Assignment 1

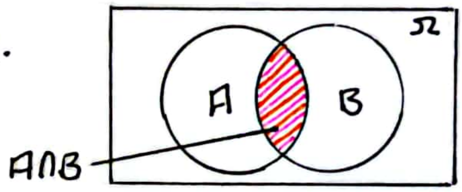
1. Provide an intuitive example to show that $P(A|B)$ and $P(B|A)$ are in general not the same.
Provide matrix examples to show $AB \neq BA$

Solution] Let A and B be two events, & Ω be sample space.

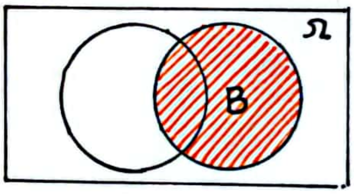
$P(A|B)$ = The probability of A, Given B has occured

$P(B|A)$ = The probability of B, Given A has occured

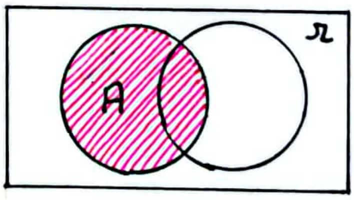
Events $A \cap B$ in sample space.



New sample space AFTER B has occured.



New sample space AFTER A has occured.



Probability of A Given B = $\dfrac{P(A \cap B)}{\text{Sample space after B occured}} = \dfrac{P(A \cap B)}{P(B)}$

Probability of B given A = $\dfrac{P(A \cap B)}{\text{Sample space after A occured}} = \dfrac{P(A \cap B)}{P(A)}$

We know that generally $P(A) \neq P(B)$, and can thus conclude that generally $P(A|B)$ and $P(B|A)$ are not the same

To show $AB \neq BA$ with matrix examples.

Let $A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$

$$AB = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} (1\times1)+(1\times3) & (1\times2)+(1\times4) \\ (1\times1)+(1\times3) & (1\times2)+(1\times4) \end{bmatrix} = \begin{bmatrix} 4 & 6 \\ 4 & 6 \end{bmatrix} \ldots\ldots (1)$$

$$BA = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} (1\times1)+(1\times2) & (1\times1)+(1\times2) \\ (3\times1)+(4\times1) & (3\times1)+(3\times4) \end{bmatrix} = \begin{bmatrix} 3 & 3 \\ 7 & 7 \end{bmatrix} \ldots\ldots (2)$$

(1) & (2) are not equal

$\therefore AB \neq BA$

2. Independance and uncorrelation.

(1) Suppose X and Y are two continuous random variables. Show that if X and Y are independent, then they are uncorrelated

solution] X and Y are continuous and independent

$\Rightarrow f(x,y) = f_x(x)\, f_y(y)$

$$\text{Cov}(X,Y) = E(XY) - E(X)E(Y)$$
$$= \iint xy\, p_{xy}(x,y)\,dx\,dy - \int x\, p_x(x)\,dx \int y\, p_y(y)\,dx$$

(since X & Y are independent)

$$= \iint xy\, p_x(x)\, p_y(y)\,dx\,dy - \int x\, p_x(x)\,dx \int y\, p_y(y)\,dy$$
$$= \int x\, p_x(x)\,dx \int y\, p_y(y)\,dy - \int x\, p_x(x)\,dx \int y\, p_y(y)\,dy$$
$$= 0.$$

Since $\text{cov}(X,Y) = 0$ we know X & Y is uncorrelated

(2) Suppose X and Y are uncorrelated, can we conclude X & Y are independent?

solution] We assume $X \sim \text{Uniform}[-1,1]$ and $Y = X^2$

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

Since $\text{Cov}(X\,Y) = 0$ we know X & Y is uncorrelated. while X and Y are dependent. Thus we conclude that uncorrelation does not conclude independance.

