National Economics University, Vietnam
Faculty of Mathematics Economics
Data Science in Economics and Business
Machine Learning 1

Homework Week 1

Student: Nguyễn Anh Tú - ID: 11207333

1 Problem 1

Consider the following bivariate distribution $p(x, y)$ of two discrete random variables X and Y

Y \ X	x_1	x_2	x_3	x_4	x_5
y_1	0.01	0.02	0.03	0.1	0.1
y_2	0.05	0.1	0.05	0.07	0.2
y_3	0.1	0.05	0.03	0.05	0.04

Compute:

(a) Compute the marginal distributions $p(x)$ and $p(y)$

(b) Compute the conditional distribution $p(x|Y = y_1)$ and $p(x|Y = y_3)$

Solution.

(a) Formula for marginal distribution $p(X = x_i) = \sum_j p(x_i, y_j)$

Thus marginal distributions of $p(x)$ are:

$$\begin{aligned} p(X = x_1) &= p(x_1, y_1) + p(x_1, y_2) + p(x_1, y_3) \\ &= 0.01 + 0.05 + 0.1 = 0.16 \\ p(X = x_2) &= 0.02 + 0.1 + 0.05 = 0.17 \\ p(X = x_3) &= 0.03 + 0.05 + 0.03 = 0.11 \\ p(X = x_4) &= 0.1 + 0.07 + 0.05 = 0.22 \\ p(X = x_5) &= 0.1 + 0.2 + 0.04 = 0.34 \end{aligned}$$

Similarly, marginal distributions of $p(y)$ are:

$$\begin{aligned}
p(Y = y_1) &= p(x_1, y_1) + p(x_2, y_1) + p(x_3, y_1) + p(x_4, y_1) + p(x_5, y_1) \\
&= 0.01 + 0.02 + 0.03 + 0.1 + 0.1 = 0.26 \\
p(Y = y_2) &= 0.05 + 0.1 + 0.05 + 0.07 + 0.2 = 0.47 \\
p(Y = y_3) &= 0.1 + 0.05 + 0.03 + 0.05 + 0.04 = 0.27
\end{aligned}$$

(b) Formula for the conditional distribution $p(x|Y = y_i) = \dfrac{p(x, y_i)}{p(Y = y_i)}$

Thus conditional distributions of $p(x|Y = y_1)$ are:

$$\begin{aligned}
p(x = x_1|Y = y_1) &= \frac{p(x_1, y_1)}{p(Y = y_1)} = \frac{0.01}{0.26} \approx 0.038 \\
p(x = x_2|Y = y_1) &= 0.02 \div 0.26 \approx 0.077 \\
p(x = x_3|Y = y_1) &= 0.03 \div 0.26 \approx 0.115 \\
p(x = x_4|Y = y_1) &= 0.1 \div 0.26 \approx 0.385 \\
p(x = x_5|Y = y_1) &= 0.1 \div 0.26 \approx 0.385
\end{aligned}$$

Similarly, conditional distributions of $p(x|Y = y_3)$ are:

$$\begin{aligned}
p(x = x_1|Y = y_3) &= \frac{p(x_1, y_3)}{p(Y = y_3)} = \frac{0.1}{0.27} \approx 0.37 \\
p(x = x_2|Y = y_3) &= 0.05 \div 0.27 \approx 0.185 \\
p(x = x_3|Y = y_3) &= 0.03 \div 0.27 \approx 0.111 \\
p(x = x_4|Y = y_3) &= 0.05 \div 0.27 \approx 0.185 \\
p(x = x_5|Y = y_3) &= 0.04 \div 0.27 \approx 0.148
\end{aligned}$$

2 Problem 2

Consider two random variables x, y with joint distribution $p(x, y)$. Show that:

$$E_X[X] = E_Y[E_X[x|y]]$$

Here, $E_X[x|y]$ denotes the expected value of x under the conditional distribution $p(x|y)$.

Solution. Suppose that X and Y are two random discrete variables.

