CS559-B Homework 1

Thanapoom Phatthanaphan
20011296

Problem 1 (10pt): Independence and un-correlation

(1) (5pt) Suppose X and Y are two continuous random variables, show that if X and Y are independent, then they are uncorrelated.

(2) (5pt) Suppose X and Y are uncorrelated, can we conclude X and Y are independent? If so, prove it, otherwise, give one counterexample. (Hint: consider $X \sim Uniform[-1, 1]$ and $Y = X^2$)

Solution

(1) if X and Y are independent, then they are uncorrelated $\rightsquigarrow Cov(X,Y) = 0$

From $Cov(X,Y) = E(XY) - E(X)E(Y)$

Since X, Y are independent, then $E(XY) = E(X)E(Y)$

Therefore, $Cov(X,Y) = E(X)E(Y) - E(X)E(Y) = 0$ //

(2) From $Cov(X,Y) = E(XY) - E(X)E(Y)$

① For $-1 \le X \le 1$ $E(X) = \mu_X = 0$, $f(x) = \frac{1}{1-(-1)} = \frac{1}{2}$

② $E(Y) = E(X^2) = \int_{-1}^{1} x^2 f(x)\,dx = \int_{-1}^{1} x^2 \cdot (\frac{1}{2})\,dx = \frac{1}{2}\left[\frac{1}{3}(1)^3 - \frac{1}{3}(-1)^3\right] = \frac{1}{3}$

③ $E(XY) = E(X^3) = \int_{-1}^{1} x^3 f(x)\,dx = \int_{-1}^{1} x^3 \cdot (\frac{1}{2})\,dx = \frac{1}{2}\left[\frac{1}{4}(1)^4 - \frac{1}{4}(-1)^4\right] = 0$

From ①, ②, and ③ : $Cov(X,Y) = E(XY) - E(X)E(Y) = 0 - 0 \cdot \frac{1}{3} = 0$

Since $Cov(X,Y) = 0$ then, X and Y are uncorrelated,
but I cannot conclude that they are independent

Let $X = 1 \rightsquigarrow Y = (1)^2 = 1$
$X = 0 \rightsquigarrow Y = 0^2 = 0$
$X = -1 \rightsquigarrow Y = (-1)^2 = 1$
} the value of Y depends on the value of X

Therefore, X and Y are uncorrelated, but they are not independent. //

Problem 2 (10pt): [Likelihood Ratio] Suppose we consider two category classification, the class conditionals are assumed to be Gaussian, i.e., $p(x|\omega_1) = \mathrm{N}(2,1)$ and $p(x|\omega_2) = \mathrm{N}(6,1)$, based on prior knowledge, we have $P(\omega_2) = \frac{1}{3}$. We do not penalize for correct classification, while for misclassification, we put 1 unit penalty for misclassifying ω_1 to ω_2 and put 2 units for misclassifying ω_2 to ω_1. Derive the bayesian decision rule using likelihood ratio.

Solution

$P(\omega_2) = \frac{1}{3}$

$P(\omega_1) = 1 - \frac{1}{3} = \frac{2}{3}$

$P(x|\omega_1) = N(2,1)$

$P(x|\omega_2) = N(6,1)$

$$\lambda = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 2 & 0 \end{bmatrix}$$

From $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}$

$$P(x|\omega_1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-2)^2}{2}}$$

$$P(x|\omega_2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-6)^2}{2}}$$

decide ω_1 : $\frac{P(x|\omega_1)}{P(x|\omega_2)} > \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)}$

$$\frac{(1/\sqrt{2\pi})(e^{-\frac{(x-4)^2}{2}})}{(1/\sqrt{2\pi})(e^{-\frac{(x-6)^2}{2}})} > \frac{(1-0)}{(2-0)} \cdot \frac{(1/3)}{(2/3)}$$

