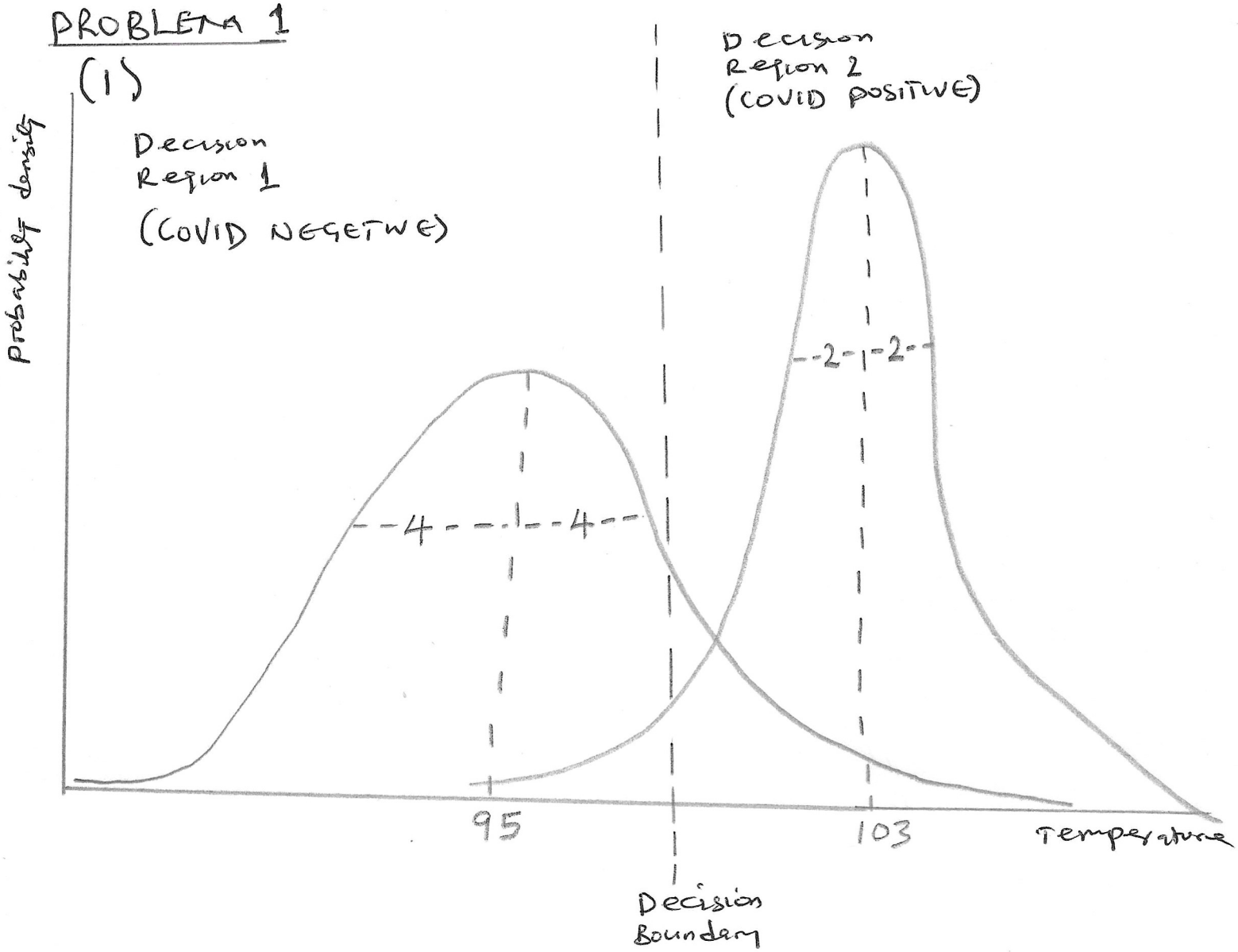
COSC 522 HOMEWORK 1 (FALL 2024

STUDENT NAME: JOSEPH OCHE AGADA

NET ID: JAGADA

PROBLEM 1

(1)