3. Let $W_{max}(x)$ be state of nature for which $P(\omega_{max}|x) \geq P(\omega_i|x)$ for all $i = 1, \ldots c$

(1) Show that $P(\omega_{max}|x) \geq 1/c$

Solution] We know that since $P(\Omega) = 1$, then $\sum_{i=0}^{c} P(\omega_i|x) = 1$

If $P(\omega_i|x) = P(\omega_j|x)$ for all i & j then we know that
$P(\omega_i|x) = \frac{1}{c}$
$\therefore P(\omega_{max}|x) = \frac{1}{c}$

Now if any $P(\omega_i|x)$ is less than $1/c$ $[P(\omega_i|x) < 1/c]$
then, $P(\omega_{max}|x) > \frac{1}{c}$

$\therefore P(\omega_{max}|x) \geq \frac{1}{c}$

(2) Show that minimum error rate decision rule, the average probability of error is given by
$P(error) = 1 - \int P(\omega_{max}|x) p(x) dx$

Solution] We know that when we minimize the average probability of error
$P(error) = \int P(error|x) p(x) dx$

We know that $P(error|x) = 1 - P(\omega_{correct}|x) dx$. Thus,
$P(error) = \int (1 - P(\omega_{max}|x)) p(x) dx$

$P(error) = 1 - \int P(\omega_{max}|x) p(x) dx.$

(3) Show that $P(\text{error}) \leq \frac{c-1}{c}$

Solution] From (1) we know that $P(\omega_{max}|x) \geq 1/c$

From (2) we know that $P(\text{error}) = 1 - \int P(\omega_{max}|x) P(x)\, dx$

Substituting (1) in (2)

$$P(\text{error}) \leq 1 - \int \frac{1}{c} P(x)\, dx$$

$\Rightarrow P(\omega_{max}|x) \geq 1/c$

$\Rightarrow 1 - P(\omega_{max}|x) \leq 1 - 1/c$

$$P(\text{error}) \leq 1 - \frac{1}{c} \int P(x)\, dx$$

$$\leq 1 - 1/c$$

$$\Rightarrow P(\text{error}) \leq \frac{c-1}{c}$$

4. In two category classification, the class conditionals are ie. $p(x|\omega_1) = N(4,1)$, $p(x|\omega_2) = N(8,1)$.
Based on prior knowledge, $P(\omega_2) = 1/4$.
We do not penalize for correct classification while for misclassification, we put 1 unit of penalty for misclassifying ω_1 to ω_2 and 3 units misclassifying ω_2 to ω_1.
Derive bayesian rule using Likelihood Ratio.

Solution] Using decision Rule., decide ω_1 if

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}$$

$$\text{LHS}: \frac{p(x|\omega_1)}{p(x|\omega_2)} = \frac{\frac{1}{\sqrt{2\pi}} e^{\frac{-(x-4)^2}{2}}}{\frac{1}{\sqrt{2\pi}} e^{\frac{-(x-8)^2}{2}}} = e^{\frac{-(x-4)^2}{2} + \frac{(x-8)^2}{2}}$$

$$\text{RHS}: \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)} = \frac{3-0}{1-0} \cdot \frac{1/4}{3/4} = \frac{3}{1} \cdot \frac{1}{3} = 1$$

$$e^{\frac{-(x-4)^2}{2} + \frac{(x-8)^2}{2}} > 1$$ → substituting LHS & RHS in decision rule

$$-\frac{(x-4)^2}{2} + \frac{(x-8)^2}{2} > \log 1$$ → taking log on both sides.

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

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

$$-8x + 48 > 0$$

$$x < 6$$

∴ Using decision rule, we decide ω_1 if
$x < 6$
Otherwise we decide ω_2

5. In many ML operations, one has the option to assign the pattern to one of the c classes or to reject it as being unrecognizable. If cost of rejection is not too high, rejection may be a desirable action. Let

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

where λ_r is the loss incurred by choosing the $(c+1)^{th}$ action - rejection, and λ_s is the loss incurred by making a substitution error

(1) Derive the decision rule with minimum risk.

We find Risk assuming ω_{max} is our correct class

$$Risk = \sum_{j\neq max} \lambda_s P(\omega_j|x)$$

$$= \lambda_s \sum_{j\neq max} P(\omega_j|x)$$

$$= \lambda_s[1 - P(\omega_{max}|x)]$$

We also find risk of a ω_k where $k \neq max$.

$$Risk = \sum_{j\neq k} \lambda_s P(\omega_j|x)$$

$$= \lambda_s \sum_{j\neq k} P(\omega_j|x)$$

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

$$\Rightarrow \lambda_s[1 - P(\omega_k|x)] \geq \lambda_s[1 - P(\omega_{max}|x)]$$

$\therefore$ We always choose maximum probability that is low risk.

With rejections our risk is λ_r

$\therefore$ We should choose ω_{max} or reject depending on which is smaller.

