1.

(a) dataset A : converges

dataset B : not converges

(b) dataset B is linear seperable,

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \log\left(1+\exp\left(-y^{(i)}\theta^T x^{(i)}\right)\right)$$

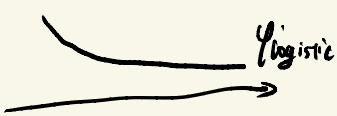
$$\nabla_\theta J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} \frac{\exp\left(-y^{(i)}\theta^T x^{(i)}\right) y^{(i)} x^{(i)}}{1+\exp\left(-y^{(i)}\theta^T x^{(i)}\right)}$$

$$= -\frac{1}{m}\sum_{i=1}^{m} \frac{y^{(i)}x^{(i)}}{\exp\left(y^{(i)}\theta^T x^{(i)}\right)+1}$$

$y^{(i)}\theta^T x^{(i)} > 0$ if the θ is already devide the dataset B into 2 part.

$\theta^{(k+1)} = \theta^{(k)} - \alpha \nabla_\theta J(\theta)$

which makes $J(\theta)$ decreases



it will converges slowly for $y^{(i)}x^{(i)T}\theta > 0$

(c) (i) no, no help to solve the problem

(ii) yes, it can make $\alpha\nabla_\theta J(\theta)$ decreases quickly

(iii) no, it just scalar θ

(iv) yes, it will make $\theta/\|\theta\|_2^2$ smaller quickly

(v) yes, it will make data not linear seperable

(d) no, $[1-z]_+$ hinge function will be 0 as $z>1$, so it will be minimized to 0 if the data set is linear seperable

use the l_2 regularization

$$\frac{\partial}{\partial\theta_i}\|\theta\|_2^2 = \frac{\partial(\theta_1^2+\theta_2^2+\cdots+\theta_n^2)}{\partial\theta_i}$$

$$= 2\theta_i$$

$$\nabla_\theta \|\theta\|_2^2 = 2\theta$$

2.

(a)

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

$$\sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)})) x_j^{(i)} = 0$$

when $j = 0$

$$\sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)})) = 0$$

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

$$\Sigma_{i \in S_{a,b}} P(y^{(i)} = 1 | x^{(i)}; \theta) = \Sigma_{i \in S_{a,b}} \mathbb{I}\{y^{(i)} = 1\}$$

(b) NO

for example, when $a = 0$, $b = 0.5$

all prediction is 0, but $P(y^{(i)} = 0 | x^{(i)}; \theta)$ is not 0. the converse is also false

(c) $\frac{\partial J(\theta)}{\partial \theta_j} = \frac{\partial l(\theta)}{\partial \theta_j} + c\theta_i = 0$

$$\frac{\partial l(\theta)}{\partial \theta_i} = -c\theta_i$$

so it will make it not perfectly calibrated

3. (a)

$$P(\theta | x, y) = \frac{P(\theta, x, y)}{P(x, y)} = \frac{P(y | x, \theta) P(x, \theta)}{P(x, y)}$$

$$= \frac{P(y | x, \theta) P(x) P(\theta)}{P(x, y)}$$

so $\arg\max_\theta P(\theta | x, y) =$

$$\arg\max_\theta P(y | x, \theta) P(\theta)$$

(b) $\theta_{MAP} = \arg\max_\theta P(y | x, \theta) P(\theta)$

$$= \arg\min_\theta -\log (P(y | x, \theta) P(\theta))$$

$$= \arg\min_\theta (-\log P(y | x, \theta) - \log (\frac{1}{(2\pi)^{\frac{n}{2}} \eta^n} \exp\{-\frac{1}{2\eta} \|\theta\|_2^2\}))$$

$$= \arg\min_\theta (-\log P(y | x, \theta) + \frac{1}{2\eta^2} \|\theta\|_2^2)$$

so for $\lambda = \frac{1}{2\eta^2}$

it equivalent to L_2 regularization

(c)

$$P(y^{(i)} | x^{(i)}, \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(y^{(i)} - \theta x^{(i)})^2}{2\sigma^2})$$

$$\arg\min_\theta (-\log P(\vec{y} | X, \theta) + \frac{1}{2\eta^2} \|\theta\|_2^2)$$

$$= \arg\min_\theta (\frac{1}{2\sigma^2} \|\vec{y} - X\theta\|_2^2 + \frac{1}{2\eta^2} \|\theta\|_2^2)$$

$$J(\theta) = \frac{1}{2\sigma^2} \|\vec{y} - X\theta\|_2^2 + \frac{1}{2\eta^2} \|\theta\|_2^2$$

$$\nabla_\theta J(\theta) = \frac{1}{\sigma^2} X^T(\vec{y} - X\theta) + \frac{1}{\eta^2}\theta$$

$$= \frac{1}{\sigma^2}(X^T X\theta - X^T\vec{y}) + \frac{1}{\eta^2}\theta = 0$$

$$\left(\frac{1}{\sigma^2}X^T X + \frac{1}{\eta^2} I\right)\theta = \frac{1}{\sigma^2} X^T\vec{y}$$

$$\theta = \left(\frac{1}{\sigma^2}X^T X + \frac{1}{\eta^2} I\right)^{-1} \frac{1}{\sigma^2} X^T\vec{y}$$