(2)

a. Given $x = 100$

$$P(w_j/x=100) = \frac{P(x=100/w_j)P(w_j)}{\sum_{j=1}^{2} P(x=100/w_j)P(w_j)}$$

But here,

$$\sum_{j=1}^{2} P(x=100/w_j)P(w_j) = P(x=100/w_1)P(w_1) + P(x=100/w_2)P(w_2)$$

$$= \frac{1}{\sigma_1\sqrt{2\pi}} e^{-(100-\mu_1)^2/2\sigma_1^2} \times P(w_1) + \frac{1}{\sigma_2\sqrt{2\pi}} e^{-(100-\mu_2)^2/2\sigma_2^2} \times P(w_2)$$

Since

$\sigma_1 = 4, \sigma_2 = 2, \mu_1 = 95, \mu_2 = 103, P(w_1) = P(w_2) = 0.5$

We have

$$\sum_{j=1}^{2} P(x=100|w_j)P(w_j) = \frac{1}{4\sqrt{2\pi}} e^{-(100-95)^2/2\times 16} \times 0.5 + \frac{1}{2\sqrt{2\pi}} e^{-(100-103)^2/2\times 4} \times 0.5$$

$$= 0.0457 \times 0.5 + 0.0647 \times 0.5$$

$$= \underline{0.0552}$$

Hence

$$P(w_1|x=100) = \frac{P(x=100|w_1)P(w_1)}{\sum_{j=1}^{2} P(x=100|w_j)P(w_j)}$$

$$= \frac{0.0457 \times 0.5}{0.0552} = \underline{0.4139}$$

$$P(w_2|x=100) = \frac{P(x=100|w_2)P(w_2)}{\sum_{j=1}^{2} P(x=100|w_j)P(w_j)}$$

$$= \frac{0.0647 \times 0.5}{0.0552} = 0.58605$$

Since $P(w_2|x=100) = 0.58605 > P(w_1|x=100) = 0.4139$

We conclude that the patient with temperature $x = 100$ belong to class 2. That implies that the patient is predicted to be COVID-19 positive. This is according to the Maximum posterior probability (MAPP) optimization method.

b. The decision boundary is the value of the predictor variable (temperature) x where posterior probability of class 1 equals the posterior probability of class 2. Since the prior probabilities are assumed to be these, the decision boundary is the value of x where

$$p(x|w_1) = p(x|w_2)$$

but we have

$\mu_1 = 95$, $\mu_2 = 103$, $\sigma_1 = 4$ and $\sigma_2 = 2$

We find the value of x where

$$p(x|w_1) = p(x|w_2)$$

$$\Rightarrow \frac{1}{\sigma_1\sqrt{2\pi}} e^{-(x-\mu_1)^2/2\sigma_1^2} = \frac{1}{\sigma_2\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma_2^2}$$

$$\Rightarrow \frac{1}{4\sqrt{2\pi}} e^{-(x-95)^2/2(4)^2} = \frac{1}{2\sqrt{2\pi}} e^{-(x-103)^2/2(2)^2}$$