$$e^{\frac{-(x-4)^2+(x-6)^2}{2}} > \frac{1}{4}$$

Apply ln : $\frac{-(x-4)^2+(x-6)^2}{2} > \ln(1/4)$

$$-x^2+8x-16+x^2-12x+36 > 2\ln(1/4)$$

$$-4x+20 > \ln(1/16)$$

$$x < \frac{\ln(1/16)-20}{-4}$$

$$x < 5.693$$

decide ω_2 : $\frac{P(x|\omega_1)}{P(x|\omega_2)} < \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)}$

$$\frac{(1/\sqrt{2\pi})(e^{-\frac{(x-4)^2}{2}})}{(1/\sqrt{2\pi})(e^{-\frac{(x-6)^2}{2}})} < \frac{(1-0)}{(2-0)} \cdot \frac{(1/3)}{(2/3)}$$

$$e^{\frac{-(x-4)^2+(x-6)^2}{2}} < \frac{1}{4}$$

Apply ln : $\frac{-(x-4)^2+(x-6)^2}{2} < \ln(1/4)$

$$-x^2+8x-16+x^2-12x+36 < 2\ln(1/4)$$

$$-4x+20 < \ln(1/16)$$

$$x > \frac{\ln(1/16)-20}{-4}$$

$$x > 5.693$$

Therefore, decide ω_1 if $x \leq 5.693$
decide ω_2 if $x > 5.693$

Note! for $x = 5.693$, I put it into the case of decide ω_1 because $P(\omega_1) > P(\omega_2)$

Problem 3 (15pt): [Minimum Probability of Error, Discriminant Function] Let the components of the vector $\mathbf{x} = [x_1, ..., x_d]^T$ be binary valued (0 or 1), and let $P(\omega_j)$ be the prior probability for the state of nature ω_j and $j = 1, ..., c$. We define

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

with the components x_i being statistical independent for all $\mathbf{x}$ in ω_j. Show that the minimum probability of error is achieved by the following decision rule:
Decide ω_k *if* $g_k(\mathbf{x}) \geq g_j(\mathbf{x})$ *for all* j *and* k, *where*

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

Solution

From the question : $P(x|\omega_j) = \sum_{i=1}^{d} P(x_i|\omega_j)$

$= \sum_{i=1}^{d} [P(x_i|\omega_j)]^{x_i} [1 - P(x_i|\omega_j)]^{1-x_i}$; Bernoulli

$= \sum_{i=1}^{d} p_{ij}^{x_i} (1 - p_{ij})^{1-x_i}$

From $g_j(x) = \ln P(x|\omega_j) + \ln P(\omega_j)$

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

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

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

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

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

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

Problem 4 (15pt): [Minimum Risk, Reject Option] In many machine learning applications, one has the option either to assign the pattern to one of c classes, or to reject it as being unrecognizable. If the cost for reject is not too high, rejection may be a desirable action. Let

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

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 error.

(1) (5pt) Derive the decision rule with minimum risk.

(2) (5pt) What happens if $\lambda_r = 0$?

(3) (5pt) What happens if $\lambda_r > \lambda_s$?

Solution

(1) From $R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|x)$, $R(\alpha_i|x) = \sum_{i \neq j} P(\omega_j|x) = 1 - P(\omega_i|x)$

For $i, j = 1, 2, \ldots, c \;/\; i \neq j$:

$$R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|x) = \sum_{\substack{j=1 \\ i=j}}^{c} (0 \cdot P(\omega_j|x)) + \sum_{\substack{j=1 \\ i \neq j}}^{c} \lambda_s \cdot P(\omega_j|x) = \lambda_s (1 - P(\omega_i|x))$$

For $i = c+1$: $R(\alpha_{i=c+1}|x) = \lambda_r$

Since $\min R(\alpha_i|x) = \max P(\omega_i|x)$ then, decide ω_i if $R(\alpha_i|x) < R(\alpha_{c+1}|x)$

$$\lambda_s(1 - P(\omega_i|x)) < \lambda_r$$
$$\lambda_s - \lambda_s P(\omega_i|x) < \lambda_r$$
$$P(\omega_i|x) < 1 - \frac{\lambda_r}{\lambda_s}$$

(2) If $\lambda_r = 0$: $P(\omega_i|x) > 1 - \frac{0}{\lambda_s} \rightsquigarrow P(\omega_i|x) > 1$

The probability cannot be greater than 1, so the option to decide ω_i will always reject.

