Camlin Page
Date / /

9.1 Independence and Uncorrelation.

(1) Suppose x and y are two continuous variables, show that if x and y are independent, then they are uncorrelated.

⇒ Proof:

Two Random variables x, y

Two random variable 'x' and 'y' are independent when their joint probability distribution is a product of their marginal probability distribution (for Σ_x and Σ_y)

i.e. $$P_{xy}(x,y) = P_x(x)\, P_y(y) \quad \text{——— (i)}$$

(where x and y are elements of variables 'x' and 'y' respectively)

Two random variables x and y are uncorrelated when their correlation co-efficient is zero.

i.e $\rho(x,y) = 0$

$$\because \rho(x,y) = \frac{\text{cov}[x,y]}{\sqrt{\text{var}[x]\,\text{var}[y]}}$$

but even $\text{Cov}[x,y]$ is zero.

$$\therefore \text{Cov}[x,y] = E[x,y] - E[x]\,E[y]$$

$$\therefore 0 = E[x,y] - E[x]E[y]$$

$$\therefore E[xy] = E[x]E[y] \quad \text{——— (ii)}$$

We know that,

$$E[x,y] = \iint x.y.P_{x,y}(x,y)\,dx\,dy$$

but from equation (i),

$$E[x,y] = \iint x.y.(P_x[x].P_y[y])\,dx\,dy$$

$$= \iint (x P_x[x] \cdot y P_y[y])\,dx\,dy$$

$$= \int x P_x[x] \int y P_y[y]\,dy\,dx$$

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

$$E[x,y] = E[x] \cdot E[y] \quad \text{---} \quad (iii)$$

So from (iii) we can conclude that when $E[x,y] = E[x].E[y]$ the two random variables are uncorrelated. And from (i) we can conclude they are also independent. Hence proved.

Q.1

(ii) Suppose x and y are uncorrelated, can we conclude x and y are independent? If so, prove it, otherwise, given one counter example.

⇒ Lets take an example for two random variables 'x' and 'y' with 0 correlation and check if they satisfy independency condition.

$x \sim$ Uniform $[-1, 1]$

$y = x^2$

let probability of $x = -1$ and $x = 1$ be $\frac{1}{2}$ and $\frac{1}{2}$

$\therefore \quad y = x^2$

$y = [1, 1]$

let probability of $y = 1, y = 1$ be $\frac{1}{2}$ and $\frac{1}{2}$

$$E[x] = -1 \times \frac{1}{2} + 1 \times \frac{1}{2}$$

$$= 0$$

$$E[y] = 1 \times \frac{1}{2} + 1 \times \frac{1}{2}$$

$$= 1$$

$$E[x, y] = \cancel{1 \times 1} \; -1 \times \frac{1}{2} + 1 \times \frac{1}{2} = 0$$

$\therefore \quad E[x, y] = E[x] \cdot E[y] = 0$

Thus x and y are uncorrelated.

Now let's check the marginal distributions

$P[X]$ = for $x=1$ = $\frac{1}{2}$ $\quad$ $P[y]$ = for $y=1$ = $\frac{1}{2}$

$P[X]$ = for $x=-1$ = $\frac{1}{2}$ $\quad$ //

$P[X,Y]$ for $[-1,1]$ = $\frac{1}{2}$ $\quad$ $P[X,Y]$ for $[1,1]$ = $\frac{1}{2}$

So for $P[x,y] = P[X].P[Y]$ (if independent)

$$\frac{1}{2} = \frac{1}{2} \times \frac{1}{2}$$

$$\frac{1}{2} \neq \frac{1}{4}$$

As we can conclude from the above example not all uncorrelated random variables are necessarily independent.

Camlin Page
Date / /

Q.2 [Minimam Error Rate Decision] Let $\omega_{max}(x)$ be state of nature for which $P(\omega_{max}/x) \geq P(\omega_i/x)$ for all $i = 1, \ldots, c$

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

$\Rightarrow$ As,

Given $P(\omega_{max}/x) \geq P(\omega_i/x)$,

$$\sum_{i=1}^{c} P(\omega_{max}/x) \geq \sum_{i=1}^{c} P(\omega_i/x)$$

but $\sum_{i=1}^{c} P(\omega_i/x) = 1$

$\therefore \quad c\,P(\omega_{max}/x) \geq 1$

$\therefore \quad P(\omega_{max}/x) \geq \frac{1}{c}$ ——— (i)

Hence proved.

