Connor Realmuto

I pledge my honor that I have abided by the Stevens Honor System - CR

1.

Class → $w_1, w_2 \dots w_c$

a.) Bayes Theorem: $P(w_i|x) = \frac{P(x|w_i)P(w_i)}{P(x)}$ $i = 1, \dots, c$

where $P(x) = \sum_{i=1}^{c} P(x|w_i)P(w_i)$

Rule: choose w_i over w_j (where $i, j \in \{1, \dots, c\}$ & $i \neq j$) if the posterior probability of w_i is greater than w_j.

i.e. if $P(w_i|x) > P(w_j|x)$, choose w_i.

Using only the class conditional density & prior would yield:

$$\frac{P(x|w_i)\,P(w_i)}{P(x)} > \frac{P(x|w_j)P(w_j)}{P(x)}$$

Since $P(x)$ is the same for both it can be ignored.

∴ decide w_i if:

likelihood ratio →

$$P(x|w_i)P(w_i) > P(x|w_j)P(w_j)$$

or

$$\frac{P(x|w_i)}{P(x|w_j)} > \frac{P(w_j)}{P(w_i)}$$

where $i, j \in \{1, \dots, c\}$ & $i \neq j$

b.) w/ $p(x|w_1) = \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{x^2}{2}\right)}$

$p(x|w_2) = 1, x\left[-\frac{1}{2}, \frac{1}{2}\right]$

w/ loss taken into respect we decide based on minimizing the risk:

$$R(a_i|x) = \sum_{j=1}^{c} \lambda_{ij} P(w_j|x)$$

where λ_{ij} is the loss for taking action a_i (choose w_i) when the true state of nature is w_j

w/ 2 classes

$R(a_1|x) = \lambda_{11} P(w_1|x) + \lambda_{12} P(w_2|x)$

$R(a_2|x) = \lambda_{21} P(w_1|x) + \lambda_{22} P(w_2|x)$

w/ zero 1 loss $\rightarrow \lambda = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

$R(a_1|x) = P(w_2|x)$

$R(a_2|x) = P(w_1|x)$ $\rightarrow$ Decision rule $\rightarrow$ Choose w_1 if (minimize risk) $P(w_2|x) < P(w_1|x)$

This is the same result as part a.) i.e. we can use:

Choose $w_1 \rightarrow \frac{P(x|w_1)}{P(x|w_2)} > \frac{P(w_2)}{P(w_1)} \rightarrow P(x|w_1) > P(x|w_2)$

Since $P(x|w_1)$ is > 0 for all $x \rightarrow P(x|w_1) > P(x|w_2)$ when x is not in the interval $[-1/2, 1/2]$

Otherwise, choose w_1 if

$p(x|w_1) > 1$

$$\frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} > 1 \rightarrow -\frac{x^2}{2} > \ln(\sqrt{2\pi}) \rightarrow -x^2 > 2\ln(\sqrt{2\pi})$$

$$\rightarrow x > \left(-2\ln(\sqrt{2\pi})\right)^{1/2} \rightarrow p(x|w_1) \text{ is never equal to } 1.$$

$\therefore$ Overall decision rule:

Rule: $\begin{cases} \text{Choose } w_2, x \in \left[-\frac{1}{2}, \frac{1}{2}\right] \\ \text{Choose } w_1, \text{ Otherwise} \end{cases}$

2. a)

$$E_p(w) = \sum_{i=1}^{N} -\{ y_i \log f(x_i) + (1-y_i)\log(1-y_i)\log(1-f(x_i))\} + \frac{\lambda}{2}\|w\|^2$$

$$\Delta_w E_p(w) = \sum_{i=1}^{N} -\frac{d}{dw}\left(y_i \log(\sigma(w^T x_i)) + (1-y_i)\log(1-\sigma(w^T x_i))\right) + \frac{d}{dw}\frac{\lambda}{2}\|w\|^2$$

$$= -\left[\frac{y_i}{\sigma(w^T x_i)}\left(\frac{d\sigma(w^T x_i)}{dw}\right) + \frac{1-y_i}{(1-\sigma(w^T x_i))}\left(-\frac{d\sigma(w^T x_i)}{dw}\right)\right]$$

$$= -\left[\frac{y_i}{\sigma(w^T x_i)} - \frac{1-y_i}{1-\sigma(w^T x_i)}\right]\frac{d(\sigma(w^T x_i))}{dw}$$

