EECE 5644 Intro to Machine Learning Midterm

Problem 1 - Maximum Likelihood Estimation

In this problem, $Y = B + 1$ with probability v, and $Y = B$ with probability $1 - v$, where v is known. You must estimate B given Y.

a.) Show that the likelihood function is given by $f_{Y|B}(y|B) = 1 - v$, for $B = y$, and $f_{Y|B}(y|B) = v$ for $B = y - 1$.

Problem 1)

Maximum likelihood estimation. In this problem, $Y = B+1$ with $Pr(v)$ and $Y = B$ with $Pr(1-v)$, where v is known. You must estimate B given Y

a) Show that likelihood function is given by

$f_{Y|B}(y|B) = 1-v$ for $B = y$ and

$f_{Y|B}(y|B) = v$ for $B = y-1$

	B	$B+1$
Y	$1-v$	v

- $Y_1 = B+1$ and $P(Y_1) = v$
- $Y_2 = B$ and $P(Y_2) = 1-v$

$$f(B|Y) = \frac{1-v}{\cancel{v}+1-\cancel{v}} = 1-v$$

$$f(B+1|Y) = \frac{v}{\cancel{v}+1-\cancel{v}} = v$$

$$f_{Y|B}(y|B) = \frac{P(B=y|y=B)P(y=B)}{\Sigma P(B=y|y=B)P(y=B)} = \frac{(1)(1-v)}{(1)(1-\cancel{v}) + \cancel{(1)}(\cancel{v})} = \boxed{1-v}$$

$$f_{Y|B}(y|B) = \frac{P(B=y-1|y=B+1)P(y=B+1)}{\Sigma P(B=y-1|y=B+1)P(y=B+1)} = \frac{(1)(v)}{(1)(1-\cancel{v}) + (1)(\cancel{v})} = \boxed{v}$$

b.) For $v = 1/4$, find the Maximum Likelihood Estimator for B using $Y = y$

b.) For $v = 1/4$, find maximum likelihood estimator for B using $Y = y$

$$p(Y|B) = \prod_{k=1}^{n} p(y_k|B)$$

$$L(B) = \prod_{k=1}^{n} f(y_k|B)$$

$$\hat{B} = \underset{B}{\arg\max}\left(\prod_{k=1}^{n} f(y_k|B)\right)$$ or can take log-likelihood

$$\hat{B} = \underset{B}{\arg\max}\left(\sum_{k=1}^{n} \ln f(y_k|B)\right)$$

$$= \underset{B}{\arg\max}\left(\ln(1-v), \ln(v)\right)$$

$$\boxed{\hat{B} = 3/4}$$

$$= \underset{B}{\arg\max}\left(\ln(3/4), \ln(1/4)\right)$$

c.) Suppose that $\{N_j\}$ for $j = 1, \ldots, D$ is an independent set of identically distributed Bernoulli trials, with $P(N_j = 1) = v$, and $P(N_j = 0) = 1 - v$. Both D and v are known, and suppose that $Y_j = B + N_j$ for $j = 1, \ldots, D$. Note that the same realization of B is used for all Y_j, and that each Y_j equals B or $B + 1$ independently for each index j. Show that $Y_{\min} = \min\{Y_1, Y_2, Y_3, \ldots, Y_D\}$ is sufficient to estimate B given the observation $Y_1, Y_2, Y_3, \ldots, Y_D$ (A statistic $t(Y_1, Y_2, Y_3, \ldots, Y_D)$ is *sufficient* for B if the likelihood function of B for the measurments $Y_1, Y_2, Y_3, \ldots, Y_D$ depends on the measurements only through t.)

