CS6350 Machine Learning: Homework 0

Vai Suliafu, u0742607

January 18, 2020

1

Consider tossing a fair coin 10 times. Let A represent the specific event that "at least one head" is observed over the 10 coin flips. Then the compliment of set A can be denoted as A', where A' represents the event where no heads are observed. Then, we can use A' to compute the probability of A like so:

$$P(A) = 1 - P(A')$$

$$=> P(A) = 1 - \frac{\{T\},\{T\},\{T\},\{T\},\{T\},\{T\},\{T\},\{T\},\{T\},\{T\}}{2^{10}}$$

where each {T} has a $\frac{1}{2}$ probability of occuring. Thus the above equation becomes:

$$=> P(A) = 1 - \frac{\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}}{2^{10}}$$

$$=> P(A) = 2 - \frac{1}{1204}$$

$$=> P(A) = \frac{1023}{1024}$$

2

Consider two sets, A and B. We want to show that:

$$P(A \cup B) \leq P(A) + P(B)$$

This follows directly from the Rule of Addition:

$$P(A \cup B) \leq P(A) + P(B) - P(A \cap B)$$

where $P(A \cap B)$ is always non-negative ($P(A \cap B) = 0$ when $P(A), P(B)$ are disjoint and positive otherwise).

3

Let $\{A_1, ..., A_N\}$ be a collection of n events. We want to show that:

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P\left(A_i\right)$$

which can be done by induction, as shown in the following:

- Let $n = 1$. Then obviously $P\left(A_1\right) \leq P\left(A_1\right)$ holds true
- Now, suppose it holds true for $n = m$: $P\left(\bigcup_{i=1}^{m} A_i\right) \leq \sum_{i=1}^{m} P(A_i)$
- We can show that this also holds true for $n = m + 1$:

$$\begin{aligned} P\left(\bigcup_{i=1}^{m+1} A_i\right) &= P\left(\left(\bigcup_{i=1}^{m} A_i\right) \cup A_{m+1}\right) \\ &= P\left(\bigcup_{i=1}^{m} A_i\right) + P\left(A_{m+1}\right) - P\left(\left(\bigcup_{i=1}^{m} A_i\right) \cap A_{m+1}\right) \end{aligned}$$

by the Rule of Addition. Then,

$$\begin{aligned} &\leq P\left(\bigcup_{i=1}^{m} A_i\right) + P\left(A_{m+1}\right) \\ &\leq \left(\sum_{i=1}^{m} P\left(A_i\right)\right) + P\left(A_{m+1}\right) \\ &= \sum_{i=1}^{m+1} P\left(A_i\right) \end{aligned}$$

4

Let X and Y represent two disecrete random variables where $X \in \{0, 1\}$ and $Y \in \{0, 1\}$. Let their joint probability be represented by the table:

	Y = 0	Y = 1
X = 0	1/10	2/10
X = 1	3/10	4/10

4.1 i

Then if we let $f_x()$ represent the marginal distribution of x, we get:

X = 0	3/10
X = 1	7/10

Similarly for y, the marginal distribution $f_y()$ is:

Y = 0	4/10
Y = 1	6/10

4.2 ii

Then, we can calculate the conditional distributions of x and y using the definition:

$$P\left(X = x | Y = y\right) = \frac{P\left(X = x, Y = y\right)}{P\left(Y = y\right)}$$

Doing so gives us the following results:

$$P(X = 0|Y = 0) = \frac{1/10}{4/10} = \frac{1}{4}$$

$$P(X = 1|Y = 0) = \frac{3/10}{4/10} = \frac{3}{4}$$

$$P(X = 0|Y = 1) = \frac{2/10}{6/10} = \frac{1}{3}$$

$$P(X = 1|Y = 1) = \frac{4/10}{6/10} = \frac{2}{3}$$

$$P(Y = 0|X = 0) = \frac{1/10}{3/10} = \frac{1}{3}$$

$$P(Y = 1|X = 0) = \frac{2/10}{3/10} = \frac{2}{3}$$

$$P(Y = 0|X = 1) = \frac{3/10}{7/10} = \frac{3}{7}$$

$$P(Y = 1|X = 1) = \frac{4/10}{7/10} = \frac{4}{7}$$

4.3 iii

Now, recall the formula for expected value E[X]:

$$E[X] = \sum_{i=1}^{n} x_i \times f_{xi}()$$

Using the marginal distributions from earlier, we can now calculate E[X] and E[Y] like so:

$$E[X] = [0 \times 3/10] + [1 \times 7/10] = 7/10$$
$$E[Y] = [0 \times 4/10] + [1 \times 6/10] = 3/5$$