(3) If $\lambda_r > \lambda_s$: $P(\omega_i|x) > 1 - \frac{\lambda_r}{\lambda_s} \rightsquigarrow 1 - \frac{\lambda_r}{\lambda_s}$ will be negative value

The probability cannot be negative value, so the option to decide ω_i will never reject.

Problem 5 (25pt): [Maximum Likelihood Estimation (MLE)] A general representation of a exponential family is given by the following probability density:

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

- 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(.) represents the exponential function.

(1) (5pt) Write down the expression of $A(\eta)$ in terms of $T(x)$ and $h(x)$.
(2) (10pt) Show that $\frac{\partial}{\partial \eta}A(\eta) = \mathrm{E}_\eta T(x)$ where $\mathrm{E}_\eta(.)$ is the expectation w.r.t $p(x|\eta)$.
(3) (10pt) Suppose we have n i.i.d samples $x_1, x_2, \ldots, x_n$, derive the maximum likelihood estimator for η. (You may use the results from part(b) to obtain your final answer)

Solution

(1) From the probability density function : $\int_{-\infty}^{\infty} p(x)dx = 1 \rightsquigarrow \int_{-\infty}^{\infty} p(x|\mu)dx = 1$

$$\int_{-\infty}^{\infty}\left[h(x)\, e^{\eta^T T(x) - A(\eta)}\right]dx = 1$$

$$\frac{1}{e^{A(\eta)}}\int_{-\infty}^{\infty}\left[h(x)\, e^{\eta^T T(x)}\right]dx = 1$$

$$\int_{-\infty}^{\infty}\left[h(x) e^{\eta^T T(x)}\right]dx = e^{A(\eta)}$$

$$A(\eta) = \ln\int_{-\infty}^{\infty}\left[h(x) e^{\eta^T T(x)}\right]dx$$

(2) $\frac{\partial}{\partial \eta}A(\eta) = \int_{-\infty}^{\infty}\frac{1}{\left[h(x)e^{\eta^T T(x)}\right]}dx\ \frac{\partial}{\partial \eta}\int_{-\infty}^{\infty}\left[h(x)e^{\eta^T T(x)}\right]dx$

$$= e^{-A(\eta)}\int_{-\infty}^{\infty}\left[h(x)\, e^{\eta^T T(x)}\frac{\partial}{\partial \eta}(\eta^T T(x))\right]dx \quad ; \quad \frac{\partial}{\partial \eta}(\eta^T T(x)) = T(x)\cdot I = T(x)$$

(I: identity matrix)

$$= e^{-A(\eta)}\int_{-\infty}^{\infty}\left[h(x)\, T(x)\cdot e^{\eta^T T(x)}\right]dx$$

$$= \int_{-\infty}^{\infty}\left[T(x)\, h(x)\, e^{\eta^T T(x) - A(\eta)}\right]dx \quad ; \quad h(x)\, e^{\eta^T T(x) - A(\eta)} = p(x|\eta)$$

$$\frac{\partial}{\partial \eta}A(\eta) = \int_{-\infty}^{\infty} T(x)\cdot p(x|\eta)dx = \mathrm{E}_\eta \cdot T(x)$$

Solution problem 5 (Cont.)

