$

$= \underset{\theta}{\text{argmin}} +\log\sqrt{2\pi}\sigma + \frac{1}{2\sigma^2}\|\vec{y} - X\theta\|^2 + \frac{1}{2\eta^2}\|\theta\|_2^2$

$= \underset{\theta}{\text{argmin}} \underbrace{\left[\frac{1}{2\sigma^2}\|\vec{y} - X\theta\|_2^2 + \frac{1}{2\eta^2}\|\theta\|_2^2\right]}_{J(\theta)}$

(d) $\theta_{MAP} = \underset{\theta}{\operatorname{argmax}}\ p(y|x,\theta)\, p(\theta)$

$= \underset{\theta}{\operatorname{argmax}}\ \log p(y|x,\theta) + \log p(\theta)$

~~$= \underset{\theta}{\operatorname{argmax}}\ \frac{1}{(2b)^m} \exp\left(\right.$~~

$= \underset{\theta}{\operatorname{argmax}} \log \frac{1}{(2\pi)^{m/2}\sigma^m} \exp\left(\frac{-\|\vec{y}-X\theta\|_2^2}{2\sigma^2}\right) + \log \frac{1}{2b} \exp\left(\frac{-\|\theta\|_1}{b}\right)$

$= \underset{\theta}{\operatorname{argmin}}\ \frac{1}{2\sigma^2}\|\vec{y}-X\theta\|_2^2 + \frac{\|\theta\|_1}{b}$

$= \underset{\theta}{\operatorname{argmin}}\ \|\vec{y}-X\theta\|_2^2 + \underbrace{\frac{2\sigma^2}{b}}_{\gamma} \|\theta\|_1$

4. (a) Yes $K(x,z) = K_1(x,z) + K_2(x,z)$

$y^T K y = y^T (K_1 + K_2) y$

$= (y^T K_1 + y^T K_2) y = \underset{\geq 0}{y^T K_1 y} + \underset{\geq 0}{y^T K_2 y} \geq 0$

(b) No, if $K_2 = 5K_1$, $y^T K y = -4 y^T K y \leq 0$ No

(c) Yes, $y^T aK y = a\, y^T K y \geq 0$

(d) No, similar to above and (b)

(e) $y^T K y = y^T K_1 K_2 y$

$$= \sum_i^m \sum_j^m y_i K_1(x^{(i)}, x^{(j)}) K_2(x^{(i)}, x^{(j)}) y_j$$

$$= \sum_i^m \sum_j^m y_i \phi_1(x^{(i)})^T \phi_1(x^{(j)}) \phi_2(x^{(i)})^T \phi_2(x^{(j)}) y_j$$

$$= \sum_i^m \sum_j^m y_i \sum_k^n \phi_{1k}(x^{(i)})^T \phi_{1k}(x^{(j)}) \sum_l^n \phi_{2l}(x^{(i)})^T \phi_{2l}(x^{(j)}) y_j$$

$$= \sum_k^n \sum_l^n \sum_i^m \sum_j^m y_i \phi_{1k}(x^{(i)})^T \phi_{1k}(x^{(j)}) \phi_{2l}(x^{(i)})^T \phi_{2l}(x^{(j)}) y_j$$

$$= \sum_k^n \sum_i \sum_j \left(y_i \phi_{1k}(x^{(i)}) \phi_{2k}(x^{(i)}) \right)^2 \geq 0$$ (Yes)

(f) $y^T K y = y^T f(x) f(z)\, y = f(x) f(z) \|y\|_2^2 \geq 0$ (Yes)

(g) $y^T K y = y^T K_3(\phi(x), \phi(z))\, y$

$= y^T \phi(\phi(x))^T \phi(\phi(z))\, y$

$= y^T \phi(x')^T \phi(z')\, y \geq 0$ (Yes)

(h) from (e) and (a), it can be inferred. (Yes)

5. (a) (i) $\theta^{(i)} = \sum_{j=1}^{i} \beta_j \phi(x^{(j)})$, $\theta^{(0)} = \sum_{j=1}^{0} \beta_j \phi(x^{(j)}) = \vec{0}$

(ii) $h_{\theta^{(i)}}(x^{(i+1)}) = g(\theta^{(i)T} \phi(x^{(i+1)}))$

$= g(\langle (\sum_{j=1}^{i} \beta_j \phi(x^{(j)}))^T \phi(x^{(i+1)}) \rangle)$

$= g(\sum_{j=1}^{i} \beta_j \phi(x^{(j)})^T \phi(x^{(i+1)}))$

$= g(\sum_{j=1}^{i} \beta_j \langle \phi(x^{(j)}), \phi(x^{(i+1)}) \rangle)$

$= \text{sign}(\sum_{j=1}^{i} \beta_j K(x^{(j)}, x^{(i+1)}))$

(iii) $\theta^{(i+1)} := \theta^{(i)} + \alpha(y^{(i+1)} - h_{\theta^{(i)}}(x^{(i+1)}))\phi(x^{(i+1)})$

$:= \sum_{j=1}^{i} \beta_j \phi(x^{(j)}) + \alpha(y^{(i+1)} - \text{sign}(\sum_{j=1}^{i} \beta_j K(x^{(j)}, x^{(i+1)})))\phi(x^{(i+1)})$ (β_{i+1})

$:= \sum_{j=1}^{i} \beta_j \phi(x^{(j)}) + \alpha y^{(i+1)} \phi(x^{(i+1)}) - \alpha \text{sign}(\sum_{j=1}^{i} \beta_j K(x^{(j)}, x^{(i+1)}))\phi(x^{(i)})$

(b) Coded

(c) The Dot kernel performs worse. RBF kernel defines a larger parameter space hence improving performance. Dot kernel essentially does a plain regression which is unsuitable for the dataset.

6. (a) (b) (c) (d) Coded