CMSC 25025 HW3 Xinyu Wei

1.

Show $\min(a,b) \le \sqrt{ab}$.

①. if $a<b$. $\min(a,b) = a = \sqrt{a \cdot a} < \sqrt{ab}$

② if $a>b$ $\min(a,b) = b = \sqrt{b \cdot b} < \sqrt{ab}$

③. if $a=b$ $\min(a,b) = a = b = \sqrt{a \cdot a} = \sqrt{bb} = \sqrt{ab}$

So $\min(a,b) \le \sqrt{ab}$ in all cases

$$P(\text{error}) = \int \mathbb{I}_1 \cdot P(X=x, Y=2)\,dx + \int \mathbb{I}_2 \cdot P(X=x, Y=1)\,dx.$$

$\mathbb{I}_i$, indicator function $= \begin{cases} 1 & \text{when predict } Y=i \\ 0 & \text{else} \end{cases}$

$$= \int \mathbb{I}_1 P(X=x \mid Y=2) \cdot \pi_2 \,dx + \int \mathbb{I}_2 P(X=x \mid Y=1)\, \pi_1 \,dx$$

$\pi_i : P(Y=i)$

$$= \int [(\mathbb{I}_1=1) \cdot P(X=x \mid Y=2) \cdot \pi_2 + (\mathbb{I}_1=0)\, P(X=x \mid Y=2)]\,dx$$
$$+ \int [(\mathbb{I}_2=1) \cdot P(X=x \mid Y=1)\, \pi_1 + (\mathbb{I}_2=0)\, P(X=x \mid Y=1)]\,dx$$

$$= \int \mathbb{I}_1 \cdot \min_{c \in \{1,2\}} (P(X=x \mid Y=c) \cdot \pi_c)\,dx + \int \mathbb{I}_2 \cdot \min_{c \in \{1,2\}} (P(X=x \mid Y=c) \cdot \pi_c)\,dx$$

since when $\mathbb{I}_1 = 1$, predict $Y=1$, $P(X=x \mid Y=2)$ must be smaller than $P(X=x \mid Y=1)$

similarly when $\mathbb{I}_2 = 2$, predict $Y=2$, $P(X=x \mid Y=1)$ must be smaller than $P(X=x \mid Y=2)$.

$$= \int \min_{c \in \{1,2\}} (P(X=x \mid Y=c) \cdot \pi_c)\,dx$$

$$\le \int \sqrt{P(X=x \mid Y=1)\,\pi_1 \cdot P(X=x \mid Y=2)\,\pi_2}\,dx$$

$$= \sqrt{\pi_1 \pi_2} \int \sqrt{P(X=x \mid Y=1) \cdot P(X=x \mid Y=2)}\,dx.$$

where $\pi_1 = P(Y=1)$

$\pi_2 = P(Y=2)$

$$\le \frac{1}{2} \int \sqrt{P(X=x \mid Y=1) \cdot P(X=x \mid Y=2)}\,dx$$

Since $\pi_1 + \pi_2 = 1$. $\sqrt{\pi_1 \pi_2} \le \frac{\pi_1 + \pi_2}{2} = \frac{1}{2}$

X is continuous,

So $P(\text{error}) \le \sqrt{P(Y=1)\,P(Y=2)} \int \sqrt{f(x \mid Y=1) \cdot f(x \mid Y=2)}\,dx$

$$\le \frac{1}{2} \int \sqrt{f(x \mid Y=1)\; f(x \mid Y=2)} \cdot dx$$

2. (a).

$f(x|Y=k) \sim N(\mu_k, \sigma^2) \quad k=1,2$

Assume $\mu_1 > \mu_2$.

Decision boundary: At boundary x_0, $f(x_0|Y=1)\pi_1 = f(x_0|Y=2)\cdot\pi_2$

$$\frac{f(x_0|Y=1)}{f(x_0|Y=2)} = \frac{\pi_2}{\pi_1}$$

$$\frac{f(x_0|Y=1)}{f(x_0|Y=2)} = \exp\left(-\frac{1}{2\sigma^2}((x_0-\mu_1)^2-(x_0-\mu_2)^2)\right)$$

$$= \exp\left(-\frac{1}{2\sigma^2}(x_0^2+\mu_1^2-2x_0\mu_1-x_0^2-\mu_2^2+2x_0\mu_2)\right)$$

$$= \exp\left(-\frac{1}{2\sigma^2}(2x_0(\mu_2-\mu_1)+\mu_1^2-\mu_2^2)\right)$$

$$= \frac{\pi_2}{\pi_1}$$

$$(2x_0(\mu_2-\mu_1)+\mu_1^2-\mu_2^2) = -\left(\log\frac{\pi_2}{\pi_1}\right)2\sigma^2$$

