Teresa Liu

CSCI 467 PS3

1) a) # data points in training set = 85 - 12 = 73
data points in test set = 12

b) The outcome Y has 3 classes, assuming that each paper has exactly one author. Those 3 classes are authored by Hamilton, Jay, or Madison. We can use a classification model to analyze the text of each paper and classify the paper into 1 of the 3 classes.

c) Formulating this problem into a multilabel problem, the labels would be authored by Hamilton, Jay, or Madison. These labels would be binary, meaning that the paper either has Y as an author or does not have Y as an author.

For a paper solely written by Jay, we label that paper as:

	Label (has as author)
Jay	Yes
Hamilton	No
Madison	No

For a paper written by Hamilton and Madison

	Label (has as author)
Jay	No
Hamilton	Yes
Madison	Yes

For a paper written by all 3 authors:

	Label (has as author)
Jay	Yes
Hamilton	Yes
Madison	Yes.

2) a) $Pr(Y = y^{(i)} \mid X = x^{(i)}) = \dfrac{1}{1+\exp(-y^{(i)}[\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)}$

$\rightarrow$ Show $Pr(Y = 1 \mid X = x^{(i)}) + Pr(Y = -1 \mid X = x^{(i)}) = 1$

$Pr(Y = 1 \mid X = x^{(i)}) = \dfrac{1}{1 + \exp(-\beta_0 - \beta_1 X_1 - \beta_2 X_2 \ldots - \beta_p X_p}$

$Pr(Y = -1 \mid X = x^{(i)}) = \dfrac{1}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p}$

$$\Rightarrow \frac{1}{1 + e^{-\beta_0 - \beta_1 X_1 \cdots - \beta_p X_p}} + \frac{1}{1 + e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p}}$$

$$= \frac{1 + e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p} + 1 + e^{-\beta_0 - \beta_1 X_1 \ldots - \beta_p X_p}}{(1 + e^{-\beta_0 - \beta_1 X_1 \cdots - \beta_p X_p})(1 + e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p})}$$

$$= \frac{2 + e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p} + e^{-\beta_0 - \beta_1 X_1 \ldots \beta_p X_p}}{1 + e^{-\beta_0 - \beta_1 X_1 \ldots - \beta_p X_p} + e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p} + \underbrace{(e^{-\beta_0 - \beta_1 X_1 \ldots - \beta_p X_p}) \cdot (e^{+\beta_0 + \beta_1 X_1 \ldots + \beta_p X_p})}}$$

$$e^{-\beta_0} e^{-\beta_1 X_1} \ldots e^{-\beta_p X_p} e^{\beta_0} e^{\beta_1 X_1} \ldots e^{\beta_p X_p}$$

$$= e^{-\beta_0 + \beta_0 - \beta_1 X_1 + \beta_1 X_1 - \ldots - \beta_p X_p + \beta_p X_p}$$

$$= e^0 = 1$$

$$\Rightarrow \frac{2 + e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p} + e^{-\beta_0 - \beta_1 X_1 \ldots - \beta_p X_p}}{2 + e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p} + e^{-\beta_0 - \beta_1 X_1 - \ldots - \beta_p X_p}} = \boxed{1}$$

//

2) b) $z^{(i)} = -y^{(i)}[\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p]$

$$Pr(Y = y^{(i)} \mid X = x^{(i)}) = \frac{1}{1+\exp(z^{(i)})}$$

① approaches 0 as $\exp(z^{(i)}) \to \infty$
this case happens as $z^{(i)} \to \infty$ which occurs
when X values move towards ∞ or the given
y value increases towards ∞.

② approaches 1 as $\exp(z^{(i)}) \to 0$
this case happens as $z^{(i)} \to -\infty$ since $e^{-\infty} = 0$
which occurs when X values move toward $-\infty$
or the y value moves towards $-\infty$

c) Decision Boundary occurs when
$Pr(Y = 1 \mid X = x^{(i)}) = 0.5$

$$\frac{1}{1+\exp(z^{(i)})} = \frac{1}{2}$$

$$2 = 1 + \exp(z^{(i)})$$
$$\Rightarrow 1 = \exp(z^{(i)})$$
$$\Rightarrow \log(1) = z^{(i)}$$
$$\Rightarrow 0 = z^{(i)} = -y^{(i)}[\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p]$$

when $y = 1$

$$z^{(i)} = -\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p = 0$$

to classify as $y = 1$,

$$z^{(i)} = -\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p > 0$$

3) $f_k(x) = \frac{1}{x\sqrt{2\pi}\,\sigma_k} e^{\left(\frac{-(\ln x - \mu_k)^2}{2\sigma_k^2}\right)}$, $x \geq 0$