Taking log of both sides and simplifying gives

$$\ln\left(\frac{1}{2}\right) = \frac{(x-95)^2}{32} - \frac{(x-103)^2}{8}$$

that is, we find the value of x where

$$\ln(1) - \ln(2) = \frac{(x-95)^2}{32} - \frac{(x-103)^2}{8}$$

$$\Rightarrow 0 - 0.6931 = \frac{(x-95)^2}{32} - \frac{(x-103)^2}{8}$$

$$\Rightarrow \frac{(x-95)^2}{32} - \frac{(x-103)^2}{8} + 0.6931 = 0$$

We divide through by 32 to eliminate the denominator

$$(x-95)^2 - 4(x-103)^2 + 32(0.6931) = 0$$

$$\Rightarrow x^2 - 190x + 9025 - 4(x^2 - 206x + 10609) + 22.1792 = 0$$

$\Rightarrow x^2 - 190x + 9025 - 4x^2 + 824x - 42436 + 22.1792 = 0$

$\Rightarrow -3x^2 + 634x - 33411 + 22.1792 = 0$

$\Rightarrow -3x^2 + 634x - 33388.8208 = 0$

We solve the quadratic equation using

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

where $a = -3$, $b = 634$ and $c = -33388.8208$

Hence

$$x = \frac{-634 \pm \sqrt{634^2 - 4(-3)(-33388.8208}}{2(-3)}$$

$$x = \frac{-634 \pm \sqrt{401956 - 400665.8496}}{-6}$$

$$= \frac{-634 \pm \sqrt{1290.1504}}{-6}$$

$$x = \frac{-634 \pm 35.92}{-6}$$

$$x = \frac{-634 + 35.92}{-6} \quad \text{or} \quad \frac{-634 - 35.92}{-6}$$

$$x = 99.68 \quad \text{or} \quad 111.65$$

Since this problem led to a quadratic equation, we have 2 possible values for x but we chose $x = 99.68$ because it is more graphically likely and correct considering the parameters of the Gaussian/Normal distribution.

c. Here, we use the decision boundary obtained analytically to solve for the overall probability of error.

The overall probability of error

$$P(\text{error}) = P(\text{error}|w_1) + P(\text{error}|w_2)$$

If $x > 99.68$, we should predict class 2, if we predict class 1 in this case, we have committed an error and this error can be calculated using

$P(x > 99.68|w_1)$ and this is $P(\text{error}|w_1)$

So

$$P(\text{error}|w_1) = P(x > 99.68|w_1)$$

Similarly

$$P(\text{error}|w_2) = P(x < 99.68|w_2)$$

but

$$P(\text{error}|w_1) = P(x > 99.68|w_1) = P\left(Z > \frac{99.68 - 95}{4}\right)$$

since $\mu_1 = 95$ and $\sigma_1 = 4$.

Note: we standardized the value of x to enable us use standard normal distribution table to obtain the probability values.

