HW-1

1) Provide an intuitive example to show that $P(A|B)$ and $P(B|A)$ are in general not the same.

→ Here to prove $P(A|B) \neq P(B|A)$ in general lets take an example of someone having COVID-19 and Testing positive.
So, Here our two events are:
A: Testing Vte
B: Having COVID-19.

~~If a person have covid-19 an~~
Suppose that COVID-19 test is 80% accurate. Which means that probability testing positive while having COVID-19 is 80%.
i.e $P(A|B) = 0.8$
But alternatively there might be only 15% chance of having being tested positive and having COVID-19 as well. This happens due to "false positive".
i.e $P(B|A) = 0.15$

Hence it is proved that in general $P(A|B) \neq P(B|A)$.

(2)(i) Here X and Y are two Continous Random Variables And its also given that X and Y are independent

So, $P(X=x, Y=y) = P(X=x) \cdot P(Y=y)$

~~Now $E(X,Y) = \int\int x\,y\,P(x,y)\,dx\,dy$~~

~~$= \int x P_x[X]\,dx \cdot y P_y[Y]\,dy$~~

~~$= \int x P_x[X]\,dx \cdot \int y P_y[Y]$~~

Now $E(X,Y) = \int\int x\,y\,P(X=x, Y=y)\,dx\,dy$

$= \int x P(X=x)\,dx \cdot y P(Y=y)\,dy$

$= \int x P(X=x)\,dx \cdot \int y P(Y=y)\,dy$

$E(X,Y) = E[X] \cdot E[Y]$ by def.

$E(X,Y) - E[X] \cdot E[X] = 0$

Now coefficient constant $= \dfrac{cov[X,Y]}{\sqrt{var[X] \cdot var[Y]}}$

$= \dfrac{E(X,Y) - E[X] \cdot E[Y]}{\sqrt{var[X] \cdot var[X]}}$

$= 0 \quad [\because E(X,Y) - E[X]\,E[Y] = 0]$

Hence proved.

(2) Suppose X and Y are uncorrelated, can we conclude X and Y are independent? If so, prove it or give one counter example.

No to prove that ↑ assume $X \sim \text{Normal}(0,1)$

and $Y = X^2$

and $E(X) = 0$ as $\mu = 0$ and $E(X^3) = 0$ (third moment mean)

i.e Y is dependent upon X.

And if we still be able to prove $\text{cov}(X,Y) = 0$ then it will be clear that X and Y are uncorrelated but not independent.

$$\text{Cov}(X,Y) = E[(X-\bar{x})(Y-\bar{Y})]$$

From question ① simplification we can write

$$\text{Cov}(X,Y) = E(XY) - E(X)E(Y)$$
$$= E(X^3) - E(X)E(Y) \quad \text{i.e } Y = X^2$$
$$= 0 - 0 \cdot E(Y)$$
$$= 0$$

Hence X and Y are uncorrelated

But from $Y = X^2$ we can say that they are dependent

Hence proved that if X and Y are uncorrelated then they may not be independent.

3) [Minimum Probability of Error, Discriminat function]
Let the components of the vector $x = [x_1, \cdots, x_d]^T$ be binary valued (0,1), and Let $P(\omega_j)$ be the prior probabilty for the state of nature ω_j and $j = 1, \cdots, c$. We define

$$P_{ij} = P(x_i = 1 | \omega_j), i = 1, \cdots, d, j = 1, \cdots, c$$

With the components x_i being statistical independent for all x in ω_j. Show the minimum probabilty of error is achived by the following decision rule.
Decide ω_k if $g_k(x) \geq g_j(x)$ for all j and k where

$$g_j(x) = \sum_{i=1}^{d} x_i \ln \frac{P_{ij}}{1-P_{ij}} + \sum_{i=1}^{d} \ln(1-P_{ij}) + \ln P(\omega_j)$$

→ Here we can use formula of ther posterior probabilty to ~~deriv~~ derive the $g_j(x)$
Let's ~~co~~ consider it as following ~~fun~~ function

$$g_j(x) = p(x/\omega_j) \cdot P(\omega_j)$$
$$= \ln[p(x/\omega_j) \cdot p(\omega_i)] \text{ taking ln.}$$
$$= \ln p(x/\omega_i) + \ln p(\omega_j) \quad ----[1]$$
(density, prior)

