HW 1

Tuesday, September 21, 2021 11:07 AM

Problem 1 (5pt): Provide an intuitive example to illustrate that $P(A|B)$ and $P(B|A)$ are in general not the same. (Math derivation is NOT needed)

The probality of A given B (P(A|B)) is not necessarily the same as the probability of B given A (P(B|A)). For example, imagine a random animal behind a curtain. If P(A) the probability that the animal is a dog & P(B) is the probability that the animal has 4 legs, the probability that the animal has 4 legs given it is a dog is basically 100%. However, if the animal has 4 leg this does not mean it is 100% a dog & it could be a number of other animals that have 4 legs. Hence, the probability that it is a dog given it has 4 legs is much less than the probability that it has 4 legs given it is a dog.

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 $Y = X^2$)

*I switched X+Y with A+B

(1) For two continuous random variables to be independent, they must be uncorrelated. This is because for two random variables to be independent, they must have no influence on each other.

P(A|B) = P(A) and P(B|A) = P(B)

no influence on each other.

$P(A|B) = P(A)$ and $P(B|A) = P(B)$

$P(A \cap B) = P(A) \cdot P(B)$

The above formulas hold true for continuous random variables.

The formula for correlation is the following:

$$\rho = \frac{Cov(A,B)}{\sqrt{Var(A)}\ \sqrt{Var(B)}}$$

This can be simplified to just the numerator since we are considering when correlation = 0.

The formula for covariance is the following:

$Cov(A,B) = E[AB] - E[A]E[B]$

$E[A]$ is the expected value or mean of A.

If $Cov(A,B) = 0$, then:

$E[AB] = E[A]E[B]$

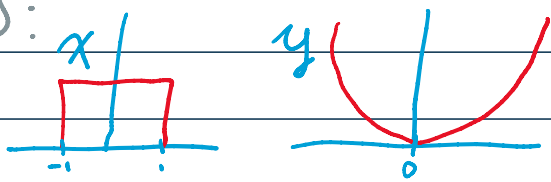
$$E[AB] = \int\int AB\, P_{AB}(A,B)\, dA\, dB$$
$$= \int\int AB\, P_A(A) P_B(B)\, dA\, dB$$
$$= \left(\int A P_A(A)\right)\left(\int B P_B(B)\right) dA\, dB$$
$$= \left[\int A \cdot P_A(A) dA\right]\left[\int B \cdot P_B(B) dB\right]$$
$$= E[A] \cdot E[B] = E[AB]$$

Therefore, if A & B are two independent continuous random variables, they must be uncorrelated.

(2) If A & B are uncorrelated, this <u>does not</u> mean they are necessarily independent. Consider the example of:

$X \sim U(-1, 1)$

$Y \sim X^2$



$x \sim U(-1, 1)$

$y \sim \chi^2$

y clearly depends on x, but these are uncorrelated.

$E[x] = \frac{(-1+1)}{2} = \frac{0}{2} = 0$

$E[y] = E[x^2] = 0$

If correlation = 0 then covariance = 0, therefore:

$E[xy] = E[x] \cdot E[y] = 0$

The correlation is 0 even though they are not independent.

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.

(1) (3pt) Interpret in words the meaning of p_{ij}.

(2) (12pt) 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*

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

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

The above expression states the probability for all i in range $1, ..., d$ & j in range $1, ..., c$ is equal to the probability that feature x (for all i) is 1 (& not 0) given the state of nature (for all j). It is the probability of a certain feature given a certain state of nature. It is

p_{ij}: It is the probability of a certain feature given a certain state of nature. It is a conditional probability.

(2) Minimum probability of error

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

↓ Discriminant function

$$g(x) = \ln(p(x|\omega_j) P(\omega_j)) = \ln(p(x|\omega_j)) + \ln(P(\omega_j))$$

Setting the above equations equals gives:

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

$$\rightarrow p(x|\omega_j) = \frac{P(\omega_j|x)\, p(x)}{P(\omega_j)} \quad *\text{Bayes}$$

$$\rightarrow \frac{p_{ij}}{1-p_{ij}} = \frac{P(x|\omega_j)}{1-P(x|\omega_j)} \qquad 1-p_{ij} = 1-P(x|\omega_j)$$

$$\rightarrow \ln(p(x|\omega_j)) = \sum_{i=1}^{d} x_i \ln\left(\frac{P(x|\omega_j)}{1-P(x|\omega_j)}\right) + \sum_{i=1}^{d} \ln(1-P(x|\omega_j))$$

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

$$= \prod_{i=1}^{d} p_{ij}^{x_i} (1-p_{ij})^{1-x_i} \quad *\text{Likelihood}$$

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

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

Problem 4 (10pt): [Likelihood Ratio] Consider two-class (i.e., ω_1, ω_2) classification with feature vector $\mathbf{x}$, suppose $p(\mathbf{x}|\omega_1)$ is standard normal distribution and $p(\mathbf{x}|\omega_2)$ is uniform distribution over $[-\frac{1}{2}, \frac{1}{2}]$, i.e.,

$$p(\mathbf{x}|\omega_1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$
$$p(\mathbf{x}|\omega_2) = 1,\ \mathbf{x} \in [-\frac{1}{2}, \frac{1}{2}]$$

Assuming zero-one loss and $P(\omega_1) = P(\omega_2)$, using likelihood ratio test to derive the corresponding decision rule.