A statistic t is said to be sufficient for B if $p(Y_1,\dots,Y_D \mid t,B)$ is independent of B, think of B as random variable

-> $$p(B \mid t, Y_1,\dots,Y_D) = \frac{p(Y_1,\dots,Y_D \mid t,B)\,p(B \mid t)}{p(Y_1 \dots Y_D \mid t)}$$

-> t is sufficient for B iff $p(Y_1 \dots Y_D \mid B)$ can be factored into product of two functions

-> 1.) 1 depending on t and B

2.) another depending only on training samples

-> Through Factorization Theorem, shift our attention away from $p(Y_1 \dots Y_D \mid t,B)$ to simpler function

$$p(Y_1 \dots Y_D \mid B) = \prod_{k=1}^{n} p(y_k \mid B)$$

$$p(Y_{min} \mid t, B) = \frac{p(Y_{min}, t, B)}{p(t,B)}$$

$$= \frac{p(B \mid t, Y_{min})\, p(Y_{min}, t)}{p(B \mid t)\, p(t)}$$

$$= \frac{p(B \mid t, Y_{min})\, p(Y_{min} \mid t)\, \cancel{p(t)}}{p(B \mid t)\, \cancel{p(t)}}$$

$$= \frac{p(B \mid t, Y_{min})\, p(Y_{min} \mid t)}{p(B \mid t)}$$

pdf of B determined by sufficient statistic which implies that
$p(B|t, Y_{min}) = p(B|t)$

$$p(Y_{min}|t, B) = \frac{p(B|t, Y_{min})\,p(Y_{min}|t)}{p(B|t)}$$

$$= \frac{p(B|t)\,p(Y_{min}|t)}{p(B|t)}$$

$p(Y_{min}|t, B) = p(Y_{min}|t)$
which doesn't depend on B thus Y_{min} is sufficient to estimate B because it is independent of B

d.) The likelihood function for B given Y_{min} is $f(y_{\text{min}}|B) = 1 - v^D$ for $y_{\text{min}} = B$, and $f(y_{\text{min}}|B) = v^D$ for $y_{\text{min}} = B + 1$. Find the maximum likelihood estimator for B given $Y_1, Y_2, Y_3, \ldots, Y_D$.

d) The likelihood function for B given Y_{min} is
$f(y_{min}|B) = 1 - v^D$ for $y_{min} = B$ and
$f(y_{min}|B) = v^D$ for $y_{min} = B + 1$.
-> Find the maximum likelihood estimator for B given $Y_1, Y_2, \ldots, Y_D$

$$\hat{B} = \underset{B}{\arg\max}\ l(B)$$

$$= \underset{B}{\arg\max}\left(\ln f(y_{min} = B|B),\ \ln(y_{min} = B+1|B)\right)$$

$$\boxed{\hat{B} = \underset{B}{\arg\max}\left(\ln(1 - v^D),\ \ln(v^D)\right)}$$

Problem 2 - Bayesian Estimation

Suppose $\{Y_j\}$ are defined as in Problem 1. Additionally, it is known that parameter B has a probability densitry function $f_B(b) = e^{-b}$ for $b \geq 0$.

a.) Using the prior density of B only, what is an estimator of B? Provide an explanation of your choice of estimator.

using only the prior density, a good estimator for B would be the expected value of the prior distribution.
recognizing that $\lambda = 1$ is the distribution $e^{-b} \Rightarrow \lambda e^{-\lambda b}$

$$E[f_B] = \int_0^\infty b e^{-b} \, db$$
$$= -be^{-b}\Big|_0^\infty + \int_0^\infty e^{-b} \, db$$
$$= [0 - e^{-b}]\Big|_0^\infty$$
$$B = -e^{-\infty} + e^{0} \Rightarrow \boxed{B = 1}$$

b.) Using the given information in 1d.), as well as the prior distribution, what is the Maximum a Posterior of B given $\{Y_1, Y_2, Y_3, \ldots, Y_D\}$? Explain your answer carefully, providing all calculations.

The maximum a posterior estimator of B given $\{Y_1, Y_2, \ldots, Y_D\}$ would be derived by solving for the derivate of the joint density $f(y, B) = f(y|B) f(B)$

$$L(B, y) = \sum_{k=1}^{n} \ln f(y_k|B) + \ln f(B)$$

$$\hat{B} = \frac{d}{dB}\left[\ln(y_{min} = B|B) + \ln(y_{min} = B+1|B) + \ln(f_B(b))\right)$$

Taking the derivative w.r.t B would translate to where the slope = 0 thus would be the maximum a posterior probability estimate of B

$$\boxed{\hat{B} = \frac{d}{dB}\left[\ln(1-v^D) + \ln(v^D) + \ln(e^{-b})\right) = 0}$$

Problem 3 - Bayesian Classification

Consider the binary Bayesian classification problem.