Hence

$$P(\text{error}|w_1) = P(Z > 1.170) = \underline{0.121}$$

Similarly

$$P(\text{error}|w_2) = P(x < 99.68|w_2)$$

$$= P\left(Z < \frac{99.68 - 103}{2}\right)$$

$$= P(Z < -1.66) = 0.0485$$

Hence

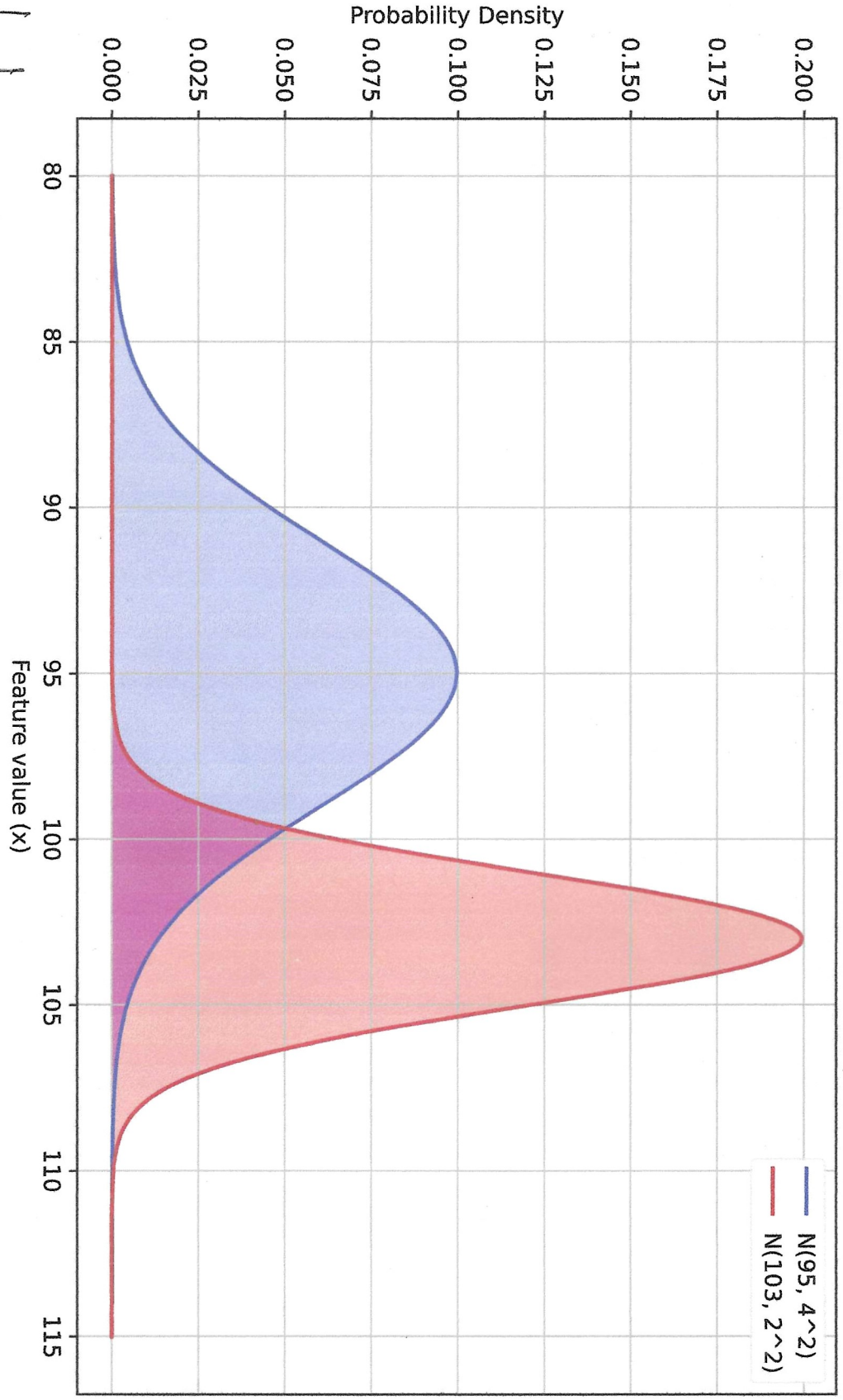
$$P(\text{error}) = P(\text{error}|w_1) + P(\text{error}|w_2) = 0.121 + 0.0485$$

$$= \underline{\underline{0.1695}}$$

PROBLEM 1

(3)

a.


Gaussian Distributions for Class 1 and 2; PROBLEM 1 (3)a.
Probability Density
Feature value (x)
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
80
85
90
95
100
105
110
115
N(95, 4^2)
N(103, 2^2)

(3) a. check the plot above

b. Given that $P(w_1) = 0.8$ and $P(w_2) = 0.2$

$$P(x=100) = \sum_{j=1}^{2} P(x=100|w_j)P(w_j)$$

$$= P(x=100|w_1)P(w_1) + P(x=100|w_2)P(w_2)$$

$$= \frac{1}{4\sqrt{2\pi}} e^{-(100-95)^2/2\times16} \times 0.8 + \frac{1}{2\sqrt{2\pi}} e^{-(100-103)^2/2\times4} \times 0.2$$

$$= 0.0457 \times 0.8 + 0.0647 \times 0.2$$

$$= 0.03656 + 0.01294$$

$$= 0.0495$$

Hence

$$P(w_1|x=100) = \frac{P(x=100|w_1)P(w_1)}{P(x=100)} = \frac{\frac{1}{4\sqrt{2\pi}} e^{-(100-95)^2/2\times16} \times 0.8}{0.0495}$$