$$x_0 = \frac{-(\log\frac{\pi_2}{\pi_1}\cdot 2\sigma^2)-\mu_1^2+\mu_2^2}{2(\mu_2-\mu_1)}$$

$$= \frac{\mu_2^2-\mu_1^2-2\sigma^2\log\frac{\pi_2}{\pi_1}}{2(\mu_2-\mu_1)}$$

So the decision boundary is $x_0 = \dfrac{\mu_2^2-\mu_1^2-2\sigma^2\cdot\log\frac{\pi_2}{\pi_1}}{2(\mu_2-\mu_1)}$

Predict $\begin{cases} \hat{Y}=1 & \text{when } x \geq x_0 \\ \hat{Y}=2 & \text{when } x < x_0 \end{cases}$

Bayes Loss: $L(\hat{Y},Y) = E(\mathbb{1}_{\{\hat{Y}\neq Y\}})$ $\quad$ $\mathbb{1}_{\{\hat{Y}\neq Y\}}$ is the indicator function $= \begin{cases} 1 & \hat{Y}\neq Y \\ 0 & \hat{Y}=Y. \end{cases}$

$$= P(\hat{Y}\neq Y)\cdot 1 + P(\hat{Y}=Y)\cdot 0$$

$$= P(\hat{Y}\neq Y)$$

$$= P(x>x_0|Y=2)\pi_2 + P(x<x_0|Y=1)\pi_1$$

$$= P\left(\frac{x-\mu_2}{\sigma} > \frac{x_0-\mu_2}{\sigma}\right)\cdot\pi_2 + P\left(\frac{x-\mu_1}{\sigma} < \frac{x_0-\mu_1}{\sigma}\right)\cdot\pi_1$$

$$= \left[1-\Phi\left(\frac{x_0-\mu_2}{\sigma}\right)\right]\pi_2 + \Phi\left(\frac{x_0-\mu_1}{\sigma}\right)\cdot\pi_1$$

is the Bayes Loss

(b).

$P(\text{error}) = P(\hat{Y} \neq Y) = [1-\Phi(\frac{x_0-\mu_2}{\sigma})]\pi_2 + \Phi(\frac{x_0-\mu_1}{\sigma})\cdot\pi_1$

As $\sigma \to 0$, $\Phi(\frac{x_0-\mu_2}{\sigma}) \to 1$ because $\mu_1 > x_0 > \mu_2$; $x_0-\mu_2 > 0$, $\frac{x_0-\mu_2}{\sigma} \to +\infty$

$\Phi(\frac{x_0-\mu_1}{\sigma}) \to 0$ because $\mu_1 > x_0 > \mu_2$; $x_0-\mu_1 < 0$, $\frac{x_0-\mu_1}{\sigma} \to -\infty$

So $P(\hat{Y}\neq Y) \to (1-1)\pi_2 + 0\cdot\pi_1$

$= 0$

So $P(\text{error}) \to 0$ as $\sigma \to 0$.

(c).

For σ fixed, as $\pi_1 \to 0$, $X_0 = \frac{\mu_2^2-\mu_1^2-2\sigma^2\log\frac{\pi_2}{\pi_1}}{2(\mu_2-\mu_1)} \to +\infty$ since $\frac{\pi_2}{\pi_1} \to +\infty$ & $\mu_2 < \mu_1$

For small π_1, we can always predict $\hat{Y} = 2$.

$P(\text{error}) = P(\hat{Y}\neq Y) = [1-\Phi(\frac{x_0-\mu_2}{\sigma})]\pi_2 + \Phi(\frac{x_0-\mu_1}{\sigma})\pi_1$

$\to (1-1)\cdot\pi_2 + 1\cdot\pi_1$ as $x_0 \to +\infty$.

$= \pi_1$

here π_1 is small, so low error rate guaranteed.

(d).

Loss of individual example: $L(h,x) = \sum_{k=1}^{K} L_{k,h(x)} P(k|x)$.

Expectation of the individual loss: $E(L(h,x)) = \int_{h(x)} f(X=x)\cdot L(h,x)\,dx$

$= \int_{h(x)} f(X=x)\cdot \sum_{k=1}^{K} L_{k,h(x)} P(k|x)\,dx$

$= \sum_{l=1}^{K}\sum_{k=1}^{K} \int_{h(x)=l} L_{k,l}\, P(k|x)\, f(X=x)\,dx$

$= \sum_{k=1}^{K}\sum_{l=1}^{K} \int_{h(x)=l} P(x,k)\cdot L_{k,l}\,dx$

$= L(h)$

So to minimize $L(h)$, we can minimize $L(h,x)$ for each given x.