Let class 1 have prior probability 1/3 and class 2 have prior 2/3. Suppose that we have uniform costs. Let the measurement x have density $f_1(x) = Ke^{-x}$, for $0 \leq x \leq 1$ and 0 otherwise under class 1. Suppose that the measurement x has density $f_2(x) = Ke^{-(1-x)}$, for $0 \leq x \leq 1$ and 0 otherwise under class 2

a.) Find the Bayesian optimal classifier in this case. Completely specify the decision region for class 1 using the x-axis. Simplify your decision rule as much as possible.

First, we start by finding the values of K that normalize the densities $f_1(x)$ and $f_2(x)$. Starting first for $f_1(x)$

$$\int_{-\infty}^{+\infty} K_1 e^{-x} dx = 1 \rightarrow K_1 \int_{-0}^{1} e^{-x} dx = 1$$

$$K_1(-e^{-x})\Big|_0^1 = 1 \rightarrow K_1(-e^{-1} + 1) = 1$$

$$K_1 = 1/(1 - e^{-1}) \rightarrow K_1 = 1.582$$

Now, for $f_2(x)$

$$\int_{-\infty}^{+\infty} K_2 e^{-(1-x)} dx = 1 \rightarrow K_2 \int_{-0}^{1} e^{-(1-x)} dx = 1$$

$$\frac{K_2}{e}(e^x)\Big|_0^1 = 1 \rightarrow \frac{K_2}{e}(e + 1) = 1$$

$$K_2 = e/(e - 1) \rightarrow K_2 = 1.582$$

Thus, $K_1 = K_2 = 1.582$

Using uniform costs and filling values for the known priors, the Bayesian optimal classifier can be defined as:

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{P(\omega_2)}{P(\omega_1)} \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \rightarrow \frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{2/3}{1/3} \frac{1 - 0}{1 - 0} \rightarrow \frac{p(x|\omega_1)}{p(x|\omega_2)} > 2$$

$\frac{p(x|\omega_1)}{p(x|\omega_2)} > 2 \rightarrow$ Decide $\omega_1 \rightarrow$Class 1 and $\frac{p(x|\omega_1)}{p(x|\omega_2)} < 2 \rightarrow$ Decide $\omega_2 \rightarrow$Class 2

Which, the likelihood ratio evaluates to:

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} > 2 \rightarrow \frac{K_1}{K_2} e^{-x+(1-x)} > 2 \rightarrow e^{-2x+1} > 2$$

So, when the Likelihood ratio $e^{-2x+1} > 2$, we decide ω_1 else we decide ω_2

The decision region for Class 1 can be found by equating the distributions and priors and solving for the boundary x_0:

$$p(x_0|\omega_1)P(\omega_1) = p(x_0|\omega_2)P(\omega_2)$$

$$\ln p(x_0|\omega_1) + \ln P(\omega_1) = \ln p(x_0|\omega_2) + \ln P(\omega_2)$$

$$-x_0 + \ln(1/3) = -1 + x_0 + \ln(2/3)$$

$$2x_0 = 1 + \ln(1/3) - \ln(2/3)$$

$$x_0 = \frac{1 + \ln(1/3) - \ln(2/3)}{2}$$

$$x_0 = 0.1534$$

Thus, the decision region for class 1 will be between $0 \leq x < 0.1534$

b.) Find an expression for the probability of error of the Bayesian optimal classifier. Simplify your result as much as possible

The expression for the probability of error for the Bayesian optimal classifier can be expressed as:

$$p_e = \int_{-\infty}^{+\infty} min[p(\omega_1|x), p(\omega_2|x)]p(x)dx$$

Simplifier further using the bounds of the distributions and definitions of the posterior probabilities:

$$p_e = \int_0^1 min[p(\omega_1|x), p(\omega_2|x)]p(x)dx$$

$$p_e = \int_0^1 min\left[\frac{p(x|\omega_1)P(\omega_1)}{p(x)}, \frac{p(x|\omega_2)P(\omega_2)}{p(x)}\right]p(x)dx$$