Since $E_X[x|y]$ denotes the expected value of x under the conditional distribution $p(x|y)$, we can write the formula $E_X[x|y] = \sum_x x \cdot p(x|y)$

The value of $\sum_x x \cdot p(x|y)$ will change for each value of y, therefore distribution of $E_X[x|y]$ values is also the distribution of y values. Then:

$$\begin{aligned} E_Y[E_X[x|y]] &= E_Y[\sum_x x \cdot p(x|y)] \\ &= \sum_y p(y) \sum_x x \cdot p(x|y) \\ &= \sum_y \sum_x p(y) \cdot p(x|y) \cdot x \\ &= \sum_x x \sum_y p(y) \cdot p(x|y) \\ &= \sum_x x \sum_y p(x, y) \end{aligned}$$

By definition of marginal distribution, $\sum_y p(x, y) = p(x)$. Thus:

$$\begin{aligned} E_Y[E_X[x|y]] &= \sum_x x \cdot p(x) \\ &= E_X[X] \quad \square \end{aligned}$$

3 Problem 3

Một cuộc điều tra cho thấy, ở 1 thành phố 20.7% dân số dùng sản phẩmX ,50% dùng loại sản phẩm Y và trong những người dùng Y thì 36.5% dùng X. Phỏng vấn ngẫu nhiên một người dân trong thành phố đó, tính xác xuất đề người ấy:

(a) Dùng cả X và Y.

(b) Dùng Y, và biết rằng người đó không dùng X.

Solution.

Gọi X là biến cố người được hỏi sử dụng sản phẩm X.
Y là biến cố người đó sử dụng sản phẩm Y.
Như vậy, theo đề bài ta có:

$$p(X) = 0.207$$

$$p(Y) = 0.5$$

$$p(X|Y) = 0.365$$

(a) Xác suất người được hỏi sử dụng cả X và Y là:

$$p(X, Y) = p(X|Y) \cdot p(Y) = 0.365 \cdot 0.5 = 0.1825$$

(b) Áp dụng định lý Bayes, xác suất người được hỏi dùng Y, biết rằng người đó không dùng X có thể được viết như sau:

$$\begin{aligned} p(Y|\overline{X}) &= \frac{p(\overline{X}|Y) \cdot p(Y)}{p(\overline{X})} \\ &= \frac{(1 - p(X|Y))p(Y)}{1 - p(X)} \\ &= \frac{(1 - 0.365) \cdot 0.5}{1 - 0.207} \approx 0.4 \end{aligned}$$

4 Problem 4

Prove the relationship: $V_X = E_X[x^2] - (E_X[X])^2$, which relates the standard definition of the variance to the raw-score expression for the variance

Solution.

$$\begin{aligned} V_X &= E_X[(X - E_X[X])^2] \\ &= E_X[X^2 - 2XE_X[X] + (E_X[X])^2] \\ &= E_X[X^2] - 2E_X[X \cdot E_X[X]] + (E_X[X])^2 \end{aligned}$$

By definition $E_X[X] = \sum xp(x)$ or $E_X[X] = \int xf(x)dx$, for random variable X, $E_X[X]$ is just a constant number, then $E[E_X[X]] = E_X[X]$. Therefore:

$$\begin{aligned} V_X &= E_X[X^2] - 2E_X[X] \cdot E_X[X] + (E_X[X])^2 \\ &= E_X[X^2] - 2(E_X[X])^2 + (E_X[X])^2 \\ &= E_X[X^2] - (E_X[X])^2 \quad \square \end{aligned}$$

5 Problem 5

Giả sử bạn đứng trước ba ô cửa mà đằng sau nó là một trong hai thứ: con dê hoặc một chiếc xe hơi giá trị. Bạn mong muốn mở trúng ô cửa có chiếc xe để được nhận nó (nếu mở trúng ô cửa có dê thì bạn phải rinh nó về nhà). Monty yêu cầu bạn chọn một trong các ô cửa. Dĩ nhiên bạn chọn một cách "hú họa" tại xác suất lúc này để nhận xe hơi ở mỗi ô cửa đều là $\frac{1}{3}$. Giả sử bạn chọn ô cửa số 1. Monty sẽ giúp bạn LOẠI TRỪ 1 ĐÁP ÁN SAI bằng cách mở một ô cửa có dê trong hai ô cửa còn lại (dĩ nhiên ông ta đã biết mỗi ô cửa có gì). Sau đó bạn được lựa chọn LẦN HAI: Giữ nguyên ô cửa ban đầu hay đổi sang ô cửa còn lại chưa được lật mở?