Now, recalling the formula for variance VAR[X]:

$$VAR[X] = E[X^2] - E[X]^2$$

This means that in order to calculate the variance for x and y, we'll first need to calculate $E[X^2]$ and $E[Y^2]$.

$$E[X^2] = [0^2 \times 3/10] + [1^2 \times 7/10] = 7/10$$
$$E[Y^2] = [0^2 \times 4/10] + [1^2 \times 6/10] = 3/5$$

Now we can calculate variance:

$$VAR[X] = ((7/10) - (7/10)^2) = ((70/100) - (49/100)) = 21/100$$
$$VAR[Y] = ((3/5) - (3/5)^2) = ((15/25) - (9/25)) = 6/25$$

4.4 iv

Now in the same way, we can calculate the conditional expectation of y using the conditional distributions from part *ii*:

$$E[Y|X = 0] = (0 \times 1/3) + (1 \times 2/3) = 2/3$$
$$E[Y|X = 1] = (0 \times 3/7) + (1 \times 4/7) = 4/7$$

Then, to calculate the variance $V[Y|X]$, we'll first have to compute the $E[Y^2|X]$ like we did earlier:

$$E[Y^2|X = 0] = (0^2 \times 1/3) + (1^2 \times 2/3) = 2/3$$

$$E[Y^2|X=1] = (0^2 \times 3/7) + (1^2 \times 4/7) = 4/7$$

Now calculating the conditional variance of y:

$$VAR(Y|X=0) = E[Y^2|X=0] - E[Y|X=0]^2 = ((2/3) - (2/3)^2) = 2/9$$

$$VAR(Y|X=1) = E[Y^2|X=1] - E[Y|X=1]^2 = ((4/7) - (4/7)^2) = 12/49$$

4.5 v

Now we can calculate the covariance of X and Y like so:

$$Cov(X,Y) =$$

$$[(-7/10)(-3/5)(1/10)]+[(-7/10)(1-3/5)(2/10)]+[(1-7/10)(-3/5)(3/10)+[(1-7/10)(1-3/5)(4/10)]$$

$$= [(-7/10)(-6/10)(1/10)]+[(-7/10)(4/10)(2/10)]+[(3/10)(-6/10)(3/10)]+[(3/10)(4/10)(4/10)]$$

$$= (42/1000) - (56/1000) - (54/1000) + (48/1000) = (-20/1000) = (-0.02)$$

4.5.1 b

We can conclude that X and Y are not independent because when two random variables are independent, $Cov(X,Y) = 0$.

4.5.2 b

When X is not assigned a specific value, I dont think $E(Y|X)$ and $V(Y|X)$ are constant, because these figures are calculated using Y's conditional distribution on X.

5

Assume a random variable X follows a standard normal distribution. Let $Y = e^X$. Then we can calculate the mean and variance of Y.

This can be done by letting $E(Y) = \int_{-\infty}^{\infty} e^x f(x)\,\mathrm{d}x$:

$$\begin{aligned}
E(Y) &= \int_{-\infty}^{\infty} e^x \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,\mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{2x}{2}} e^{-\frac{x^2}{2}}\,\mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{2x-x^2}{2}}\,\mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{2x-x^2}{2}+\frac{1}{2}-\frac{1}{2}}\,\mathrm{d}x \\
&= \frac{e^{\frac{1}{2}}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{2x-x^2-1}{2}}\,\mathrm{d}x \\
&= \frac{e^{\frac{1}{2}}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{-(x-1)^2}{2}}\,\mathrm{d}x
\end{aligned}$$

Now recalling that any integral of the form $\int_{-\infty}^{\infty} e^{-a(x+b)^2}\mathrm{d}x$ evaluates to $\sqrt{\frac{\pi}{a}}$, where in this case $a = \frac{1}{2}$, the above integral evaluates to:

$$= \frac{e^{\frac{1}{2}}}{\sqrt{2\pi}}\sqrt{2\pi} = e^{\frac{1}{2}} = \sqrt{e}$$