$p(x|\omega_1) \sim N(1,1)$ $\quad p(x|\omega_2) \sim N(-\frac{1}{2}, \frac{1}{2})$

$p(x|\omega_1) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ $\quad p(x|\omega_2) = 1,\ x \in |-\frac{1}{2}, \frac{1}{2}|$

$p(\omega_1) = p(\omega_2)$

Since zero-one loss → Loss-Matrix $\lambda = \begin{bmatrix} \lambda_{11}=0 & \lambda_{12}=1 \\ \lambda_{21}=1 & \lambda_{22}=0 \end{bmatrix}$

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

★Decide ω_1 if LHS $\frac{p(x|\omega_1)}{p(x|\omega_2)}$ > RHS

RHS → $\frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)} = \frac{1-0}{1-0} \cdot 1 = 1$

$\Rightarrow \frac{p(x|\omega_1)}{p(x|\omega_2)} > 1 \rightarrow p(x|\omega_1) > p(x|\omega_2)$

$\rightarrow \frac{1}{\sqrt{2\pi}} e^{-x^2/2} > 1,\ x \in |-\frac{1}{2}, \frac{1}{2}|$

$\rightarrow e^{-x^2/2} > \sqrt{2\pi} \rightarrow \ln(e^{-x^2/2}) > \ln(\sqrt{2\pi})$

$\rightarrow \frac{-x^2}{2} > \ln(\sqrt{2\pi}) \rightarrow x^2 < -2\ln(\sqrt{2\pi})$

$\rightarrow x < \sqrt{-2\ln(\sqrt{2\pi})}$

→ No value of $x \in [-\frac{1}{2}, \frac{1}{2}]$ satisfy this

⇒ Choose ω_2 if $x \in [-\frac{1}{2}, \frac{1}{2}]$

If x is not within $[-\frac{1}{2}, \frac{1}{2}]$

$\rightarrow \frac{p(x|\omega_1)}{p(x|\omega_2)} > 0 \rightarrow p(x|\omega_1) > 0$

$\rightarrow \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2} > 0 \rightarrow e^{-x^2/2} > 0$

⇒ Choose ω_1 if $x \notin [-\frac{1}{2}, \frac{1}{2}]$

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

Loss function →

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

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 $\lambda_r > \lambda_s$?

(1) Decision rule w/ minimum risk

Loss function $\lambda(\alpha_i|\omega_j)$ = Loss incurred for deciding ω_i, action α_i, when the true state of nature is ω_j

Expected Loss (Conditional Risk) is the sum of all the possible losses each multiplied by the probability of that

Expected Loss (Conditional Risk) is the sum of all the possible losses each multiplied by the probability of that loss occuring.

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

Zero-One Loss → $\lambda(\alpha_i|\omega_j) = \begin{cases} 0, \text{if } i=j \\ 1, \text{if } i \neq j \end{cases}$

→ Condition Risk → $R(\alpha_i|x) = \sum_{i \neq j} P(\omega_j|x)$

→ $R(\alpha_i|x) = 1 - P(\omega_i|x)$ * 1 - P(correct choice)

1) $i = 1, \ldots, c$

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

$R(\alpha_i|x) = \lambda_s \sum_{j=1}^{c}{}_{j \neq 1} P(\omega_j|x)$

$R(\alpha_i|x) = \lambda_s (1 - P(\omega_i|x))$

2) $i = c+1$

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

=> Decide ω_i if $R(\alpha_i|x) < R(\alpha_{c+1}|x)$

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

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

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

=> Decision Rule: Decide ω_i if $P(\omega_i|x) > 1 - \lambda_r/\lambda_s$ + reject otherwise

(2) $\lambda_r = \emptyset$

$P(\omega_i|x) > 1 - 0/\lambda_s$

→ $P(\omega_i|x) > 1$

$P(w_i|x)$ can never be > 1.

$\Rightarrow$ Always reject

(3) $\lambda_r > \lambda_s$

$P(w_i|x) > 1 - \frac{\lambda_r}{\lambda_s}$
If $\lambda_r > \lambda_s \rightarrow \frac{\lambda_r}{\lambda_s} > 1$
$\Rightarrow 1 - \frac{\lambda_r}{\lambda_s} < 1$
$P(w_i|x)$ will never be $< 1 - \frac{\lambda_r}{\lambda_s}$
since it cannot be < 0.