(d) $\log P(\theta)$

$$= -\log 2b - \frac{|z-\mu|}{b} \quad |_{\mu=0}$$

$$= -\log 2b - \frac{|z|}{b}$$

$$\theta_{MAP} = \arg\min_\theta \left(\frac{1}{2\sigma^2}\|\vec{y} - X\theta\|_2^2 + \frac{|\theta|}{b}\right)$$

$$= \arg\min_\theta \left(\|\vec{y} - X\theta\|_2^2 + \frac{2\sigma^2}{b}\|\theta\|_1\right)$$

so $J(\theta) = \|X\theta - \vec{y}\|_2^2 + \gamma\|\theta\|_1$

where $\gamma = \frac{2\sigma^2}{b}$

4.

(a) $K_{ij} = k(x^{(i)}, x^{(j)}) = k_1(x^{(i)}, x^{(j)}) + k_2(x^{(i)}, x^{(j)})$

$= k_{1ij} + k_{2ij}$

so $k = k_1 + k_2$

$z^T k z = z^T k_1 z + z^T k_2 z \geq 0$

so it is a kernel

(b) it isn't a kernel

for $k_2 = 2k_1$

$k = -k_1$

$z^T k z = -z^T k_1 z \leq 0$

(c) $k = ak_1$

$z^T k z = a z^T k z \geq 0$

so it's a kernel

(d) it isn't a kernel

(e) $K_{ij} = k_{1ij} k_{2ij}$

$$z^T k z = \sum_{i=1}^m \sum_{j=1}^m z_i k_{ij} z_j$$

$$= \sum_{i=1}^m \sum_{j=1}^m z_i k_{1ij} k_{2ij} z_j$$

$$= \sum_{i=1}^m \sum_{j=1}^m z_i \phi_1^T(x_i)\phi_1(x_j)\phi_2^T(x_i)\phi_2(x_j) z_j$$

$$= \sum_i \sum_j \sum_a \sum_b z_i \phi_{1a}(x_i)\phi_{1a}(x_j)\phi_{2b}(x_i)\phi_{2b}(x_j) z_j$$

$$= \sum_a \sum_b \sum_i \sum_j z_i \phi_{1a}(x_i)\phi_{1a}(x_j)\phi_{2b}(x_i)\phi_{2b}(x_j) z_j$$

$$= \sum_a \sum_b \sum_i \left(z_i \phi_{1a}(x_i)\phi_{2b}(x_i)\right) \sum_j \left(z_j \phi_{1a}(x_j)\phi_{2b}(x_j)\right)$$

$$= \sum_a \sum_b \left(\sum_i z_i \phi_{1a}(x_i)\phi_{2b}(x_i)\right)^2 \geq 0$$

so it is a kernel

(f) $K_{ij} = f(x^{(i)}) f(x^{(j)})$

so $K = F(x) F^T(x)$

where $F: R^n \to R^n$

and $F\left(\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}\right) = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}$

$z^T K z = z^T F(x) F^T(x) z$

$= \|F(x)^T z\|_2^2 \geq 0$

so it's a kernel

(g) $K_{ij} = k_3(\phi(x^{(i)}), \phi(x^{(j)}))$

for any set $\{y^{(1)}, \dots y^{(m)}\}$

we can find $\{x^{(1)}, \dots x^{(m)}\}$

where $\phi(x^{(i)}) = y^{(i)}$

so it's a kernel

(h) $k(x,z) = \sum_{k=0}^{n} c_k (k_1(x,z))^k$

$K_{ij} = \sum_{k=0}^{n} c_k (k_1(x^{(i)}, x^{(j)}))^k$

according to (e), $K(x,z)$ is valid

(a), (c) and

5. (a)

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

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

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

$\theta^{(0)} = \sum_{j=1}^{0} \beta_j \phi(x^{(j)}) = 0$

(ii) $h_{\theta^{(i)}}(x^{(i+1)}) = g\left(\sum_{j=1}^{i} \beta_j \phi(x^{(j)})^T \phi(x^{(i+1)})\right)$

$= g\left(\sum_{j=1}^{i} \beta_j k(x^{(j)}, x^{(i+1)})\right)$

(iii)

$\beta_{i+1} = \alpha (y^{(i+1)} - g(\sum_{j=1}^{i} \beta_j k(x^{(j)}, x^{(i+1)})))$

(c) dot product kernel is bad, because this data set is not linearly seperable

6.

(b) $\log \frac{p(y=1)}{p(y=0)} \frac{p(x \mid y=1)}{p(x^{(i)} \mid y=1)}$

$= (\log p(y=1) + \sum_{j=1}^{d} \log p(x_j \mid y=1))$

$- (\log p(y=0) + \sum_{j=1}^{d} \log p(x_j \mid y=0))$

$= \log \frac{\phi_y}{1-\phi_y} + \sum_{j=1}^{d} \log \frac{\phi_{j|y=1}}{\phi_{j|y=0}}$

using this method can easily predict