~~Here write is given that x~~ Here it is given that x are statistically independent for all x in ω_j
So we can write the the density as product over $i = 1$ to j.

$$p(x|\omega_j) = \prod_{i=1}^{d} p(x_i|\omega_j)$$

Now we can convert above equation in terms of p_{ij}

$$= \prod_{i=1}^{d} p_{ij}^{x_i}(1-p_{ij})^{1-x_i} \quad - - [2]$$

plug in [2] in [1]

$$g_j(x) = \sum_{i=1}^{d} \ln\left[p_{ij}^{x_i}(1-p_{ij})^{1-x_i}\right] + \ln p(\omega_j)$$

$$= \sum_{i=1}^{d} \left[x_i \ln p_{ij} + (1-x_i)\ln(1-p_{ij})\right] + \ln p(\omega_j)$$

$$= \sum_{i=1}^{d} x_i \ln\left(\frac{p_{ij}}{1-p_{ij}}\right) + \sum_{i=1}^{d} \ln(1-p_{ij}) + \ln p(\omega_j)$$

Hence derived.

4) [Likelyhood Ratio] Suppose we consider two category classification, the class condition are assumed to be Gaussian. i.e $p(x/\omega_1) = N(4,1)$ and $p(x/\omega_2) = N(8,1)$ based on prior knowledge we have $p(\omega_2) = 1/4$. We do not penalize for correct classification, while for misclassification, we put 1 unit penalty for misclassifying ω_1 to ω_2 and put 3 unit for misclassifying ω_2 to ω_1. Deaive bayesian decision rule using likely hood ratio.

→ Here we are given following informations

$p(x/\omega_1) = N(4,1)$ $\quad p(\omega_2) = 1/4$

$p(x/\omega_2) = N(8,1)$ $\quad p(\omega_1) = 3/4$

and penalty $\lambda = \begin{matrix} \lambda_{11} \\ \lambda_{21} \end{matrix} \begin{bmatrix} 0 & 3 \\ 1 & 0 \end{bmatrix} \begin{matrix} \lambda_{12} \\ \lambda_{22} \end{matrix}$

we know that

$$N = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-(x-a)^2}{2\sigma}$$

So, from above equation we can write

$$p(x/\omega_1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-4)^2}{2}} \quad --- [1]$$

$$p(x/\omega_2) = \frac{1}{\sqrt{2\pi}} e^{-1} \frac{(x-8)^2}{2} \quad --- [2]$$

From Likelyhood ratio test we also know following is true

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{12}} \frac{p(\omega_2)}{p(\omega_1)}$$

Now plug in value of [1] & [2] in above formula

$$\frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{(x-4)^2}{2}}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{(x-8)^2}{2}}} > \frac{3-0}{1-0} \times \frac{1}{4} \times \frac{4}{3}$$

$$e^{-\frac{(x-4)^2}{2} + \frac{(x-8)^2}{2}} > 1$$

$$\log\left(e^{\frac{-(x-4)^2}{2} + \frac{(x-8)^2}{2}}\right) > \log(1)$$

$$\frac{-(x-4)^2}{2} + \frac{(x-8)^2}{2} > 0$$

$$-(x-4)^2 + (x-8)^2 > 0$$

$$-(x^2-8x+16) + (x^2-16x+64) > 0$$

$$-x^2+8x-16+x^2-16x+64 > 0$$

$$-8x+48 > 0$$

$$48 > 8x$$

$$\boxed{x < 6}$$

So if $x < 6$ then we can conclude following

if $x < 6$ then we go for ω_1.
other wise we go for ω_2.

5) [Minimum Risk, Reject option] In many machine learning applications, one has either option to assign the pattern to one of c classes, or to reject it as being unrecognizable. If cost of reject is not too high rejection may be desirable action. Let

$$\lambda(\alpha_i|\omega_j) = \begin{cases} 0 & , i=j \text{ and } i,j=1,2,\cdots c \\ \lambda_r & , i=c+1 \\ \lambda_s & , \text{otherwise.} \end{cases}$$

When λ_r is loss incurred for choosing the (c+1)th action, rejection, and λ_s is the loss incurred for making any substitution error.