$\lambda_s[1 - P(\omega_{max}|x)]$ or λ_r

We reject if

$$\lambda_r \leq \lambda_s[1 - P(\omega_{max}|x)]$$

$$\frac{\lambda_r}{\lambda_s} \leq 1 - P(\omega_{max}|x)$$

$$P(\omega_{max}|x) \geq 1 - \frac{\lambda_r}{\lambda_s}$$

Else we accept.

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

Solution] Since $\lambda_r = 0$ then risk is ~~afs~~ rejection is 0

$\therefore$ We should reject.

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

Solution] If $\lambda_r > \lambda_s$ then risk of reject is more.

$\therefore$ We should not not reject

6. A general representation of a exponential family is given by the following probability dencity.

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

- η is natural parameter.
- $h(x)$ is the base density which ensures x is in right space
- $T(x)$ is the sufficient statistics
- $A(\eta)$ is the log normalizer which is determined by $T(x)$ and $h(x)$
- $\exp(\cdot)$ represents the exponential function.

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

Solution] We know that $\int P(x|\eta)\,dx = 1$ ①

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

② in ①

$$\int h(x)\exp\{\eta^T T(x) - A(\eta)\}\,dx = 1$$

$$\exp\{-A(\eta)\}\int h(x)\exp\{\eta^T T(x)\}\,dx = 1$$

$$\exp\{-A(\eta)\} = \frac{1}{\int h(x)\exp\{\eta^T T(x)\}\,dx} \quad ③$$

$$\log(\exp\{-A(\eta)\}) = \log\left(\frac{1}{\int h(x)\exp\{\eta^T T(x)\}dx}\right)$$

$$-A(\eta) = \overset{0}{\cancel{\log 1}} - \log\left(\int h(x)\exp\{\eta^T T(x)\}dx\right)$$

$$\therefore A(\eta) = \log\int h(x)\exp\{\eta^T T(x)\}\,dx$$

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

solution] $\frac{\delta A(\eta)}{\delta \eta} = \frac{\delta}{\delta\eta}\left[\log\left(\int h(x)\exp\{\eta^T T(x)\}\,dx\right)\right]$

$= \frac{\frac{\delta}{\delta\eta}\int h(x)\exp\{\eta^T T(x)\}dx}{\int h(x)\exp\{\eta^T T(x)\}dx}$

$= \frac{1}{\int h(x)\exp\{\eta^T T(x)\}dx} \cdot \frac{\delta}{\delta\eta}\int h(x)\exp\{\eta^T T(x)\}dx$

from ①

$= \exp\{-A(\eta)\} \cdot \frac{\delta}{\delta\eta}\int h(x)\exp\{\eta^T T(x)\}dx$

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

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

from ②

$= \int T(x)\, P(x|\eta)\,dx$

$\{ E(h(x)) = \int h(x) P(x)]$

$= E_\eta T(x)$

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

solution] We know that

$\Rightarrow l(\theta) = \log L(\theta) = \log \prod_{i=1}^{n} f(x_i|\theta)$

$\Rightarrow l(\eta) = \log L(\eta) = \log \prod_{i=1}^{n} p(x_i|\eta)$

We know that $P(x_i|\eta) = h(x_i)\exp\{\eta^T T(x_i) - A(\eta)\}$

likelihood $l(\eta) = \log \prod_{i=1}^{n} p(x_i|\eta)$

$= \log \prod_{i=1}^{n} h(x_i) + \eta^T \sum_{i=1}^{n} T(x_i) - nA(\eta)$

Diffrentiate on both sides with respect to η

$$\frac{\delta}{\delta\eta} l(\eta) = \sum_{i}^{n} T(x_i) - n \frac{\delta}{\delta\eta} A(\eta)$$

We know that $\nabla_\theta l = \frac{\delta}{\delta\eta} l(n) = 0$

$$\sum_{i=1}^{n} T(x_i) - n \frac{\delta}{\delta\eta} A(\eta) = 0$$

$$n \frac{\delta}{\delta\eta} A(\eta) = \sum_{i=1}^{n} T(x_i)$$

$$\frac{\delta}{\delta\eta} A(\eta) = \frac{\sum_{i=1}^{n} T(x_i)}{n}$$

From (2) we know that $\frac{\delta}{\delta\eta} A(\eta) = E_\eta T(x)$

$$\therefore E_\eta T(x) = \frac{\sum_{i=1}^{n} T(x_i)}{n}$$

CamScanner