$$= -\left[\frac{y_i}{\sigma(w^T x_i)} - \frac{1-y_i}{1-\sigma(w^T x_i)}\right]\sigma(w^T x_i)(1-\sigma(w^T x_i))\, x_i$$

$$= -\left[y_i(1-\sigma(w^T x_i)) - (1-y_i)(\sigma(w^T x_i))\right] x_i$$

$$= -y_i + y_i\sigma(w^T x_i) - \sigma(w^T x_i) + y_i\sigma(w^T x_i)$$

$$= -\left[y_i - \sigma(w^T x_i)\right] x_i$$

$$= \left(\sigma(w^T x_i) - y_i\right) x_i$$

$$\frac{d}{dw}\frac{\lambda}{2}\|w\|^2 = \lambda w$$

$\therefore$

$$\Delta_w E_p(w) = \sum_{i=1}^{N} (\sigma(w^T x_i) - y_i) x_i + \lambda w$$

b)

$$p(y|w) = \prod_{i=1}^{N} \sigma(w^T x_i)^{y_i} (1 - \sigma(w^T x_i))^{(1-y_i)}$$

$$p(w) = ?$$

$p(y|w)$ is equivalent to the above calculation w/o the negative. So if $\Delta_w \ln(p(w)) = \lambda w$ then you would get the same results as in part A if you maximize $p(y|w)p(w)$.

↓

(I was spending time on part 3 because I thought I needed a corrected result, I ran out of time to look back at this)

3.)

Instance	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
label	+	-	+	-	+	-	+	+
Data (x_1, x_2)	(10,10)	(0,0)	(8,4)	(3,3)	(4,8)	(0.5, 0.5)	(4,3)	(2,5)

Positive $\rightarrow t = 1$, negative $t = -1$

Start at $x_1 + x_2 = 0$ i.e. $w = [0\ 1\ 1]^T$ $x = [1\ x_1\ x_2]^T$

$t_i w^T x_i > 0 \rightarrow$ classification correct

$t=0$ $P(1)$, $x_1 = 10$, $x_2 = 10$, $t_1 = 1$

$$t_1 w^T x_1 = 1 \begin{bmatrix} 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 10 \\ 10 \end{bmatrix} = (10+10) = 20$$

$t_1 w^T x_1 > 0 \rightarrow$ keep $w \rightarrow w^1 = w^0$

$t=1$ $P(2)$ $x_1 = 0$, $x_2 = 0$, $t_2 = -1$

$$t_2 w^T x_2 = -1 \begin{bmatrix} 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = 0$$

$t_2 w^T x_2 \leq 0 \rightarrow$ update w

$$w^2 = w^1 + t_2 x_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}$$

$t=2$ $P(3)$ $x_1 = 8$, $x_2 = 4$, $t_3 = 1$

$$t_3 w^T x_3 = 1 \begin{bmatrix} -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 8 \\ 4 \end{bmatrix} = -1+8+4 = 11$$

$t_3 w^T x_3 > 0 \rightarrow$ keep $w \rightarrow w^3 = w^2$

$\gamma = 3$ $\quad P(4) \quad x_1 = 3, x_2 = 3, t_4 = -1$

$$t_4 w^T x_4 = -1 \begin{bmatrix} -1 & -1 & -1 \end{bmatrix} \begin{pmatrix} 1 \\ 3 \\ 3 \end{pmatrix} = -(-1-3-3) = 5$$

$t_4 w^T x_4 > 0 \Rightarrow$ update w

$$w^4 = \begin{bmatrix} -1 \\ -1 \\ -1 \end{bmatrix} - \begin{bmatrix} 1 \\ 3 \\ 3 \end{bmatrix} = \begin{bmatrix} -2 \\ -4 \\ -4 \end{bmatrix}$$

$\gamma = 4$ $\quad P(5) \quad x_1 = 4, x_2 = 8 \quad t_5 = 1$

$$t_5 w^T x_5 = 1 \begin{bmatrix} -2 & -4 & -4 \end{bmatrix} \begin{pmatrix} 1 \\ 4 \\ 8 \end{pmatrix} < 0 \Rightarrow \text{update } w$$

$$w^5 = \begin{bmatrix} -2 \\ -4 \\ -4 \end{bmatrix} + \begin{bmatrix} 1 \\ 4 \\ 8 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \\ 4 \end{bmatrix}$$