$\min L(h,x) = \sum_{k=1}^{K} L_{k,h_B(x)} P(k|x)$ by definition of $h_B(x)$.

So $h_B(x) = \underset{j=1,\dots,K}{\operatorname{argmin}} \sum_{k=1}^{K} L_{k,j} P(k|x)$ gives the lowest loss.

(e). Expected loss $L(h) = \sum_{k=1}^{2}\sum_{l=1}^{2} \int_{h(x)=l} p(x,k)\, L_{k,l}\, dx$

$$= \sum_{k=1}^{2}\left(\int_{h(x)=1} p(x,k) L_{k,1}\, dx + \int_{h(x)=2} p(x,k)\, L_{k,2}\, dx\right)$$

$$= \int_{h(x)=1} \left(p(x,1) L_{1,1} + p(x,2) L_{2,1}\right) dx + \int_{h(x)=2} \left(p(x,1)\cdot L_{1,2} + p(x,2)\, L_{2,2}\right) dx$$

$$h_B(x) = \underset{j=1,2}{\operatorname{argmin}} \sum_{k=1}^{2} L_{k,j}\, p(k|x)$$

decision boundary: $(L_{1,1} P(1|x_0)) + L_{2,1}\cdot P(2|x_0) = L_{1,2} P(1|x_0) + L_{2,2} P(2|x_0)$

$$(L_{1,1} - L_{1,2}) P(1|x_0) = (L_{2,2} - L_{2,1}) P(2|x_0)$$

$$P(1|x_0) = \frac{P(x_0|1)\cdot \pi_1}{f_X(x_0)}$$

$$P(2|x_0) = \frac{P(x_0|2)\cdot \pi_2}{f(x_0)}$$

$$\frac{P(1|x_0)}{P(2|x_0)} = \frac{L_{2,2} - L_{2,1}}{L_{1,1} - L_{1,2}}$$

$$LHS = \frac{P(x_0|1)\cdot \pi_1}{P(x_0|2)\cdot \pi_2} = \exp\left(-\frac{1}{2\sigma^2}\left(2x_0(\mu_2-\mu_1) + \mu_1^2 - \mu_2^2\right)\right)\cdot \frac{\pi_1}{\pi_2}$$

$$= \frac{L_{2,2} - L_{2,1}}{L_{1,1} - L_{1,2}}$$

$$-\frac{1}{2\sigma^2}\left(2x_0(\mu_2-\mu_1) + \mu_1^2 - \mu_2^2\right) = \log\left(\frac{L_{2,2}-L_{2,1}}{L_{1,1}-L_{1,2}}\cdot\frac{\pi_2}{\pi_1}\right)$$

$$x_0 = \frac{-2\sigma^2 \cdot \log\left(\frac{L_{2,2}-L_{2,1}}{L_{1,1}-L_{1,2}}\cdot \frac{\pi_2}{\pi_1}\right) + \mu_2^2 - \mu_1^2}{2(\mu_2-\mu_1)}$$

By definition of $L_{k,j}$, when $k=j$, prediction is correct,
So we can let $L_{k,j} = 0$ when $k=j$.

$$x_0 = \frac{2\sigma^2 \cdot \log\left(\frac{L_{1,2}\cdot\pi_1}{L_{2,1}\cdot\pi_2}\right) + \mu_2^2 - \mu_1^2}{2(\mu_2-\mu_1)}$$ is the decision boundary.

Let $L_{1,2} = \frac{1}{\pi_1}$, $L_{2,1} = \frac{1}{\pi_2}$, x_0 becomes $= \frac{2\sigma^2\cdot\log(1) + \mu_2^2 - \mu_1^2}{2(\mu_2-\mu_1)}$

$$= \frac{\mu_2^2 - \mu_1^2}{2(\mu_2-\mu_1)}$$

$= \frac{\mu_2+\mu_1}{2}$ is independent of π_1 and π_2.

So this will remedy the effect of small π_1 or π_2.

3. (a).