(ii) Show that for minimam-error-rate decision rule, the average probability of error is given by,

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

$$P(error) = \int_{i}^{c} P(error/x)\, p(x)\, dx$$

$$= \int_{i}^{c} (1 - P(\omega_{max}/x))\, p(x)\, dx$$

$$= 1 - \int_{i}^{c} P(\omega_{max}/x)\, p(x)\, dx$$ ——— (ii)

Hence proved

(iii) Show that $P(\text{error}) \leq \frac{C-1}{C}$

From (i) and (ii)

$$* \int_{-\infty}^{\infty} P(\omega_{max}/x) p(x)\, dx \geq * \int_{-\infty}^{\infty} \frac{1}{C} p(x)\, dx$$

$$\therefore 1 - \int_{-\infty}^{\infty} P(\omega_{max}/x)\, p(x)\, dx \leq 1 - \int_{-\infty}^{\infty} \frac{1}{C} p(x)\, dx$$

$$\therefore P(\text{error}) \leq 1 - \int_{-\infty}^{\infty} \frac{1}{C} p(x)\, dx$$

$$\leq 1 - \frac{1}{C}$$

$$\therefore P(\text{error}) \leq \frac{C-1}{C}$$

// Hence proved.

Camlin | Page
Date / /

Q3 [Likelihood Ratio] Suppose we consider two category classification, the class conditionals 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 $P(\omega_2) = \frac{1}{2}$. 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. Derive the bayesian decision rule using likelihood ratio

⇒

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

$P(\omega_2) = \frac{1}{4}$

$$\lambda = \begin{matrix} \\ \omega_1 \\ \omega_2 \end{matrix} \begin{matrix} \begin{matrix} \lambda_{11} & \lambda_{12} \end{matrix} \\ \begin{bmatrix} 0 & 3 \\ 1 & 0 \end{bmatrix} \\ \begin{matrix} \lambda_{21} & \lambda_{22} \end{matrix} \end{matrix}$$

By definition,

$$\frac{P(x/\omega_1)}{P(x/\omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{22} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)}$$

The Posterior Probability = $P(x/\omega_1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-h)^2}{2}}$ —(i)

(In Gaussian form)

$= P(x/\omega_2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-8)^2}{2}}$ —(ii)

From (i), (ii) and (iii),

$$\frac{P(x/\omega_1)}{P(x/\omega_2)} = \frac{\frac{1}{\sqrt{2\pi}} \times e^{-\frac{(x-h)^2}{2}}}{\frac{1}{\sqrt{2\pi}} \times e^{-\frac{(x-8)^2}{2}}}$$

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

$$\theta = \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}$$

$$= \frac{3-0}{1-0} \times \frac{1/4}{3/4}$$

$$= 1 * 1$$

$$= 1$$

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

Take logs,

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

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

$$-8x + 48 < 0$$

$$48 < 8x$$

$$x > 6$$

When x is smaller than 6 we choose ω_1

when x is larger than 6 we choose ω_2

Camlin Page
Date / /

Q.4 [Minimum Risk, Reject Option] In many machine learning applications, one has the option either to assign the pattern to one of c class, or to reject it as being unrecognizable. If the cost for reject is not too high, rejection maybe a desirable action. Let

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

where λ_r is the loss incurred for choosing the $(c+1)$-th action, rejection, and λ_s is the loss incurred for making any substitution ever.

(1) Derive the decision rule with minimum risk.

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

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

(1) For $i = 1, 2, 3, \ldots, c$

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

$$= \lambda_s \sum_{j=1, j\neq i}^{c} P(\omega_j/x)$$

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

for $i = c+1$

$$R(\alpha_{c+1}/x) = \lambda_r$$

∴ the minimum risk is achieved if we decide ω_i if

$R(\alpha_i/x) \leq R(\alpha_{i+1}/x)$

i.e. $P(\omega_i/x) \geq 1 - \frac{\lambda_r}{\lambda_s}$ and reject otherwise.

2. If $\lambda_r = 0$, we always reject the class

3. If $\lambda_r > \lambda_r$, we always accept the class.

Q.5 [Maximum Likelihood Estimation] Suppose we have training samples $\{x_1, x_2, \ldots x_n\}$. Consider the following distributions.