$\gamma = 5$ $\quad P(6) \quad x_1 = \frac{1}{2}, x_2 = \frac{1}{2} \quad t_6 = -1$

$$t_6 w^T x_6 = -1 \begin{bmatrix} -1 & 0 & 4 \end{bmatrix} \begin{pmatrix} 1 \\ \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} = -1(-1+0+2) < 0 \Rightarrow \text{update } w$$

$$w^6 = \begin{bmatrix} -1 \\ 0 \\ 4 \end{bmatrix} - \begin{bmatrix} 1 \\ \frac{1}{2} \\ \frac{1}{2} \end{bmatrix} = \begin{bmatrix} -2 \\ -\frac{1}{2} \\ \frac{7}{2} \end{bmatrix}$$

$\gamma = 6$ $\quad P(7) \quad x_1 = 4, x_2 = 3 \quad t_7 = 1$

$$t_7 w^T x_7 = 1 \begin{bmatrix} -2 & -\frac{1}{2} & \frac{7}{2} \end{bmatrix} \begin{pmatrix} 1 \\ 4 \\ 3 \end{pmatrix} = -2 - 2 + \frac{21}{2} > 0 \Rightarrow w^8 = w^7$$

$x=7$ $p(8)$ $x_1=2$ $x_2=5$ $t_8=1$

$t_8 w^t x_8 = 1 \begin{bmatrix} -2 & 1.5 & 5.5 \end{bmatrix} \begin{pmatrix} 1 \\ 2 \\ 5 \end{pmatrix} > 0 \rightarrow w^9 = w^8$

Final result $\rightarrow w = \begin{bmatrix} -2 & 1.5 & 5.5 \end{bmatrix}^T$

4.

a.) $p(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} k\left(\frac{x - x_i}{h}\right)$

$$k(u) = \begin{cases} 1, & |u| \le 1/2 \\ 0, & \text{o.w.} \end{cases}$$

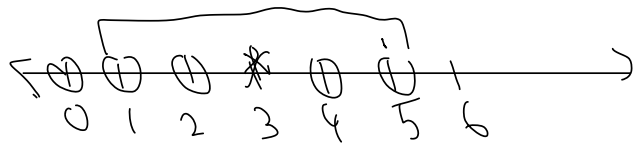
$$p(3) = \frac{1}{5} \sum_{i=1}^{5} \frac{1}{5} k\left(\frac{3 - x_i}{5}\right) \quad x_i \in \{0,1,2,4,5\}$$

$$\frac{1}{25}\left(k\left(\frac{3}{5}\right) + k\left(\frac{2}{5}\right) + k\left(\frac{1}{5}\right) + k\left(\frac{-1}{5}\right) + k\left(\frac{-2}{5}\right)\right)$$

$\frac{3}{5} > 1/2 = 0$ $\quad \frac{2}{5} < 1/2 = 1$ $\quad \frac{1}{5} < 1/2 = 1$ $\quad \frac{-1}{5} > -1/2 = 1$ $\quad \frac{-2}{5} > -1/2 = 1$

$$p(3) = \frac{1}{25}(4) = \frac{4}{25}$$

4 total around 3 w/ $h = 5$



b.) used Euclidian distance we can visually see which points are closest to X

i.) w/ k=1 → 1NN → x is from red class

↳ $\frac{k_{Red}}{k} = \frac{1}{14}$, $\frac{k_{Blue}}{k} = \frac{0}{14}$ $\frac{k_{Red}}{k} > \frac{k_{Blue}}{k}$

ii.) w/ k=7 → 7NN → x is from blue class

↳ $\frac{k_{Red}}{k} = \frac{3}{14}$, $\frac{k_{Blue}}{k} = \frac{4}{14}$, $\frac{k_{Blue}}{k} > \frac{k_{Red}}{k}$

iii.) I don't believe we should use k=9 since the other training samples are relatively far away from the test sample when compared to a 1NN or 7NN classifier, If k=9, the degree of smoothing is likely to be too large and the classifier could become inaccurate.