Newton's Iteration $\theta_{new} = \theta_{old} - \frac{\nabla J(\theta)}{H(\theta)}$

To find $\nabla J(\theta)$ and $H(\theta)$:

$sigm(\eta) = \frac{1}{1+e^{-\eta}} = \frac{e^{\eta}}{e^{\eta}+1}$

$P(y|X,\theta) = \prod_{i=1}^{n} \left(sigm(x_i\theta)^{y_i} \cdot (1-sigm(x_i\theta))^{1-y_i} \right)$ where $x_i\theta = \theta_0 + \sum_{j=1}^{d} \theta_j x_j$

$P(y|X,\theta) \in (0,1)$

n: number of data point

θ: coefficients and intercept

The negative loglikelihood $J(\theta) = -\log P(y|X,\theta)$

$\nabla J(\theta) = \frac{dJ(\theta)}{d\theta} = \sum_{i=1}^{n} x_i^T(sigm(x_i\theta) - y_i) = X^T(sigm(X\theta) - y)$ $\quad X \in \mathbb{R}^{n\times d}$

$\theta \in \mathbb{R}^{d\times 1}$

$y \in \mathbb{R}^{n\times 1}$

$\nabla J(\theta) \in \mathbb{R}^{d\times 1}$

$H(\theta) = D(\nabla J(\theta))$

$= D(X^T(sigm(X\theta) - y))$

$= X^T diag[sigm(x_i\theta)\cdot(1-sigm(x_i\theta))]X \quad i=1,2,\ldots,n$

$\theta_{new} = \theta_{old} - \frac{\nabla J(\theta)}{H(\theta)}$

$= \theta_{old} - H^{-1}(\theta)\cdot\nabla J(\theta)$ $\quad$ Let $H(\theta) = X^T S X$ where $S = diag(sigm(x_i\theta)(1-sigm(x_i\theta)))$

$= \theta_{old} - (X^T S X)^{-1} X^T(\pi - y)$ $\quad$ Let $\nabla J(\theta) = X^T(\pi - y)$ where $\pi = sigm(X\theta)$

$= (X^T S X)^{-1}[(X^T S X)\theta_{old} + X^T(y-\pi)]$

$= (X^T S X)^{-1} X^T [S X \theta_{old} + y - \pi]$

θ_{new} is the solution of a weighted least square problem.

$WLS(\theta, S) = \frac{1}{n}\sum_{i=1}^{n} s_i (y_i - x_i\theta)^2 \quad s_i = sigm(x_i\theta)(1-sigm(x_i\theta))$

$\hat{\theta}_{WLS} = (X^T S X)^{-1} X^T [S X \theta_{old} + y - \pi]$

$= \theta_{new}$.

(b). If the maximum conditional likelihood estimator exist, let it be $\hat{\theta}$,

then $\nabla J(\hat{\theta}) = 0$.

$X^T(sigm(X\hat{\theta}) - y) = 0$.

$X \in \mathbb{R}^{n\times d}$

$\hat{\theta}, \theta \in \mathbb{R}^{d\times 1}$

Let θ be such that if $y_i = 0$, $x_i\theta < 0$

if $y_i = 1$, $x_i\theta > 0$.

Since $X^T(sigm(X\hat{\theta}) - y) = 0$,

$\theta^t X^T(sigm(X\hat{\theta} - y) = 0$.

$$\theta^t X^T(\operatorname{sigm}(X\hat{\theta})-y)) = [\theta^t x_1^t, \theta^t x_2^t, \dots, \theta^t x_n^t] \begin{bmatrix} \operatorname{sigm}(x_1\hat{\theta})-y_1 \\ \operatorname{sigm}(x_2\hat{\theta})-y_2 \\ \vdots \\ \operatorname{sigm}(x_n\hat{\theta})-y_n \end{bmatrix}$$

$$= \sum_{y_i=0} \theta^t x_i^t \cdot \operatorname{sigm}(x_i\hat{\theta}) + \sum_{y_i=1} \theta^t x_i^t (\operatorname{sigm}(x_i\hat{\theta})-1)$$

< 0 since $\theta^t x_i^t < 0$ when $y_i = 0$, $\operatorname{sigm}(x_i\hat{\theta}) > 0$, $\sum_{y_i=0} \theta^t x_i^t \operatorname{sigm}(x_i\hat{\theta}) < 0$

$\theta^t x_i^t > 0$ when $y_i = 1$, $\operatorname{sigm}(x_i\hat{\theta}) - 1 < 0$, $\sum_{y_i=1} \theta^t x_i^t (\operatorname{sigm}(x_i\hat{\theta})-1) < 0$

Contradicts that $\theta^t X^T(\operatorname{sigm}(X\hat{\theta})-y) = 0$.

So the maximum conditional likelihood estimator doesn't exist.

References: Problem 3(a) referenced from www.cs.ox.ac.uk/people/nando.defreitas/machinelearning
www.stat.cmu.edu/~cshalizi/mreg/15/lectures/24

Problem 3(b) referenced form Zihao Wang