(1) Derive the decision rule with minimum risk.

Risk is defined as $R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|x)$

Now for $i = 1,2,\cdots,c$

$$\text{So } R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|x)$$

$$= \sum_{j=i} 0 + \sum_{\substack{j=1, \\ j\neq i}}^{c} \lambda_s P(\omega_j|x)$$

$$= 0 + \lambda_s \sum_{\substack{j=1 \\ j\neq i}} P(\omega_j|x)$$

$$= 0 + \lambda_s [1 - P(\omega_i|x)]$$

$$= \lambda_s [1 - P(\omega_i|x)]$$

Now for $i=c+1$

$$R(\alpha_{c+1}|x) = \lambda_r \quad (\because \text{given})$$

So, we can achive min risk if we decide ω_i if

$$R(\alpha_i|x) \leq R(\alpha_{c+1}|x)$$

Here we know $R(\alpha_i|x)$ and $R(\alpha_{c+1}|x)$

$$\text{So} \quad \lambda_s [1 - P(\omega_i|x)] \leq \lambda_r$$

$$1 - P(\omega_i|x) \leq \lambda_r/\lambda_s$$

$$\boxed{P(\omega_i|x) \geq 1 - \lambda_r/\lambda_s}$$

(2) What happens if $\lambda_r = 0$

if $\lambda_r = 0$ then $P(\omega_i|x) \geq 1$

And we always REJECT.

(3) What happens if $\lambda_r > \lambda_s$?

then $P(\omega_i|x) > 1$

And we will NEVER REJECT.

6) [Maximum Likelyhood Estimation (MLE)] A general representation of a exponential family given by the following probability function:

$$p(x|\eta) = h(x) \exp\{\eta^T T(x) - A(\eta)\}$$

η is natural parameter.
$h(x)$ is the base density which ensure x in right space.
$T(x)$ is the sufficient statistics.
$A(\eta)$ is the log normalizer which is determined by $T(x)$ & $h(x)$
$\exp(.)$ represent the exponential function.

(1) Write down the expression of $A(\eta)$ in terms of $T(x)$ & $h(x)$

$$p(x|\eta) = h(x) \cdot \exp\{\eta^T T(x) - A(\eta)\}$$

Taking limit on the both sides. in term of x.

$$\int_x p(x|\eta) = \int_x h(x) \cdot \exp\{\eta^T T(x) - A(\eta)\}$$

Now we can assign integration of probability to 1.

$$1 = \int_x h(x) \cdot \exp\{\eta^T T(x) - A(\eta)\}$$

Taking log on both side

$$\log(1) = \log\left[\int_x h(x) \cdot \exp\{\eta^T T(x) - A(\eta)\}\right]$$

$$0 = \log \left[\int_x h(x) \cdot \exp(\eta^T T(x)) \cdot \exp(-A(\eta)) \right]$$

$$0 = \log \int_x h(x) \cdot \exp(\eta^T T(x)) + \log (\exp\{-A(\eta)\}$$

[$\because$ $A(\eta)$ constant to $\int_x$]

$$0 = \log \int_x h(x) \cdot \exp(\eta^T T(x)) - A(\eta)$$

$$\boxed{A(\eta) = \log \int_x h(x) \cdot \exp(\eta^T T(x))}$$

(2) Show that $\partial / \partial \eta A(\eta) = E_\eta T(x)$ where $E_\eta(\cdot)$ is the expectation w.r.t $p(x|\eta)$

$$A_\eta = \log \int_x (h(x) \cdot \exp(\eta^T T(x))$$

$$= e^{\log \int_x h(x)} \cdot$$

$$\frac{\partial A(\eta)}{\partial \eta} = \left[\frac{\partial}{\partial \eta} \int_x h(x) \exp(\eta^T T(x)\, dx \right] \cdot \frac{1}{\int_x h(x) \exp(\eta^T T(x)}$$

$$= \left[\frac{\partial}{\partial \eta} \int_x h(x) \cdot \exp(\eta^T T(x)) \cdot dx \right] \cdot \frac{1}{e^{A(\eta)}}$$

$$= \int_x h(x) \cdot \exp\{\eta^T T(x)\}\, T(x)\, dx$$

$\{\because 1/e^{A(\eta)}$ is independent of $x\}$

$$= \int_x h(x) \cdot \exp\{\eta^T T(x) - A(\eta)\}^2\, T(x)\, dx$$

$$= \int_x p(x/\eta) \cdot T(x)\, dx$$

$$\boxed{= E_\eta\, T(x)}$$

(3) Suppose we have n i.i.d samples $x_1, x_2, \cdots, x_n$ derive the maximum likelihood estimator for η.