$Pr(Y=k \mid X=x) = p_k(x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$

a) k^* is the class with max $p_k(x)$

$k^* = \text{argmax } p_k(x)$

$\Rightarrow \underset{k}{\text{argmax}}\ \pi_k \frac{1}{x\sqrt{2\pi}\,\sigma_k} e^{\left(\frac{-(\ln x - \mu_k)^2}{2\sigma_k^2}\right)}$

$= \underset{k}{\text{argmax}}\ \log \pi_k - \log x - \cancel{\log\sqrt{2\pi}} - \cancel{\log \sigma_k} - \frac{(\ln x - \mu_k)^2}{2\sigma_k^2}$

$= \underset{k}{\text{argmax}}\ \log \pi_k - \log x - \frac{(\ln x)^2 - 2\mu_k \ln x + \mu_k^2}{2\sigma_k^2}$

$= \text{argmax } \log \pi_k - \log x - \frac{(\ln x)^2}{2\sigma_k^2} - \frac{\mu_k \ln x}{\sigma_k^2} + \frac{\mu_k^2}{2\sigma_k^2}$

$\delta(x) = \log \pi_k - \log x - \frac{(\ln x)^2}{2\sigma_k^2} - \frac{\mu_k \ln x}{\sigma_k^2} + \frac{\mu_k^2}{2\sigma_k^2}$

not linear because of $(\ln x)^2$ term

b) $K = 2$ $\pi_1 = \pi_2$ show $x^* = \sqrt{\exp(\mu_1)}\sqrt{\exp(\mu_2)}$

$\Rightarrow \cancel{\log \pi_1} - \cancel{\log x} - \frac{(\ln x)^2}{2\sigma_1^2} - \frac{\mu_1 \ln x}{\cancel{\sigma_1^2}} + \frac{\mu_1^2}{\cancel{2\sigma_1^2}} = \cancel{\log \pi_2} - \cancel{\log x} - \frac{(\ln x)^2}{2\sigma_2^2} - \frac{\mu_1 \ln x}{\sigma_1^2} + \frac{\mu_1^2}{\cancel{2\sigma_1^2}}$

$-\frac{1}{2}(\cancel{\ln x})^2 - \mu_1 \ln x + \frac{1}{2}\mu_1^2 = -\frac{1}{2}(\cancel{\ln x})^2 - \mu_2 \ln x + \frac{1}{2}\mu_2^2$

$\Rightarrow \frac{1}{2}\mu_1^2 - \frac{1}{2}\mu_2^2 = \mu_1 \ln x - \mu_2 \ln x$

$\Rightarrow \frac{1}{2}(\mu_1^2 - \mu_2^2) = \ln x(\mu_1 - \mu_2)$

$\Rightarrow \frac{1}{2}\cancel{(\mu_1 - \mu_2)}(\mu_1 + \mu_2) = \ln x \cancel{(\mu_1 - \mu_2)}$

$\Rightarrow \frac{1}{2}(\mu_1 + \mu_2) = \ln x$

$x = e^{\frac{1}{2}(\mu_1 + \mu_2)}$

$x = \sqrt{e^{(\mu_1 + \mu_2)}} = \sqrt{e^{\mu_1} e^{\mu_2}}$

$\boxed{x = \sqrt{\exp(\mu_1)}\sqrt{\exp(\mu_2)}}$

c) $x = \sqrt{\exp(1)}\sqrt{\exp(5)}$, $\frac{1}{x\sqrt{2\pi}} e^{\frac{-(\ln x - 1)^2}{2}}$, $\frac{1}{x\sqrt{2\pi}} e^{\frac{-(\ln x - 5)^2}{2}}$

$$f_k(x_i) = Pr(x_i = x \mid y = k) = \frac{1}{\sqrt{2\pi}\,\sigma_{ik}} e^{\left(-\frac{(x_i - \mu_{ik})^2}{2\sigma_{ik}}\right)}$$

4) a) $\pi_1 = \pi(\text{No}) = 7/10$

$\pi_2 = \pi(\text{Yes}) = 3/10$

b) for continuous feature "Taxable Income"

$Pr(\text{Income} \mid \text{Evade} = \text{No})$	$Pr(\text{Income} \mid \text{Evade} = \text{Yes})$
$\mu = 109.143$	$\mu = 91$
$s^2 = 2539.476$	$s^2 = 13$
$f_k(x) = \frac{1}{\sqrt{2\pi \cdot 2539.476}} \cdot e^{\left(\frac{-(x-109.143)^2}{2\sqrt{2539.476}}\right)}$	$f_k(x) = \frac{1}{\sqrt{2\pi \cdot 13}} \cdot e^{\left(\frac{-(x-91)^2}{2\sqrt{13}}\right)}$

c) for discrete feature "Refund"

