(1.a)

dataset A converged but dataset B didn't.

(1.b)

When dataset is linearly separable,the algorithm will keep increasing functional margin until it goes to infinite.

(1.c)

ii. iv. v.

Decrease learning rate or restrict the size of parameters or make the dataset linearly unseparable will help.

(1.d)

SVM could converged w.r.t dataset B, because it uses geometric margin.

(2.a)

$$\sum_{i \in I_{ab}} P(y^{(i)} = 1 | x^{(i)}; \theta) = \sum_{i \in I_{ab}} \mathbb{1}(y^{(i)} = 1)$$

$$when\ y = 1 \qquad \frac{\partial J(\theta)}{\partial \theta} = -x \frac{1}{1 + e^{-\theta^T x}} = -x(1 - \frac{1}{1 + e^{-\theta^T x}})$$

$$when\ y = 0 \qquad \frac{\partial J(\theta)}{\partial \theta} = -x(0 - \frac{1}{1 + e^{-\theta^T x}})$$

so $\frac{\partial J(\theta)}{\partial \theta} = -x^T(y - h_\theta(x)) = 0$

then $\sum y = \sum h_\theta(x)$

(2.b)

Both the answer are no. Suppose 50% of the data is positive, then a model which always outputs 0.5 will also be perfectly calibrated. On the other hand, a model that outputs 0.75 for positive examples and 0.25 for negative examples will have perfect accuracy, but is not perfectly-calibrated.

(2.c)

When a regularization $\lambda||\theta||$ is added, the equation in (b) becomes

$$\sum_{i=1}^{m} y^{(i)} = \sum_{i=1}^{m} h_\theta(x^{(i)}) + 2\lambda\theta_0$$

where θ_0 is the parameter for the intercept. In general, we will not penalize this term, and in this case regularization will have no effect.

(3.a)

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

(3.b)

$$p(\theta) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2}) = \frac{1}{\sqrt{2\pi}\eta I}\exp(-\frac{\theta^2}{2\eta^2 I})$$

$$p(\theta|x,y) = p(y|x,\theta)\cdot p(\theta) = p(y|x,\theta)\cdot\frac{1}{\sqrt{2\pi}\eta I}\exp(-\frac{\theta^2}{2\eta^2 I})$$

$$\theta_{MAP} = \arg\max_\theta(\log p(\theta|x,y)) = \arg\max_\theta \ \log p(y|x,\theta) + log\frac{1}{\sqrt{2\pi}\eta I} - \frac{\theta^2}{2\eta^2 I}$$

$$= \arg\min_\theta \ -logp(y|x,\theta) + \frac{\theta^2}{2\eta^2 I}$$

$$\lambda = \frac{1}{2\eta^2 I}$$

(3.c)

$$\varepsilon = y - \theta^T x \sim \mathbb{N}(0,\sigma^2)$$

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

$$theta_{MAP} = \arg\max_\theta \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(y-\theta^T x)^2}{2\sigma^2}) * \frac{1}{\sqrt{2\pi}\eta I}\exp(-\frac{\theta^2}{2\eta^2 I})$$

(3.d)

$$\theta_{MAP} = \arg\max_\theta \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(y-\theta^T x)^2}{2\sigma^2}) * \frac{1}{2b}\exp(-\frac{|\theta|}{b})$$

$$= \arg\min_\theta \frac{(y-\theta^T x)^2}{2\sigma^2} + \frac{|\theta|}{b}$$

$$\gamma = b$$

(4.a)

(a) Kernel. The sum of 2 positive semidefinite matrices is a positive semidefinite matrix:
$\forall z z^T G_1 z \geq 0, z^T G_2 z \geq 0$ since K_1, K_2 are kernels. This implies $\forall z z^T G z = z^T G_1 z + z^T G_2 z \geq 0$

(b) Not a kernel. Counterexample: let $K_2 = 2K_1$ (we are using (1c) here to claim K_2 is a kernel). Then we

have $\forall z z^T G z = z^T (G_1 - 2G_1) z = -z^T G_1 z \leq 0$

(c) Kernel. $\forall z z^T G_1 z \geq 0$, which implies $\forall z a z^T G_1 z \geq 0$

