1. The true risk of a binary classification is:

$$L_D(h) = P_{(x,y)\sim D}[h(x) \neq y]$$

let h be any predictor, and $f_D(x)$ be Bayes predictor.

To prove $L_D(f_D) \leq L_D(h)$

we get $P[f_D \neq y | X=x] - P[h(x) \neq y | X=x] \overset{\leq}{\cancel{\geq}} 0$ (sorry, ~~write~~ wrong order so I had to use "$\leq$")

$= [1-2P(y=1|x)]\mathbb{1}_{h(x)=1} + P(y=1|x) - [1-2P(y=1|x)]\mathbb{1}_{f_D(x)=1} - P(y=1|x)$

$= [1-2P(y=1|x)][\mathbb{1}_{h(x)=1} - \mathbb{1}_{f_D(x)=1}]$

We need to prove these two parts are both positive or negative.

If $1-2P(y=1|x) > 0$, then $P(y=1|x) < \frac{1}{2}$, in this case $f_D(x) = -1$

Thus, $\mathbb{1}_{f_D(x)} = 0$ and $\mathbb{1}_{h(x)=1} - \mathbb{1}_{f_D(x)=1} = \mathbb{1}_{h(x)=1} \in \{0,1\}$, is positive.

If $1-2P(y=1|x) < 0$, then $P(y=1|x) > \frac{1}{2}$, in this case $f_D(x) = 1$,

Thus, $\mathbb{1}_{f_D(x)} = 1$, and $\mathbb{1}_{h(x)=1} - \mathbb{1}_{f_D(x)=1} = \mathbb{1}_{h(x)=1} - 1 \in \{-1,0\}$ is negative.

Therefore the quantity is positive, since it is the product of two positive or two negative numbers.

So, $L_D(f_D) \leq L_D(h)$ and $f_D(x)$ is the optimal predictor.

2.(a) Decision boundary should be the intersect between two normal distribution curve.

$$\frac{1}{1\cdot\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-0}{1}\right)^2} = \frac{1}{\sqrt{0.5}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\frac{2}{3}\pi}{\sqrt{0.5}}\right)^2}$$

$$e^{-\frac{1}{2}x^2} = 0.5^{-\frac{1}{2}} e^{-\frac{1}{2}\left(\frac{x-\frac{2}{3}\pi}{\sqrt{0.5}}\right)^2}$$

$$-\frac{x^2}{2} = \ln 0.5^{-\frac{1}{2}} - \frac{1}{2}\left((x-\tfrac{2}{3}\pi)\cdot 0.5^{-\frac{1}{2}}\right)^2$$

$$x^2 = \ln^{0.5} + 2(x-\tfrac{2}{3}\pi)^2$$

$$x^2 = \ln^{0.5} + 2(x^2 - \tfrac{4}{3}\pi x + \tfrac{4}{9}\pi^2)$$

$$x^2 = \ln^{0.5} + 2x^2 - \tfrac{8}{3}\pi x + \tfrac{8}{9}\pi^2$$

$$x^2 - \tfrac{8}{3}\pi x + \tfrac{8}{9}\pi^2 + \ln^{0.5} = 0$$

we can get $x_1 = 1.11208273$ $x_2 = 7.26549$.

(b) Shown in code file. $L_D(f_D) = 0.10771867033630828$

(c) At hypothesis space $H = \{\frac{i}{400} : i \in [1200]\}$ and it only has one parameter b. to get h^*, we wish b^* close to 1.11208273.

$$\frac{i}{400} = 1.11208273 \rightarrow i \approx 445$$

$$b^* = \frac{445}{400} = 1.1125$$

(d) Shown in code file. $L_D(h^*) = 0.10771869912674997$.

(e)(f) Shown in code. We got $b^* = 1.325$ for the best hypothesis h_S. and $L_D(h_S) = 0.11543174391346$

4. Derivative rules we need: $\frac{d}{dx}(\ln(x)) = \frac{1}{x}, x>0$, $\frac{d}{dx}(e^x) = e^x$

$$\frac{dL(h_w,S)}{dw} = \frac{1}{m}\sum_{i=1}^{m} \frac{1}{(1+\exp(-y_i\langle w,x_i\rangle))} \cdot \exp(-y_i\langle w,x_i\rangle) \cdot (-y_i \cdot x_i)$$

exp( ) always returns positive so we are all good. Only change part has w.

$$\frac{dL(h_w,S)}{dw} = \frac{1}{m}\sum_{i=1}^{m} \frac{\exp(-y_i\langle w,x_i\rangle)}{(1+\exp(-y_i\langle w,x_i\rangle))} \cdot (-y_i \cdot x_i)$$

5. Take derivative for equation (8) (just like equation 51 in slides)

$$\frac{dL_{S,L_2}(h_w)}{dw} = 2\sum_{i=1}^{m}(\langle w,x_i\rangle - y_i)x_i + 2\lambda w$$

set it to zero

$$\cancel{2}\sum_{i=1}^{m}(\langle w,x_i\rangle - y_i)x_i + \cancel{2}\lambda w = 0$$

$$\sum_{i=1}^{m}\langle w,x_i\rangle x_i - \sum_{i=1}^{m} y_i x_i + \lambda w = 0$$

We know $A = \sum_{i=1}^{m} x_i x_i^T$ $b = \sum_{i=1}^{m} y_i x_i$, isolate w we have $\langle w,x_i\rangle x_i = (x_i x_i^T)$

$$Aw - b + \lambda w = 0$$

$w(A+\lambda) = b \rightarrow$ matrix transform for λ we need I times λ

$w(A+\lambda I) = b \rightarrow$ and based on $(A+\lambda I)$ is invertible

$$w \cdot (A+\lambda I) \cdot (A+\lambda I)^{-1} = b \cdot (A+\lambda I)^{-1}$$

$$w = (A+\lambda I)^{-1} b$$