CSDS 440: Assignment 1

Shaochen (Henry) ZHONG, sxz517

Mingyang Tie, mxt497

Due on and submitted on 09/11/2020

Problem 1

For a dice roll, let $A = \{1, 2\}$, $B = \{2, 3, 4\}$, and $C = \{1, 3\}$. We have $P(A) = \frac{1}{3}$, $P(B) = \frac{1}{2}$, and $P(C) = \frac{1}{3}$.

Now we have:

$$P(A, B) = \{2\} = \frac{1}{6} = P(A)P(B) \text{ Thus } A \text{ is independent of } B.$$
$$P(A \mid C) = \frac{\{1\}}{\frac{1}{3}} = \frac{1}{2}$$
$$P(B \mid C) = \frac{\{3\}}{\frac{1}{3}} = \frac{1}{2}$$
$$P(A, B \mid C) = \emptyset = 0 \neq P(A \mid C) \cdot P(B \mid C)$$

And this proven the statement.

Problem 2

We can view this problem as having two points x_1 and x_2 uniformaly distributed on a line with a length of $\sqrt{2}$, since this is the length of function $x + y = 1$ in interval $(0, 1)$ is $\sqrt{2}$). Let x be a random varianle $\in [0, \sqrt{2}]$, the PDF of this x would be:

$$f(x) = \begin{cases} \frac{1}{\sqrt{2}} & x \in [0, \sqrt{2}] \\ 0 & \text{otherwise} \end{cases}$$

For x_1 and x_2, since the placement of two points are independent, we have the joint PDF of x_1, x_2 to be:

$$f(x_1, x_2) = \begin{cases} \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} = \frac{1}{2} & x_1, x_2 \in [0, \sqrt{2}] \\ 0 & \text{otherwise} \end{cases}$$

Since the square distance is $D = (x_1 - x_2)^2$, its expected value is $E[(x_1 - x_2)^2]$, which is:

$$\begin{aligned} E[(x_1 - x_2)^2] &= \int_0^{\sqrt{2}} \int_0^{\sqrt{2}} (x_1 - x_2)^2 \cdot f(x_1, x_2) \cdot dx_1 dx_2 \\ &= \frac{1}{2} \int_0^{\sqrt{2}} \int_0^{\sqrt{2}} (x_1^2 - 2x_1x_2 + x_2^2) \cdot dx_1 dx_2 \\ &= \frac{1}{2} \int_0^{\sqrt{2}} \frac{x_1^2}{3} - 2\frac{x_1^2}{2}x_2 + x_2x_1 \Big|_0^{\sqrt{2}} \cdot dx_2 \\ &= \frac{1}{2} \cdot \frac{2}{3} \\ &= \frac{1}{3} \end{aligned}$$

Problem 3

First Task

Goal: Determine if an English sentence is grammatically correct.

Example: A lot of correctly and incorrectly written sentences, each labeled accordingly.

Performance Measure: Run the program and see how many sentences did the program correctly recognized.

Hypothesis Space: The program will probably try out many different types of possible language grammar during the training, this will be the hypothesis space. The program will, if sucessfully implemented, eventually capture the target concept (English grammar) among these tried hypothe-

ses.

Method: Supervised learning, as we may have pre-labbled sentences as examples.

Second Task

Goal: Determine the right time to buy or sell a particular stock to maximal profit.

Example: News articles related to the stock.

Performance Measure: If the result trading strategy may perform better than the stock market average return, e.g. S&P 500.

Hypothesis Space: The program will probably try out different conjuctions of news articles properties – e.g. emotion of vocabularies, frequency of words, etc – and connect them to the buy or sell behavior.

Method: Unsurpervised learning might be better for this task as we have no pre-labbled example. The program will hopefully be able to recognize some pattern of the cluster of article-stock combos which will grown up (and also drop down), and connect the right trading decision to these clusters.

Problem 4

(i)

Because you may not have every possible input to be in your traning example, thus a pure memerization approach won't be useful when facing input outside of seen examples. We should capture a concept that is applicable to general cases, base on the features and patterns of traning example, but not the just try to map the input with its "memorized" examples without any intelligent reasoning in between.

A human learning example is that a human student may memorize the area of a retangle with x length and y width, but unless the student may capture the concept of $area = x \cdot y$, this student may not be able to find out the area of a retangle with edge of x' and y' if such x', y' is not in the example represenation.

(ii)

It is possible the target concept of a task is not contained in its hypothesis space. If we set the hypothesis space to be big enough to contain every possible hypotheses, so that the target hypothesis is guaranteed to be include; this hypothesis space will also include the concept of simply memorize all your examples. This memerization concept will perform just as well – if not better – than your target concept and might became the final concept produced by the program, in this case the program is not learning.

A human learning example, still on area of retangle with edge x and y would be if we expose the human to every hypotheses – $x+y$, $x-y$, xy, $f(x,y)$, filling smaller squares, etc – the hypothsis of simply remember the area of every given example will be in this hypothesis space. Since the performance of this memerization concept is just as good as calculating xy (the target concept) on examples, the human might just do the memorization without any actual learning.

(iii)

Because if the set of example represenation is heavily lacked or flawed, the produced concept upon it might also be lacked or flawed. Say we have a set of example represenation of calculating abosulte value of a number X ($|X|$). If this example representation includes no example of X being negative, the produced concept might be $|X| = X$ and it will perform perfectly. However, when given $X < 0$ the output of this concept will be false, and we may avoid this mistake by include example of $X < 0$ in our example represenation in first place.