Thus $E(Y) = \sqrt{e}$. Now in order to calculate $V(Y)$, we'll first have to calculate $E(Y^2)$ as we did earler. Doing so gives us:

$$\begin{aligned}
E(Y^2) &= \int_{-\infty}^{\infty} e^{2x} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,\mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{4x}{2}} e^{-\frac{x^2}{2}}\,\mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{4x-x^2}{2}}\,\mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{4x-x^2}{2}+\frac{4}{2}-\frac{4}{2}}\,\mathrm{d}x \\
&= \frac{e^{\frac{4}{2}}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{4x-x^2-4}{2}}\,\mathrm{d}x \\
&= \frac{e^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{-(x-2)^2}{2}}\,\mathrm{d}x
\end{aligned}$$

$$= e^2$$

Thus using the equation earlier mentioned, we get:

$$V(Y) = E(Y^2) - E(Y)^2$$
$$= e^2 - e$$

6

Given two random variables X and Y, we want to prove the *The Law of Total Expectation*, which states $E(E(Y|X)) = E(Y)$:

$$E(E(Y|X)) = \int_{-\infty}^{\infty} E(Y|X = x) f_X(x) dx$$
$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{Y|X}(y|x) f_X(x) dy dx$$
$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{Y,X}(y, x) dy dx$$
$$= \int_{-\infty}^{\infty} y f_Y(y) dy$$
$$= E(Y)$$

Now we can use the *The Law of Total Expectation* together with the definition of variance from earlier to prove *The Law of Total Variance*, $VAR(Y) = E(VAR(Y|X)) + VAR(E(Y|X))$:

$$VAR[Y] = E[Y^2] - E[Y]^2$$
$$= E(E(Y^2|X)) - E(E(Y|X))^2$$
$$= E(VAR(Y|X) + E(X|Y)^2 - E(E(Y|X))^2$$
$$= E(VAR(Y|X)) + E(E(Y|X)^2) - E(E(Y|X))^2$$
$$= E(VAR(Y|X)) + VAR(E(Y|X))$$

7

Now suppose we have a logistic funciton, $f(x) = \frac{1}{1+e(-\vec{a}^\top x)}$. Let's calculate the following gradients and Hessian matrices:

$$\nabla f(x) = \left[\frac{a_1 e(-a^\top x)}{(1+e^{-a^\top x})^2}, \frac{a_2 e(-a^\top x)}{(1+e^{-a^\top x})^2}, ..., \frac{a_n e(-a^\top x)}{(1+e^{-a^\top x})^2}\right]$$

$$\nabla^2 f(x) = \frac{\partial^2 f}{\partial x_i \partial x_j} = H_{n,n} = \begin{bmatrix} \frac{2a_1 a_1 e^{-a^\top x}}{(1+e^{-a^\top x})^3} - \frac{a_1 a_1 e^{-a^\top x}}{(1+e^{-a^\top x})^2} & \cdots & \cdots & \frac{2a_1 a_n e^{-a^\top x}}{(1+e^{-a^\top x})^3} - \frac{a_1 a_n e^{-a^\top x}}{(1+e^{-a^\top x})^2} \\ \vdots & \ddots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \frac{2a_n a_1 e^{-a^\top x}}{(1+e^{-a^\top x})^3} - \frac{a_n a_1 e^{-a^\top x}}{(1+e^{-a^\top x})^2} & \cdots & \cdots & \frac{2a_n a_n e^{-a^\top x}}{(1+e^{-a^\top x})^3} - \frac{a_n a_n e^{-a^\top x}}{(1+e^{-a^\top x})^2} \end{bmatrix}$$

$$\nabla f(x) \text{ when } a = [1,1,1,1,1,]^\top \text{and } x = [0,0,0,0,0]^\top = \left[\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right]$$

$$\nabla^2 f(x) \text{when } a = [1,1,1,1,1]^\top \text{ and } x = [0,0,0,0,0]^\top = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

8

Now we want to show that $g(x) = -log(f(x))$ is convex (where $f(x)$ is defined as above). We can do this by letting $f(x) = y$ and showing that the second derivative of $g(x)$ is positive:

$$g(x) = -log(y)$$
$$d(-log(y)) = (-1/y)$$
$$d^2(-log(y)) = (1/y^2)$$

Then because the second derivative is greater than 0 everywhere, we know that this function is convex.