(3) From $L(\eta) = \prod_{i=1}^{n} p(x_i|\eta)$

log-likelihood : $\ln L(\eta) = \sum_{i=1}^{n} \ln p(x_i|\eta)$; $p(x_i|\eta) = h(x_i)\, e^{\eta^T T(x_i) - A(\eta)}$

$$= \sum_{i=1}^{n} \ln \left[h(x_i)\, e^{\eta^T T(x_i) - A(\eta)} \right]$$

Find MLE ⇝ maximize log-likelihood with respect to η

$$\frac{\partial}{\partial \eta} (\ln L(\eta)) = \frac{\partial}{\partial \eta} \left[\sum_{i=1}^{n} \ln h(x_i)\, e^{\eta^T T(x_i) - A(\eta)} \right] = 0$$

$$\underbrace{\frac{\partial}{\partial \eta} \left[\sum_{i=1}^{n} \ln h(x_i) \right]}_{0} + \frac{\partial}{\partial \eta} \left[\sum_{i=1}^{n} \ln e^{\eta^T T(x_i)} \right] - \frac{\partial}{\partial \eta} \left[n \cdot \ln e^{A(\eta)} \right] = 0$$

$$\sum_{i=1}^{n} T(x_i) - \frac{\partial}{\partial \eta} \left[n \cdot A(\eta) \right] = 0$$

From (2) : $\frac{\partial}{\partial \eta} \left[A(\eta) \right] = E_\eta (T(X))$ then,

$$\sum_{i=1}^{n} T(x_i) - n \cdot E_\eta (T(x_i)) = 0$$

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

A solution of the equation above might represent local / global minimum / maximum.

Problem 6 (25pt): [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) (5pt) Suppose the classifier is $y = x^T\theta$, where θ contains the weight as well as bias parameters. The log-likelihood function is $LL(\theta)$, what is $\frac{\partial LL(\theta)}{\partial \theta}$?

Solution

(1) From the log-likelihood function for logistic regression :

$$LL(\theta) = \sum_{i=1}^{n} [y_i \ln p_i + (1-y_i)\ln(1-p_i)] \quad ; \; p_i = 1/(1+e^{-(x^T\theta)})$$

$$\frac{\partial LL(\theta)}{\partial \theta} = \sum_{i=1}^{n} \left[y_i \left(\frac{1}{p_i}\right) \frac{\partial}{\partial \theta}(p_i) + (1-y_i)\left(\frac{1}{1-p_i}\right) \frac{\partial}{\partial \theta}(-p_i) \right]$$

Find $\frac{\partial p_i}{\partial \theta}$ $\rightsquigarrow$ $\frac{\partial p_i}{\partial \theta} = \frac{\partial}{\partial \theta}\left[\frac{1}{1+e^{-x^T\theta}}\right] = \frac{x^T e^{-x^T\theta}}{(1+e^{-x^T\theta})^2}$

then, $\frac{\partial LL(\theta)}{\partial \theta} = \sum_{i=1}^{n} \left[y_i (1+e^{-x_i^T\theta}) \left(\frac{x_i^T e^{-x_i^T\theta}}{(1+e^{-x_i^T\theta})^2}\right) - (1-y_i)\left(\frac{1+e^{-x_i^T\theta}}{e^{-x_i^T\theta}}\right)\left(\frac{x^T e^{-x_i^T\theta}}{(1+e^{-x_i^T\theta})^2}\right)\right]$

$$= \sum_{i=1}^{n} \left[\frac{y_i x_i^T e^{-x_i^T\theta}}{1+e^{-x_i^T\theta}} - (1-y_i)\left(\frac{x_i^T}{1+e^{-x_i^T\theta}}\right)\right]$$

$$= \sum_{i=1}^{n} \left[\frac{y_i x^T e^{-x_i^T\theta}}{1+e^{-x_i^T\theta}} - \frac{x_i^T}{1+e^{-x_i^T\theta}} + \frac{y_i x_i^T}{1+e^{-x_i^T\theta}}\right]$$

$$= \sum_{i=1}^{n} \left[\frac{y_i x^T (e^{-x_i^T\theta}+1)}{1+e^{-x_i^T\theta}} - \frac{x_i^T}{1+e^{-x_i^T\theta}}\right]$$

$$\frac{\partial LL(\theta)}{\partial \theta} = \sum_{i=1}^{n} \left(y_i x^T - \frac{x_i^T}{1+e^{-x_i^T\theta}}\right)$$