		Laplace
$Pr(\text{Refund} = \text{Yes} \mid \text{Evade} = \text{No})$	$= 3/7$	$4/9$
$Pr(\text{Refund} = \text{No} \mid \text{Evade} = \text{No})$	$= 4/7$	$5/9$
$Pr(\text{Refund} = \text{Yes} \mid \text{Evade} = \text{Yes})$	$= 0$	$1/5$
$Pr(\text{Refund} = \text{No} \mid \text{Evade} = \text{Yes})$	$= 1$	$2/5$

valid PMF

for discrete feature "Marital Status"

$Pr(\text{MS} = \text{Single} \mid \text{Evade} = \text{No})$	$= 2/7$	$3/10$
$Pr(\text{MS} = \text{Married} \mid \text{Evade} = \text{No})$	$= 4/7$	$5/10$
$Pr(\text{MS} = \text{Divorced} \mid \text{Evade} = \text{No})$	$= 1/7$	$2/10$
$Pr(\text{MS} = \text{Single} \mid \text{Evade} = \text{Yes})$	$= 2/3$	$3/6$
$Pr(\text{MS} = \text{Married} \mid \text{Evade} = \text{Yes})$	$= 0$	$1/6$
$Pr(\text{MS} = \text{Divorced} \mid \text{Evade} = \text{Yes})$	$= 1/3$	$2/6$

valid PMF

d) Laplace correction solves the problem of zero probability

Does not create a valid PMF since $\Sigma \neq 1$.

5)

	AN	AP
PN	8826	30
PP	23	456

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{456}{456+23} = 0.9520$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{456}{456+30} = 0.9383$$

$$TNR = \frac{TN}{TN+FP} = \frac{8826}{8826+23} = 0.9974$$

$$NPV = \frac{TN}{TN+FN} = \frac{8826}{8826+30} = 0.9966$$

$$\text{F1 score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = 0.9451$$

4e) (Yes, Single, 110k)

$$P(Y=Yes \mid X) = \frac{P(\text{Single} \mid Y=Yes)\, P(110k \mid Y=Yes)\, P(Yes \mid Y=Yes)}{P(Yes)}$$

$$P(Y=No \mid X) = \frac{P(\text{Single} \mid Y=No)\, P(110k \mid Y=No)\, P(Yes \mid Y=No)}{P(No)}$$

$$P(Y=Yes \mid X) \propto \frac{3}{6} \cdot \left(\frac{1}{\sqrt{2\pi \cdot 13}} e^{\frac{-(110-91)^2}{2(13)}} \right) \cdot \frac{1}{5}$$

$$\frac{3}{6} \cdot 1.0326 \times 10^{-7} \cdot \frac{1}{5} = 1.0326 \times 10^{-8}$$

$$P(Y=No \mid X) \propto \frac{3}{10} \cdot \left(\frac{1}{\sqrt{2\pi \cdot 50.393}} e^{\frac{-(110-109.43)^2}{2(50.393)}} \right) \cdot \frac{4}{9}$$

$$\frac{3}{10} \cdot 0.00715 \cdot \frac{4}{9} = 0.00106$$

$$P(Y=No \mid X) > P(Y=Yes \mid X)$$

So classification is $\boxed{\text{No}}$

4e) (No, Divorced, 93k)

$$P(Y = Yes \mid X) = \frac{P(\text{Divorced} \mid Y = Yes)\, P(93k \mid Y = Yes)\, P(No \mid Y = Yes)}{P(Yes)}$$

$$P(Y = No \mid X) = \frac{P(\text{Divorced} \mid Y = No)\, P(93k \mid Y = No)\, P(No \mid Y = No)}{P(No)}$$

$$P(Y = Yes \mid X) \propto \frac{2}{6} \cdot \left(\frac{1}{\sqrt{2\pi 13}}\, e^{-\frac{(93-91)^2}{13}}\right) \cdot \frac{2}{5}$$

$$= \frac{2}{6} \cdot 0.09487 \cdot \frac{2}{5} = 0.0126$$

$$P(Y = No \mid X) = \frac{2}{10} \cdot \left(\frac{1}{\sqrt{2\pi \cdot 50.393}}\, e^{-\frac{(93-109.143)^2}{2(50.393)}}\right) \cdot \frac{5}{9}$$

$$= \frac{2}{10} \cdot 0.00752 \cdot \frac{5}{9} = 0.0008356$$

$P(Y = Yes \mid X) > P(Y = No \mid X)$

So classification is $\boxed{Yes}$

3c)
Decision Boundary at x = 20