(d) Not a kernel. Counterexample: $a = 1$. Then we have $\forall z - z^T G_1 z \leq 0$

(e) Kernel. K_1 is a kernel, thus $\exists \phi^{(1)} K_1(x,z) = \phi^{(1)}(x)^T \phi^{(1)}(z) = \sum_i \phi_i^{(1)}(x) \phi_i^{(1)}(z)$ Similarly, K_2 is a kernel, thus $\exists \phi^{(2)} K_2(x,z) = \phi^{(2)}(x)^T \phi^{(2)}(z) = \sum_j \phi_j^{(2)}(x) \phi_j^{(2)}(z)$

$$\begin{aligned} K(x,z) &= K_1(x,z) K_2(x,z) \\ &= \sum_i \phi_i^{(1)}(x) \phi_i^{(1)}(z) \sum_i \phi_i^{(2)}(x) \phi_i^{(2)}(z) \\ &= \sum_i \sum_j \phi_i^{(1)}(x) \phi_i^{(1)}(z) \phi_j^{(2)}(x) \phi_j^{(2)}(z) \\ &= \sum_i \sum_j \left(\phi_i^{(1)}(x) \phi_j^{(2)}(x) \right) \left(\phi_i^{(1)}(z) \phi_j^{(2)}(z) \right) \\ &= \sum_{(i,j)} \psi_{i,j}(x) \psi_{i,j}(z) \end{aligned}$$

Where the last equality holds because that's how we define ψ. We see K can be written in the form $K(x,z) = \psi(x)^T \psi(z)$ so it is a kernel. Here is an alternate super-slick linear-algebraic proof. If G is the Gram matrix for the product $K_1 \times K_2$, then G is a principal submatrix of the Kronecker product $G_1 \otimes G_2$ where G_i is the Gram matrix for K_i. As the Kronecker product is positive semi-definite, so are its principal submatrices.

(f) Kernel. Just let $\psi(x) = f(x)$, and since $f(x)$ is a scalar, we have $K(x,z) = \phi(x)^T \phi(z)$ and we are done.

(g) Kernel. since K_3 is a kernel, the matrix G_3 obtained for any finite set $\left\{ x^{(1)}, \ldots, x^{(m)} \right\}$ is positive semidefinite, and so it is also positive semidefinite for the sets $\left\{ \phi\left(x^{(1)}\right), \ldots, \phi\left(x^{(m)}\right) \right\}$

(h) Kernel. By combining (1a) sum,(1c) scalar product, (1e) powers,(1f) constant term, we see that any polynomial of a kernel K_1 will again be a kernel.

(5.a)

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

$$\theta^{(i)} = \sum_{l=1}^{i} \beta_l^{(i)} \phi(x^{(l)})$$

$$h_\theta(x) = g(\theta^T x) = g(\theta^T \phi(x))$$

$$= g(\sum_{l=1}^{i} \beta_l^{(i)} \phi(x)) = g(\alpha(y^{(l)} - h_\theta^{(l-1)}(\phi(x^{(l)})) \cdot \phi(x))$$

$$= g(\sum_{l=1}^{i} \alpha(y^{(l)} - h_\theta^{(l-1)}(\phi(x^{(l)})\phi(x))))$$

$$= g(\sum_{l=1}^{i} \alpha(y^{(l)} - h_\theta^{(l-1)}(K(x^{(l)}, x))))$$

$$= g(\sum_{l=1}^{i} \beta_l^{(i)} K(x^{(l)}, x))$$

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

if $x^{(i)}$ *is misclassified then*

$$y^{(i)} - h_\theta^{(i-1)}(\phi(x^{(i)})) = \pm 1$$

$$\theta = \sum_{\{i: y^{(i)} \neq h_\theta^{(i)}(\phi(x^{(i)}))\}} \alpha(2y^{(i)} - 1)\phi(x^{(i)})$$