$\Rightarrow$ Always accept

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.* (Vector)
- $A(\eta)$ 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) Expression for $A(\eta)$ in terms of $T(x)$ & $h(x)$.

Since it's a probability density function: $\int_{-\infty}^{\infty} f(t)dt = 1$

$\rightarrow \int_{-\infty}^{\infty} p(x|\eta)dx = 1$

$\Rightarrow \int_{-\infty}^{\infty} h(x) e^{\{\eta^T T(x) - A(\eta)\}} dx = 1$ * take ln of both sides

$\rightarrow -A(\eta) = -\ln\{\frac{1}{\int_{-\infty}^{\infty} h(x) e^{\{\eta^T T(x)\}} dx}\}$

$\Rightarrow \boxed{A(\eta) = \ln\{\int_{-\infty}^{\infty} h(x) e^{\{\eta^T T(x)\}} dx\}}$

(2) $\frac{\partial}{\partial \eta} A(\eta) = E_\eta T(x)$ * $E_\eta(.)$ = expectation w.r.t. $p(x|\eta)$

$$\begin{aligned}
\frac{\partial}{\partial \eta} A(\eta) &= \frac{\partial}{\partial \eta} \ln\left(\int_{-\infty}^{\infty} h(x) e^{(\eta^T T(x))} dx\right) \\
&= \frac{1}{\int_{-\infty}^{\infty} h(x) e^{(\eta^T T(x))} dx} \\
&= \frac{\int_{-\infty}^{\infty} h(x) e^{(\eta^T T(x) - A(\eta))} dx}{\int_{-\infty}^{\infty} h(x) e^{(\eta^T T(x))} dx} \\
&= \frac{\int_{-\infty}^{\infty} h(x) e^{(\eta^T T(x))} dx}{e^{A(\eta)}} \\
&= \int_{-\infty}^{\infty} h(x) e^{(\eta^T T(x) - A(\eta))} T(x) dx \\
&= \int_{-\infty}^{\infty} p(x|\eta) T(x) dx \\
&= E_\eta T(x)
\end{aligned}$$

$\Rightarrow \boxed{\frac{\partial}{\partial \eta} A(\eta) = E_\eta T(x)}$

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

Likelihood of η: $L(\eta) = \prod_{i=1}^{n} p(x_i|\eta)$

$\rightarrow \ln(L(\eta)) = \ln(\prod_{i=1}^{n} p(x_i|\eta))$ * ln of both sides

$= \ln(\prod_{i=1}^{n} (h(x_i) e^{(\eta^T T(x_i) - A(\eta))})$ * plug in $p(x|\eta)$

Maximize log likelihood $\ln(L(\eta))$ w.r.t. η

Take derivative w.r.t. η + set = 0

$\rightarrow \frac{\partial}{\partial \eta} \ln(L(\eta)) = 0$

$\rightarrow \frac{\partial}{\partial \eta} (e^{(\eta^T \sum_{i=1}^{n} T(x_i) - nA(\eta))}) = 0$

$\rightarrow \frac{\partial}{\partial \eta}\left(e^{(\eta^T \sum_{i=1}^n T(x_i) - nA(\eta))}\right) = 0$

$\rightarrow \frac{\partial}{\partial \eta}\left(\eta^T \sum_{i=1}^n T(x_i) - nA(\eta)\right) = 0$

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

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

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

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

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 $\frac{\partial LL(\theta)}{\partial \theta}$? → Find gradient of LL(θ) (expression of LL in week 4)

(2) (15pt) Write the codes to build and train the classifier on Iris plant dataset (https://archive.ics.uci.edu/ml/datasets/iris). The iris dataset contains 150 samples with 4 features for 3 classes. To simplify the problem, we only consider: (a) two classes, i.e., virginica and non-virginica; (b) The first 2 types of features for training, i.e., sepal length and sepal width. Based on these simplified settings, train the model using gradient descent. Please show the classification results. (Note that (1) you could split the iris 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 could use sklearn or other packages to load and process the data, but you **can not** use the package to train the model).

(1) The Likelihood Function is the following:

$p(t|w) = \prod_{n=1}^N y_n^{t_n}(1-y_n)^{1-t_n}$

Taking the negative log of this gives the Log Likelihood Error Function:

$E(w) = -\ln p(t|w)$

$= -\sum_{n=1}^N [t_n \ln y_n + (1-t_n)\ln(1-y_n)]$

We can then take the gradient of that, $LL(\theta)$, where $\theta = w$, to get the following:

$\frac{\partial LL(\theta)}{\partial \theta} = \sum_{n=1}^N (y_n - t_n)\phi_n$, where: $y_n = \sigma(w^T\phi_n)$, $\phi_n = x_n$

t_n = target vector

t_n = target vector
Φ_n = feature matrix

(2) Done in Python in a Jupyter Notebook submitted separately.