$$p(x_i|\eta) = \prod p(x_i|\eta) \quad [\because \text{prob (3)}]$$

$$p(x|\eta) = h(x) \exp\{\eta^T T(x) - A(\eta)\}^2$$

Hence $p(x_i|\eta) = \prod h(x_i) \exp\{\eta^T \sum_i T(x_i)\} - n\, A(\eta)^2$

Now lets take Log of Likelyhood.

$$\log(\eta : x_1, x_2, \cdots x_n) = \log(p(x_1, x_2, \cdots x_n | \eta)$$

$$= \log(h(x_1, \cdots x_n) + \eta^T \sum T(x_i) - n\, A(\eta)^2$$

Takeing gradients on both side

$$\frac{\partial}{\partial \eta} \log(\eta : x_1, x_2, \cdots, x_n) = \frac{\partial}{\partial \eta}\left[h(x_1, x_1 \cdots x_n) + \eta^T \sum T(x_i) - \eta A(\eta) \right]$$

$$0 = 0 + \sum_i T(x_i) - n \frac{\partial A(\eta)}{\partial \eta}$$

$$0 = \sum_i T(x_i) - n \frac{\partial A(\eta)}{\partial \eta}$$

$$E_\eta(T(x)) = \frac{\sum_i T(x_i)}{n}$$

$$\mu_m = \sum_i \frac{T(x_i)}{n}$$

$$= \frac{\partial A(\eta)}{\partial(\eta)}$$

7) (1) Suppose the classifier is $y = x^T\theta$, where θ contains the weights as well as bias parameters. The Log likelyhood function is $LL(\theta)$, what is $\frac{\partial LL\theta}{\partial \theta}$

$$y = x^T\theta$$

So, $\sigma(y) = \frac{1}{1+e^{-x^T\theta}}$ [Logistic function]

Now let's assume following.

$$p(y=1 \mid x_i\theta) = \sigma(x^T\theta)$$

$$p(y=0 \mid x_i\theta) = 1 - \sigma(x^T\theta)$$

Then likely hood fun will be

~~$P(\log [x_i\theta) = \prod_1^N (\sigma(x^T\theta))^{y_i} (1-\sigma(x^T\theta))$~~

$$p(y \mid x_i\theta) = \prod_{n=1}^{N} y_n^{y_n} (1-y_n)^{1-y_n}$$

$$= \prod_1^N (\sigma(x^T\theta))^{y_i} (1-\sigma(x^T\theta))$$

Log likely hood

$$LL(\theta) = \sum_{i}^{N} y_i \log \sigma(x^T\theta) + (1-y_i) \log (1-\sigma(x^T\theta))$$

Taking gradient on both side

$$\frac{\partial LL(\theta)}{\partial(\theta)} = \frac{-y_i}{\sigma(x^T\theta)} \frac{\partial}{\partial\theta} \sigma(x^T\theta) - (1-y_i) \frac{1}{(1-\sigma(x^T\theta))} \cdot \frac{\partial}{\partial\theta} (1-\sigma(x^T\theta))$$

$$= \frac{-y_i}{\sigma(x^T\theta)} \cdot \sigma(x^T\theta)(1-\sigma(x^T\theta)) \cdot x^T$$

$$- \frac{(1-y_i)}{(1-\sigma(x^T\theta))} \cdot (\sigma(x^T\theta)(1-\sigma(x^T\theta)) x^T$$

taking x^T common

$$= -x^T [y_i - y_i \sigma(x^T\theta) - \sigma(x^T\theta) + y_i \sigma(x^T\theta)]$$

$$= -x^T [y_i - \sigma(x^T\theta)]$$

Finally $$\frac{\partial LL\theta}{\partial\theta} = \sum_{i}^{N} [y_i - \sigma(x^T\theta)] x_i^T$$