$$= \underline{0.7386}$$

Similarly

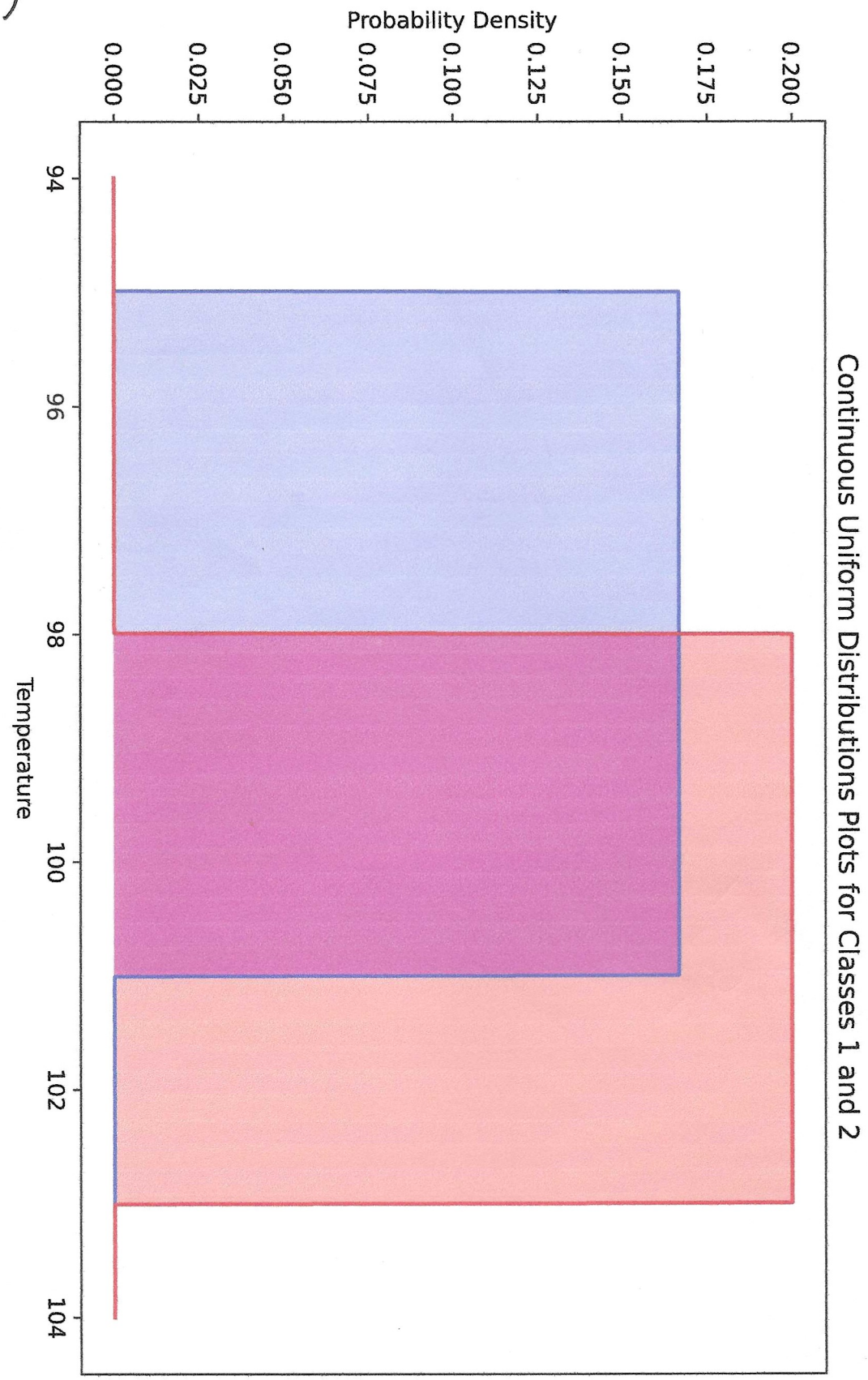
$$P(w_2|x=100) = \frac{P(x=100|w_2)P(w_2)}{P(x=100)} = \frac{\frac{1}{2\sqrt{2\pi}} e^{(100-103)^2/2\times4} \times 0.2}{0.0495}$$

$$= \frac{0.0647 \times 0.2}{0.0495} = \underline{0.2614}$$

With $P(w_1) = 0.8$ & $P(w_2) = 0.2$
the prediction for patient with temperature $x = 100$ is now class 1 (No COVID) using the maximum Baobab Posterior probability optimization method.