5.) $f(x|\theta) = 2\theta x e^{-\theta x^2}$ $x > 0$ $X \rightarrow \{x_1, x_2 \ldots x_n\}$

↑ Likelihood

$$f(x|\theta) = \prod_{i=1}^{n} 2\theta x_i e^{-\theta x_i^2}$$

Log-Likelihood $\rightarrow$ $\ln(f(x|\theta)) = \sum_{i=1}^{n} \ln(2\theta x_i e^{-\theta x_i^2})$

$$LL(\theta) = \sum_{i=1}^{n} (\ln(2\theta x_i) - \theta x_i^2)$$

$$\Delta_\theta LL(\theta) = \sum_{i=1}^{n} \left(\frac{2x_i}{2\theta x_i} - x_i^2\right) = \sum_{i=1}^{n} \left(\frac{1}{\theta} - x_i^2\right) = \frac{n}{\theta} - \sum_{i=1}^{n} x_i^2$$

Set $= 0$

$$\hookrightarrow \quad \frac{n}{\theta} - \sum_{i=1}^{n} x_i^2 = 0 \rightsquigarrow \frac{1}{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i^2$$

$$\text{MLE} \rightarrow \quad \theta = \frac{n}{\sum_{i=1}^{n} x_i^2} \rightarrow \frac{1}{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i^2$$

b.) $\underset{x_i}{\text{Unif}}(\theta_1, \theta_2)$ $\begin{cases} \frac{1}{\theta_2 - \theta_1}, & x_i \text{ between } \theta_1 \& \theta_2 \text{ for } x_i = \{x_1 \ldots x_n\} \\ 0, & 0, w. \end{cases}$

↓

$$X_i \sim Unif(\theta_1, \theta_2) = \prod_{i=1}^{n} \frac{1}{\theta_2 - \theta_1} \quad X_i \text{ b/t } \theta_1 \& \theta_2, 0 \text{ o.w.}$$

Likelihood $\rightarrow = \frac{1}{(\theta_2 - \theta_1)^n}$

$$LL(\theta) \rightarrow \ln\left((\theta_2 - \theta_1)^{-n}\right) = -n\left(\ln(\theta_2 - \theta_1)\right)$$

$$\Delta_{\theta_1} LL(\theta) = -n\left(\frac{1}{\theta_2 - \theta_1}\right)\frac{d}{d\theta_1}(\theta_2 - \theta_1) = \frac{n}{\theta_2 - \theta_1}$$

$$\Delta_{\theta_2} LL(\theta) = -n\left(\frac{1}{\theta_2 - \theta_1}\right)\frac{d}{d\theta_2}(\theta_2 - \theta_1) = \frac{-n}{\theta_2 - \theta_1}$$

For θ_1, $\Delta_{\theta_1} LL(\theta) \frac{n}{\theta_2 - \theta_1} \rightarrow$ Since $X_i \sim Unif[\theta_1, \theta_2]$ we maximize θ_1 by choosing the smallest X value

$\theta_1 = \min\{X\}$ $X \in \{X_1, X_2, \ldots X_n\}$

For θ_2, $\Delta_{\theta_2} LL\theta = \frac{-n}{\theta_2 - \theta_1}$, to maximize θ_2, we need to choose the largest X value, $\theta_2 = \max(X)$ $X \in \{X_1, X_2, \ldots X_n\}$

6.)

a.) The discriminative approach makes decisions by directly modeling the posterior probability of the classes, ie. $P(C_k|x)$. The generative approach focuses on modeling the class conditional densities $P(x|C_k)$ along with the priors $P(C_k)$. The discriminative function approach utilizes the creation of a function, $g(x)$ which maps a sample to a particular class/label

And uses Bayes Theory to find posterior.

b.) When performing MLE, the assumption is that the parameters are a fixed value, while Bayesian estimation assumes the parameter comes from an underlying distribution. This makes MLE computationally efficient while Bayes can possibly allow for more accurate results the prior information is reliable.

the best model is a weighted avg of models

c.) PCA & FLD are fairly similar techniques but with different goals. PCA is an unsupervised technique that aiming to reduce dimensionality without losing information. FLD is supervised and requires the class labels. This allows FLD to be used to find the best separation across

classes.

d.) Batch gradient descent is performed using all the training data at once, while mini-batch iterates through small portions of the data to perform updates. Stochastic gradient descent iterates through samples one at a time to make updates.

Batched GD is generally unbiased, accurate, and can converge to a global optimum. However, it is slow and unreliable w/ large datasets & can easily fall into traps w/ local modes.