HW1

Saturday, October 1, 2022 9:43 PM

Omar Abdelmoteleb
I pledge my honor that I have abided by the Stevens Honor System. Omar A.

CS559-B HW1

Due: Oct. 4th, 2022

Problem 1 (5pt): 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 $AB \neq BA$. (No math derivation is needed).

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}, \quad P(B|A) = \frac{P(B \text{ and } A)}{P(A)}$$

Say we have a basket with an apple, orange, banana, potato and we randomly pick from it.

A = getting fruit
B = getting orange or vegetable

so,

A = {apple, banana, orange}
B = {orange, potato}

A and B $(A \cap B)$ = {orange}

$P(A \cap B) = 1/4$ $P(A) = 3/4$ $P(B) = 2/4$

$P(A \cap B) = 1/4$, $P(A) = \frac{3}{4}$, $P(B) = \frac{2}{4}$

$$P(A|B) = \frac{1/4}{3/4} = \frac{1}{3}$$

$$P(B|A) = \frac{1/4}{2/4} = \frac{1}{2}$$

Thus generally, not the same

$AB \neq BA$ $\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{bmatrix}$

Let $A = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$ $B = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$

$$AB = \begin{bmatrix} 1\cdot1+0\cdot3 & 1\cdot2+0\cdot4 \\ 1\cdot1+0\cdot3 & 1\cdot2+0\cdot4 \end{bmatrix}$$

$$BA = \begin{bmatrix} 1\cdot1+2\cdot1 & 1\cdot0+2\cdot0 \\ 3\cdot1+4\cdot1 & 3\cdot0+4\cdot0 \end{bmatrix}$$

$$AB = \begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix} \qquad BA = \begin{bmatrix} 3 & 0 \\ 7 & 0 \end{bmatrix}$$

So $AB \neq BA$

Problem 2 (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 $Y = X^2$)

(1) $E(XY) = E(X) \cdot E(Y)$

$$\begin{aligned} COV(X,Y) &= E(XY) - E(X)E(Y) \\ &= E(X)E(Y) - E(X)E(Y) \\ &= 0 \end{aligned}$$

Since $cov(x,y) = 0$, x and y are uncorrelated

(2) Consider $X \sim$ Uniform $[-1, 1]$ and $Y = X^2$

$E[X] = 0$ $\qquad E[XY] = E[X \cdot X^2] = E[X^3] = 0$

$$E[X^2 \mid X \geq 0] = \int_0^1 x^2 dx = \frac{1}{3}$$

$E[X^1 | X \le 0] = \int_0^{-1} x^1 dx = \frac{1}{3}$

$E(X^4 | X \le 0) = \int_{-1}^{0} -x^4 dx = -\frac{1}{5}$

Thus a proof is presented. The variables are uncorrelated yet independent.

Problem 2 (15pt): [Minimum Error Rate Decision] Let $\omega_{max}(x)$ be state of nature for which $P(\omega_{max}|x) \geq P(\omega_i|x)$ for all $i = 1, \ldots, c$.

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

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

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

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

(1) Given $P(w_{max}|x) \geq P(w_i|x)$ for all $i = 1, \ldots, c$

$$\sum_{i=1}^{c} P(w_{max}|x) \geq \sum_{i=1}^{c} P(w_i|x)$$

$$\Rightarrow c P(w_{max}|x) \geq 1 \quad \text{Divide } c$$

$$\Rightarrow \boxed{P(w_{max}|x) \geq \frac{1}{c}}$$

$\Rightarrow P(w_{max}|x) \geq \frac{1}{C}$

(2) Goal is $P(error) = 1 - \int P(w_{max}|x)p(x)dx$

Choose w_{max}. Error should be area that w_{max} is not true state of nature, so $1 - P(w_{max}|x)$

$$P(error) = \int P(error|x)p(x)dx$$

$$\Rightarrow \int 1 - P(w_{max}|x)p(x)dx$$

$$\Rightarrow 1 - \int P(w_{max}|x)p(x)dx$$

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

Given $P(w_{max}|x) \geq \frac{1}{C}$ and $P(error)$ from prior

$$P(error) \leq 1 - \int P(w_{max}|x)p(x)dx = 1 - \int \frac{1}{C}p(x)dx$$

$\int p(x)dx = 1$, so $P(error) = 1 - \frac{1}{C} = \frac{C-1}{C}$

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

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

Problem 4 (10pt): [Likelihood Ratio] Suppose we consider two category classification, the class conditionals are assumed to be Gaussian, i.e., $p(x|\omega_1) = \mathrm{N}(4,1)$ and $p(x|\omega_2) = \mathrm{N}(8,1)$, based on prior knowledge, we have $P(\omega_2) = \frac{1}{4}$. 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. Derive the bayesian decision rule using likelihood ratio.