$$p_e = \int_0^1 min[p(x|\omega_1)P(\omega_1), p(x|\omega_2)P(\omega_2)]dx$$

This can be simplified even further by using the decision boundary to solve for the min probability error by:

$$p_e = \int_0^{x_0} p(x|\omega_2)P(\omega_2)dx + \int_{x_0}^1 p(x|\omega_1)P(\omega_1)dx$$

$$p_e = \frac{2K_2}{3e}\int_0^{0.1528} e^x dx + \frac{K_1}{3}\int_{0.1528}^1 e^{-x}dx$$

$$p_e = \frac{2K_2}{3e}e^x\Big|_0^{0.1528} + \frac{K_1}{3}e^{-x}\Big|_{0.1528}^1$$

$$p_e = \frac{2K_2}{3e}(e^{0.1528} - 1) + \frac{K_1}{3}(e^{-1} - e^{-0.1528})$$

$$p_e = 0.3227$$

Thus, the min $P(\text{error}) = 0.3227$

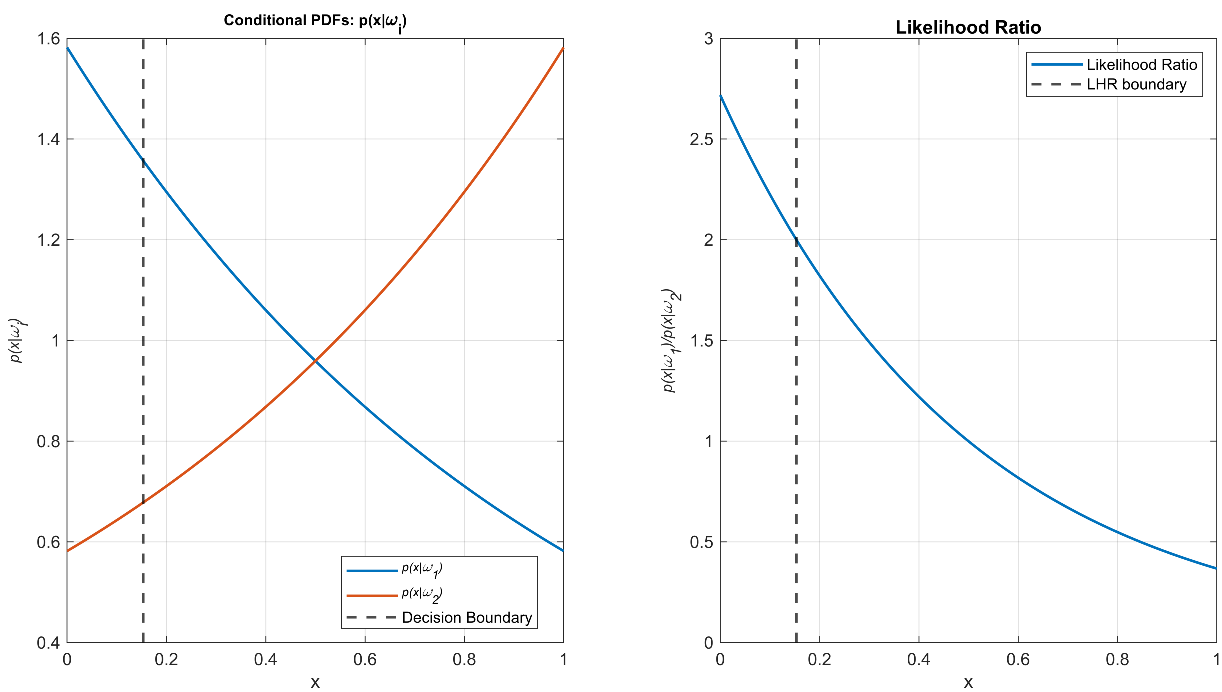
Double Checking Calculations in MATLAB

```
Pw1 = 1/3;
```