(1) Exponential density: $F(x;\theta) = \theta e^{-\theta x}$, $x \geq 0$, $\theta > 0$

Find MLE for θ

$\Rightarrow$

Likelihood function

$$L(\theta, x) = \prod_{i=1}^{n} F(x;\theta)$$

Take log likelihood to maximize,

$$\ln L(\theta, x) = \sum_{i=1}^{n} \ln F(x;\theta)$$

$$\ln L(\theta, x) = \sum_{i=1}^{n} \ln (\theta e^{-\theta x_i})$$

$$\ln L(\theta, x) = \sum_{i=1}^{n} -(\theta x_i) + n \ln \theta$$

$$\ln L(\theta, x) = -\left(\sum_{i=1}^{n} x_i\right)\theta + n \ln \theta$$

$$\frac{d}{d\theta} \ln L(\theta, x) = -\left(\sum_{i=1}^{n} x_i\right) + \frac{n}{\theta}$$

..... (Taking derivatives)

But,

$$\frac{d}{d\theta} \ln L(\theta, x) = 0$$

$$\therefore 0 = -\left(\sum_{i=1}^{n} x_i\right) + \frac{n}{\theta}$$

$$\therefore \frac{n}{\theta} = \sum_{i=1}^{\theta} x_i$$

$$\therefore \theta = \frac{n}{\sum_{i=1}^{n} x_i}$$

$$\therefore \text{MLE of } \theta = \hat{\theta} = \frac{n}{\sum_{i=1}^{n} x_i} \quad //$$

(2) Uniform distribution Unif (θ_1, θ_2), find MLE for θ_1, θ_2

$$f(x/(\theta_1,\theta_2)) = \begin{cases} \frac{1}{\theta_2-\theta_1} & \text{for } \theta_1 \leq x \leq \theta_2 \\ 0 & \text{otherwise} \end{cases}$$

$\therefore$ Likelihood function has the form.

$$L(\theta_1,\theta_2) = \begin{cases} \frac{1}{(\theta_2-\theta_1)^n} & \text{for } \theta_1 \leq x_i \leq \theta_2 \ (i=1,\dots n) \\ 0, & \text{otherwise.} \end{cases}$$

We can see that for the given function we need to maximize $\frac{1}{(\theta_2-\theta_1)^n}$, which is a decreasing function.

$\therefore (\theta_2-\theta_1)^n$ should be minimum.

i.e. θ_2 should be minimum such that $\theta_2 \geq x_i$

θ_1 should be maximum such that $x_i \leq \theta_1$

So it can be concluded that

MLE of θ_2 is $\hat{\theta}_2 = \max [x_1 \ldots\ldots x_n]$

MLE of θ_1 is $\hat{\theta}_1 = \min [x_i \ldots\ldots x_n]$

Camlin Page
Date / /

Q.6 [Logistic Regression, MLE] In this problem, you need to use MLE to derive and build a logistic regression classifier (suppose the target/response $y \in \{0,1\}$):

(1) Suppose the classifier is $y = x^T\theta$, where θ contains the weight as well as bias parameter. The log-likelihood function is $LL(\theta)$, what is $\frac{\partial LL(\theta)}{d\theta}$

⇒ Logistic regression is actually at classification method, LR introduces an extra non-linearity over a linear classifier

$f(x) = x^T\theta$ by using a logistic function

$$x^T\theta = m\Sigma_{i=1}\ x_i\theta_i = x_1\theta_1 + \dots + x_m\theta_m$$

$$\sigma(f(x)) = 1/1+e^{-f(x)} \quad \text{i.e. } \sigma(z) = 1/1+e^{-z}$$

This is called the logistic function

Assume that,

$$p(y=1|x;\theta) = \sigma(x^T\theta)$$
$$p(y=0|x;\theta) = 1 - \sigma(x^T\theta)$$

Then the likehood (assuming data independence) is,

$$p(y|x;\theta) \sim \prod_i [\sigma(x^T\theta))^{y_i}(1-\sigma(x^T\theta))]$$

The log likelihood equation is,

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

$$\partial LL(\theta)/\partial \theta_j = \partial LL(\theta)/\partial p \cdot \partial p/\partial \theta_j = [(y/p) - (1-y/1-p)] \cdot \sigma(z)[1-\sigma(z)] \cdot x_j$$

$$= [y/p - 1 - y/1-p] \cdot p\left[\frac{1-p}{}\right] \cdot x_j$$

$$= \left[\frac{y(1-p) - y(p-1)}{(p-1)(1-p)} \times (1-p)\right] x_j$$

$$= [y-p]x_j$$

$$\therefore \partial LL(\theta)/\partial \theta_j = n \sum_{i=1} [y_i - \sigma(x^T\theta)] x_{ij}$$

$$\frac{\partial LL(\theta)}{\partial \theta_j} = \sum_{i=1}^{n} [y_i - \sigma(x^T\theta)] x_{ij}$$