Likelihood Ratio Test:

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

Since $P(\omega_2) = \frac{1}{4}$, $P(\omega_1) = \frac{3}{4}$

$$\lambda = \begin{bmatrix} 0 & 3 \\ 1 & 0 \end{bmatrix}$$

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

$P(x|\omega_1) = N(4,1)$ $\sigma = 1$, $\mu = 4$

$$\Rightarrow \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-4)^2}$$

$$\Rightarrow \frac{1}{\sqrt{2\pi}} e^{\cdot}$$

$$P(x|w_2) = N(8,1) \quad \sigma = 1, \mu = 8$$

$$\Rightarrow \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-8)^2}$$

LHS:

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

RHS: $\frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \frac{P(w_2)}{P(w_1)}$

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

$$= 1$$

Decide w_1 if LHS > 1

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

$$\Rightarrow -\frac{(x-4)^2}{2}+\frac{(x-8)^2}{2} > \log 1 = 0$$

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

$$\Rightarrow \underline{-x^2+8x-16} + \underline{x^2-16x+64} > 0$$

$$\Rightarrow \frac{-8x+48}{-8} \quad \frac{>0}{-8}$$

$$\Rightarrow x-6<0$$

$$\Rightarrow x<6$$

Decide w_1 if $x<6$, else w_2

Problem 5 (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 $\lambda_r > \lambda_s$?

Conditional Risk Zero-one :

$$R(a_i|x) = \sum_{j=1}^{c} \lambda(a_i|w_j) P(w_j|x)$$

$$= \sum_{i \neq j} P(w_j|x)$$

$$= 1 - P(w_i|x)$$

Expected loss on conditional risk:

$$R(a_i|x) = \sum_{j=1}^{c} \lambda(a_i|w_j) P(w_j|x)$$

(1) For $i, j = 1, \ldots, c$

$$R(a_i|x) = \sum_{j=1}^{c} \lambda(a_i|w_j) P(w_j|x)$$

$$R(a_i|x) = \sum_{j=1}^{c} \lambda(a_i|w_j) P(w_j|x)$$

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

$$= \lambda_s (1 - P(a_i|x))$$

(2) For $i = c+1$

$$R(a_{c+1}|x) = \lambda_r$$

Decide w_i if $R(a_i|x) \leq R(a_{c+1}|x)$

$$\Rightarrow P(w_i|x) \geq 1 - \frac{\lambda_r}{\lambda_s}$$, reject otherwise

(3) for $\lambda_r = 0$ /to accept:

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

If $\lambda_r = 0$, then $P(w_i|x) \geq 1$.

Since it is ≥ 1, we always reject.

Problem 6 (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)

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

Since this is a density function.

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

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

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

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

$$= \int p(x|\eta) T(x) dx$$

$$= E_\eta T(x)$$

(3) Consider n i.i.d samples $x_1, x_2, \ldots, x_n$
Derive MLE for η

Likelihood function for whole set of n i.i.d samples:

$$p(x_1, x_2, \ldots, x_n | \eta) = \prod_{k=1}^{n} p(x_k | \eta)$$

$$= \left(\prod_{k=1}^{n} h(x_k | \eta) \right) \exp\left\{ \eta^T \sum_{k=1}^{n} T(x_k) - n A(\eta) \right\}$$

$$L(\eta; x_1, x_2, \ldots, x_n) = \log\left(p(x_1, \ldots, x_n | \eta) \right)$$

$$= \log h(x_1, \ldots, x_n) + \eta^T \sum_{n=1}^{n} T(x_k) - n A(\eta)$$

$$= \log h(x_1, \ldots, x_n) + \eta' \sum_{k=1} T(x_k) - n A(\eta)$$

$$\Rightarrow \frac{\partial}{\partial \eta} (\eta; x_1, \ldots, x_n) = \sum_{k=1}^{n} T(x_k) - n \frac{\partial}{\partial \eta} A(\eta)$$

$$\Rightarrow \frac{\partial A(\eta)}{\partial \eta} = \frac{\sum_{k=1}^{n} T(x_k)}{n}$$

Problem 7 (20pt): [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 $\frac{\partial LL(\theta)}{\partial \theta}$?

(2) (15pt) Write the codes to build and train the classifier on Diabetes dataset (attached in Canvas). The Diabetes dataset contains 768 samples with 9 features for 2 outcomes. To simplify the problem, we only consider: **Glucose** and **BMI** as our features. Based on the simplified settings, train the model using gradient descent. Please show the classification results. (Note that (1) you could split the Diabetes dataset into train/test set. (2) You could visualize the results by showing the trained classifier overlaid on the train/test data. (3) You could tune several hyperparameters, e.g., learning rate, weight initialization method etc, to see their effects. (3) you **can not** use the package to directly train the model (e.g., sklearn.linear_model.LogisticRegression)).