[PS 2]

#1.

(a) Training of data_B doesn't converge, whereas that of data_A does.

(b) Because data_B is linearly seperable, so scale of $\theta^T x$ gets larger to decrease loss until $|\theta|$ reaches ∞

(c) iii) → θ won't go ∞.

(O) v) → this prevents data_B from being linearly seperable.

(d) No, because it maximises geometric margin, with $|\theta|$ fixed.

#2.

(a) Learned θ of logistic regression model meets:

$$\frac{\partial J(\theta)}{\partial \theta} = \sum (y^i - h_\theta(x^i))\, x^i = 0$$

⇒ matrix form :

$$\underbrace{\begin{bmatrix} 1 & & 1 \\ \vdots & \cdots & \vdots \\ x_n^{(1)} & & x_n^{(m)} \end{bmatrix}}_{\substack{X \\ (n+1 \times m)}} \underset{(m\times 1)}{(Y - h(X))} = 0$$

consider 1st row of eq.

$\sum (y^i - h_\theta(x^i)) = 0$ $\quad (y^i \in \{0,1\})$

$\Rightarrow \sum P(y^i = 1 \mid x^i;\theta) = \sum h_\theta(x^i) = \sum y^i = \sum 1\{y^i = 1\}$

this is equivalent to question's condition $((a,b) = (0,1))$

(b) Perfect calibration doesn't necessarily imply perfect accuracy because calibration is about probability. Converse is necessarily true.

(c) $X(Y - h(X)) + 2\lambda\theta = 0$

Consider 1st row of eq.

$\sum (y^i - h_\theta(x^i)) + 2\lambda\theta_0 = 0 \Rightarrow \sum P(y^i = 1 \mid x^i;\theta) = \sum 1\{y^i = 1\} + 2\lambda\theta_0$

∴ prediction will be biased, proportional to θ_0

#3. Prove by contradiction.

assume $|\theta_{MAP}|_2 > |\theta_{ML}|_2$

then $p(\theta_{MAP}) < p(\theta_{ML}) \quad \because \theta \sim N(0, \tau^2 I)$

$\Rightarrow P(\theta_{MAP}) \prod p(y^i|x^i;\theta_{MAP}) < P(\theta_{ML}) \prod p(y^i|x^i;\theta_{MAP})$

$(\because \theta_{ML}$ maximises $\prod p(y^i|x^i;\theta)) \quad < P(\theta_{ML}) \prod p(y^i|x^i;\theta_{ML})$

$\Rightarrow$ contradicts to the fact that θ_{MAP} maximises $p(\theta)\prod p(y^i|x^i;\theta)$ //

#4.

(a) $u^T K u = \sum_{i,j} z_i z_j (K_1(x^i, x^j) + K_2(x^i, x^j))$

$= u^T K_1 u + u^T K_2 u \geq 0 \Rightarrow$ Kernel //

(b) not Kernel

(c) $u^T K u = a\, u^T K_1 u \quad \therefore$ Kernel

(d) not kernel

(e) Since $K_1(x,z), K_2(x,z)$ is Kernel, $\exists \phi_1 \in \mathbb{R}^{d_1}, \exists \phi_2 \in \mathbb{R}^{d_2}$

s.t. $K_1(x,z) = \phi_1(x)^T \phi_1(z)$, $K_2(x,z) = \phi_2(x)^T \phi_2(x)$

Then $K(x,z)$

$= \sum_i \phi_1(x)_i \phi_1(z)_i \sum_j \phi_2(x)_j \phi_2(z)_j$

$= \sum_i \sum_j (\phi_1(x)_i \phi_2(x)_j)(\phi_1(z)_i \phi_2(z)_j) = \underbrace{\text{flat}(\phi_1(x)\phi_2(x)^T)}_{\phi_3(x)}{}^T \underbrace{\text{flat}(\phi_1(z)\phi_2(z)^T)}_{\phi_3(z)}$

, $\phi_3 \in \mathbb{R}^{d_1 d_2} \Rightarrow K$ is Kernel //

(f) $u^T K u = \sum_{i,j} u_i u_j f(x^i) f(x^j) = (\sum u_i f(x^i))^2 \geq 0 \Rightarrow$ kernel

(g) $u^T K u = \sum u_i u_j K_3(\phi(x^i), \phi(x^j)) \geq 0 \Rightarrow$ Kernel

(h) $u^T K u = \sum u_i u_j (\sum \alpha_k (K_1(x^i, x^j))^k) \geq 0 \quad (\because (a), (c), (e))$

#5. Since θ is too high dim, $\theta^T \phi(x)$ is represented as Kernel $K(\theta, \phi(x))$ always. (a) To predict, $h_{\theta^i}(x^{i+1}) = g(K(\theta^i, \phi(x^{i+1})))$ (b) And param update rule is modified to

$K(\theta^{i+1}, \phi(x^{i+1})) := K(\theta^i, \phi(x^{i+1})) + \alpha 1\{K(\theta^i, \phi(x^{i+1}))y^{i+1} < 0\} K(\theta^i, \phi(x^{i+1})) y^{i+1}$

(c). Only dot product in kernel form need: $O(n^2) \to O(n)$

(#6. look at PS2-6.ipynb)