PROBLEM 2

(1)


Continuous Uniform Distributions Plots for Classes 1 and 2
Probability Density
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Temperature
94
96
98
100
102
104

PROBLEM 2

(2) For class 1

$X_1 \sim U(95, 101)$ (W_1)

for class 2

$X_2 \sim U(98, 103)$ (W_2)

$P(W_1) = P(W_2) = 0.5$ (equal prior probability)

Negative class is Class 1, false negative means predicting negative class when the actual class is positive class - that is

$P(X > 99 | W_1)$ (Probability of predicting class 1 ~~when~~ when we should predict class 2)

$$P(X > 99 | W_1) = \int_{99}^{101} f(x)\, dx$$

for W_1, $f(x) = \frac{1}{101-99}$

Hence

$$P(X > 99 | W_1) = \int_{99}^{101} \frac{1}{101-99} dx = \int_{99}^{101} \frac{1}{6} dx = \frac{1}{6}\int_{99}^{101} dx$$

$$= \frac{x}{6}\Big|_{99}^{101} = \left(\frac{101-99}{6}\right) = \underline{0.333}$$

Hence probability of false negative is 0.333

(3) positive class is class 2, false positive means predicting positive class when the actual class is negative. that is

$$P(X < 99 | W_2) = \int_{98}^{99} f(x) dx = \int_{98}^{99} \frac{1}{103-98} dx$$

$$= \int_{98}^{99} \frac{1}{5} dx = \frac{x}{5}\Big|_{98}^{99} = \frac{99-98}{5} = \frac{1}{5} = 0.20$$

so $P(X < 99 | W_2) = 0.20$

(4) The optimal decision boundary that minimizes the overall probability of error in the Bayesian sense is 99.68

(5) Yes, with more accurate domain knowledge (prior probability value) the optimum probability of error can be improved upon.

PROBLEM 3

On THE PAPER TITLED "STATISTICAL MODELING: THE 2 CULTURES" BY PROFESSOR LEO BREIMAN

In this paper, the author examined the problems associated with the over-insistence of core statisticians on data models. According to the paper, 98% of statisticians are in the data modeling culture while only 2% are in the algorithmic model community. The author argued that the focus of the statistical community on data models has led to irrelevant theories, kept statisticians from taking advantage of algorithmic models, and prevented them from working on exciting new projects. The author's opinion and interest in this discourse emanated from his initial academic experience before he moved into consulting, his experience as a consultant in the industry, and his experience at UC, Berkeley after he returned to the academic community. Based on this movement between academia and industry, the author was able to experience both cultures and form an informed opinion.

The author discussed a few of the projects he executed while in the industry as a consultant and how these projects shaped his cultural beliefs. He emphasized his Ozone Prediction Project and Chlorine Project. The Zone Prediction Project, although failed, the author believes the project will be a success if revisited today considering the now available sophisticated algorithmic models. Following the author's experience in the industry as a consultant, he went back into the academic community with the perceptions that consultants focus on finding solutions, searching for models that give better solutions, be it algorithmic or data models, and accuracy of prediction of test data as criteria for assessing model. On the other hand, statisticians in the academic community continued to emphasize the goodness of fit and residual assessment of data models as the determinant of model performance.

The author identified some of the common problems with data model which are failure of the goodness of fit test and residual analysis in model assessment. The failure of goodness of fit test is due to the yes/no nature of the test while residual analysis also fail when data dimension is more the 4 to 5. The author agreed that predictive accuracy which is the parameter for measuring the performance of the algorithmic models is most appropriate for measuring the performance of models.

Finally, the author pointed out the increasing popularity and applicability of algorithmic model especially with the birth of neural networks and decision tree. Emphases of these algorithmic models are in achieving predictive accuracy and this accuracy is measured by cross-validation which the author believe is more effective than its equivalent in the data modeling community. The popularity of algorithmic models has continued to grow with the recent emergence of ensemble techniques like random forest which involve the combination of the predictive capability of multiple models to improve model performance. The author wants statisticians to reconsider their preference for data models.