Solution.

Cách 1 Không mất tính tổng quát, giả sử ban đầu người chơi chọn ô cửa thứ nhất, bảng sau sẽ mô tả kết quả win (người chơi nhận được xe) hay lose (nhận được dê)

khi người chơi lựa chọn giữ/đổi lựa chọn của mình trong trường hợp xe nằm ở từng ô:

Lựa chọn	*Xe ở ô cửa 1*	*Xe ở ô cửa 2*	*Xe ở ô cửa 3*	*Xác suất thắng*
Giữ	win	lose	lose	$\frac{1}{3}$
Đổi	lose	win	win	$\frac{2}{3}$

Như vậy, theo bảng trên ta thấy xác suất người chơi nhận được xe ô tô là cao hơn khi lựa chọn là đổi sang một trong hai ô cửa còn lại. Vì vậy người chơi **nên đổi sang ô cửa còn lại** .

Cách 2 Không mất tính tổng quát, giả sử ban đầu người chơi chọn ô cửa thứ nhất.

Gọi A là biến cố xe nằm ở ô cửa thứ nhất.

Gọi B là biến cố một trong hai ô cửa còn lại được mở ra (và tất nhiên trong ô cửa đó sẽ có dê).

Vậy xác suất người chơi nhận được xe khi giữ nguyên lựa chọn của mình sau khi một trong hai ô cửa còn lại được mở cũng là xác suất xe nằm ở ô thứ nhất sau khi mở 1 trong 2 ô còn lại và bằng $p(A|B)$

Xác suất người chơi không nhận được xe khi giữ nguyên lựa chọn của mình sau khi mở 1 trong 2 ô cửa còn lại là $p(\overline{A}|B) = 1 - p(A|B)$

Áp dụng định lý Bayes ta có:

$$p(A|B) = \frac{p(B|A) \cdot p(A)}{p(B)}$$

Biết rằng khi xe nằm ở ô thứ nhất, Monty có thể chọn ngẫu nhiên 1 trong 2 ô còn lại để mở nên xác suất một ô được mở khi xe nằm ở ô thứ nhất $p(B|A) = \frac{1}{2}$

Xác suất xe nằm ở ô cửa thứ nhất $p(A) = \frac{1}{3}$

Xác suất một ô nào đó trong hai ô còn lại được lật mở:

$$\begin{aligned} p(B) &= p(\text{ô đó được mở khi xe nằm ở ô cửa 1}) \\ &+ p(\text{ô đó chắc chắn được mở vì xe nằm ở ô còn lại}) \\ &= p(\text{xe nằm ở ô cửa 1}) \cdot p(\text{một trong hai ô còn lại được mở}) \\ &+ p(\text{xe nằm ở ô cửa còn lại}) \cdot p(\text{một trong hai ô cửa còn lại chắn chắn được mở}) \\ &= \frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot 1 = \frac{1}{2} \end{aligned}$$

Vì vậy:

$$\begin{aligned} p(A|B) &= (\frac{1}{2} \cdot \frac{1}{3}) \div \frac{1}{2} = \frac{1}{3} \\ p(\overline{A}|B) &= 1 - p(A|B) = 1 - \frac{1}{3} = \frac{2}{3} \end{aligned}$$

Ta thấy xác suất người chơi không nhận được xe khi giữ nguyên lựa chọn của mình sau khi mở 1 trong 2 ô cửa còn lại cao hơn, vì vậy người chơi **nên đổi sang ô còn lại**.