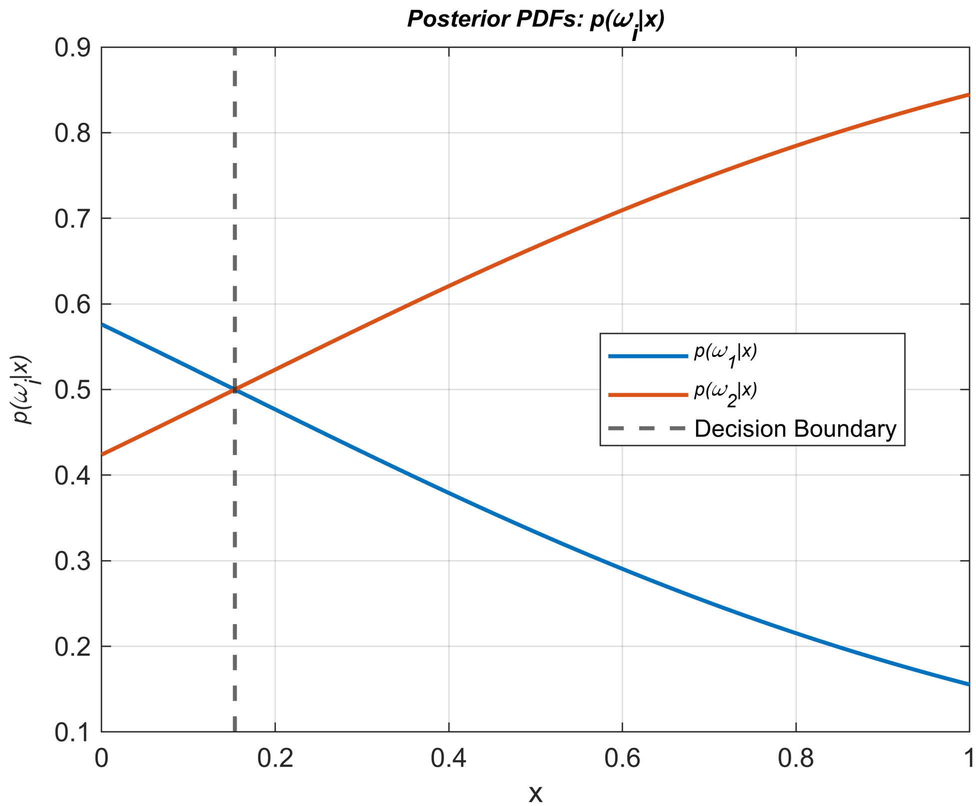
```
Pw2 = 2/3;
K = 1/(1-exp(-1))
```

```
K = 1.5820
```

```
x = 0:0.001:1;
x0 = (1+log(Pw1)-log(Pw2))/2;
pxw1 = K*exp(-x);
pxw2 = K*exp(-(1-x));
likelihoodRatio = pxw1./pxw2;
figure('Units','inches','Position',[0,0,12,6]);
subplot(1,2,1),plot(x,pxw1,'LineWidth',1.5), hold on, title('Conditional PDFs: p(x|\omega_i)')
plot(x,pxw2,'LineWidth',1.5), xlabel('\it{x}'), ylabel('\it{p(x|\omega_i)}'), grid on
xline(x0,'k--','LineWidth',1.5),hold off, legend('\it{p(x|\omega_1)}','\it{p(x|\omega_2)}','Dec
subplot(1,2,2),plot(x,likelihoodRatio,'LineWidth',1.5), grid on, xlabel('\it{x}'), ylabel('\it{
hold on, xline(x0,'k--','LineWidth',1.5), title('Likelihood Ratio'), legend('Likelihood Ratio',
xlim([0 1])
```




```
px = pxw1.*Pw1 + pxw2.*Pw2;
pw1x = (pxw1.*Pw1)./px;
pw2x = (pxw2.*Pw2)./px;
figure
plot(x,pw1x,'LineWidth',1.5), hold on, plot(x,pw2x,'LineWidth',1.5), grid on, xline(x0,'--','Li
xlabel('\it{x}'), ylabel('\it{p(\omega_i|x)}'), title('Posterior PDFs: \it{p(\omega_i|x)}')
legend('\it{p(\omega_1|x)}','\it{p(\omega_2|x)}', 'Decision Boundary','Location','best')
```




```
% Min P(error)
pe = Pw2*(K/exp(1))*(exp(x0)-1)- (K*Pw1)*(exp(-1)-exp(-x0))
```

```
pe = 0.3227
```