4. The following data set was collected to classify people who evade taxes:

Tax ID	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	122 K	No
2	No	Married	77 K	No
3	No	Married	106 K	No
4	No	Single	88 K	Yes
5	Yes	Divorced	210 K	No
6	No	Single	72 K	No
7	Yes	Married	117 K	No
8	No	Married	60 K	No
9	No	Divorced	90 K	Yes
10	No	Single	85 K	Yes

Considering relevant features in the table (only one feature is not relevant), assume that the features are *conditionally independent.* (25 pts)

(a) Estimate prior class probabilities.

(b) For continuous feature(s), assume conditional Gaussianity and estimate class conditional pdfs $p(x|\omega_i)$. Use Maximum Likelihood Estimates.

(c) For each discrete feature X, assume that the number of instances in class ω_i for which $X = x_j$ is n_{ji} and the number of instances in class ω_i is n_i. Estimate the probability mass $p_{X|\omega_i}(x_j|\omega_i) = P(X = x_j|\omega_i)$ as n_{ji}/n_i for each discrete feature. Is this a valid estimate of the pmf?

(a) $P(\text{evade}) = 3/10$

$P(\text{not evade}) = 7/10$

(b) continuous feature: Taxable Income

$$p(x|w_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2\right)$$

$\mu(\text{evade}) = (88+90+85)/3 \approx 87.67\text{K}$

$\sigma^2(\text{evade}) = [(88-87.67)^2 + (90-87.67)^2 + (85-87.67)^2]/3 = (0.1089 + 5.4289 + 7.1289)/3 \approx 4.22\text{K}$

$\mu(\text{not evade}) = (122+77+106+210+72+117+60)/7 \approx 109.14\text{K}$

$\sigma^2(\text{not evade}) = [(122-109.14)^2 + (77-109.14)^2 + (106-109.14)^2 + (210-109.14)^2 + (72-109.14)^2 + (117-109.14)^2 + (60-109.14)^2]/7$

$= (165.3796 + 1032.9796 + 9.8596 + 10172.7396 + 1379.3796 + 61.7796 + 2414.7396)/7$

$\approx 2176.69\text{K}$

$$\therefore p(x|w_{evade}) = \frac{1}{\sqrt{2\pi}\times 4.22\text{K}} \exp\left(-\frac{1}{2}\frac{(x-87.67\text{K})^2}{4.22\text{K}}\right)$$

$$p(x|w_{not\ evade}) = \frac{1}{\sqrt{2\pi}\times 2176.69\text{K}} \exp\left(-\frac{1}{2}\frac{(x-109.14\text{K})^2}{2176.69\text{K}}\right)$$

(c) For "Refuse": $P(X=\text{Yes}|w_{evade}) = 0/3 = 0$

$P(X=\text{No}|w_{evade}) = 3/3 = 1$

$P(X=\text{Yes}|w_{not\ evade}) = 3/7$

$P(X=\text{No}|w_{not\ evade}) = 4/7$

For "Marital Status": $P(X=\text{Single}|w_{evade}) = 2/3$

$P(X=\text{Married}|w_{evade}) = 0$

$P(X=\text{Divorced}|w_{evade}) = 1/3$

$P(X=\text{Single}|w_{not\ evade}) = 2/7$

$P(X=\text{Married}|w_{not\ evade}) = 4/7$

$P(X=\text{Divorced}|w_{not\ evade}) = 1/7$

It's not a valid estimate:

$\because P(X=\text{Yes}|w_{evade}) = 0$, $P(X=\text{Married}|w_{evade}) = 0$, these can cause the posterior equals to 0

also these are due limited samples. they cannot present true possibilities.

(d) There is an issue with using the estimate you calculated in 4c. Explain why the laplace correction $(n_{ji}+1)/(n_i+l)$, where l is the number of levels X can assume,[1] solves the problem with the estimate given in 4c. Is this a valid estimate of the pmf?

(e) Estimate the minimum error rate decision rule for classifying tax evasion using Laplace correction.

(d) For "Refuse": $P(X=\text{Yes}\mid W_{evade}) = (0+1)/(3+2) = 1/5$
$P(X=\text{No}\mid W_{evade}) = (3+1)/(3+2) = 4/5$
$P(X=\text{Yes}\mid W_{not\ evade}) = (3+1)/(7+2) = 4/9$
$P(X=\text{No}\mid W_{not\ evade}) = (4+1)/(7+2) = 5/9$

For "Marital Status": $P(X=\text{Single}\mid W_{evade}) = (2+1)/(3+3) = 3/6 = 1/2$
$P(X=\text{Married}\mid W_{evade}) = (0+1)/(3+3) = 1/6$
$P(X=\text{Divorced}\mid W_{evade}) = (1+1)/(3+3) = 2/6 = 1/3$
$P(X=\text{Single}\mid W_{not\ evade}) = (2+1)/(7+3) = 3/10$
$P(X=\text{Married}\mid W_{not\ evade}) = (4+1)/(7+3) = 5/10 = 1/2$
$P(X=\text{Divorced}\mid W_{not\ evade}) = (1+1)/(7+3) = 2/10 = 1/5$

Valid

(e) $g_i(x) = P(w_i\mid x) = \dfrac{P(x\mid w_i)P(w_i)}{P(x)}$

$\therefore g_i(x) = P(x\mid w_i)P(w_i)$

$= P(x_1, x_2, x_3\mid w_i)P(w_i)$

$= P(x_1\mid w_i)P(x_2\mid w_i)P(x_3\mid w_i)P(w_i)$

$\therefore g_{evade}(x) = P(\text{Refund}\mid W_{evade})P(\text{Marital Status}\mid W_{evade})P(\text{Taxable Income}\mid W_{evade})P(W_{evade})$

$g_{not\ evade}(x) = P(\text{Refund}\mid W_{not\ evade})P(\text{Marital Status}\mid W_{not\ evade})P(\text{Taxable Income}\mid W_{not\ evade})P(W_{not\ evade})$

choose $g_{evade}(x) \underset{\text{not evade}}{\overset{\text{evade}}{\gtrless}} g_{not\ evade}(x)$