Written by: Professor Wu
DS 4420: Machine Learning 2

Assignment 3

Question 1. (20 pts) Given the probability table below where X represents the probability of you having a good or bad date and Y represents the potential topics you could talk about during a date.

	$x = good$	$x = bad$
$y = ex$	0	0.1
$y = food$	0.4	0.1
$y = travel$	0.2	0.1
$y = weather$	0.0	0.1

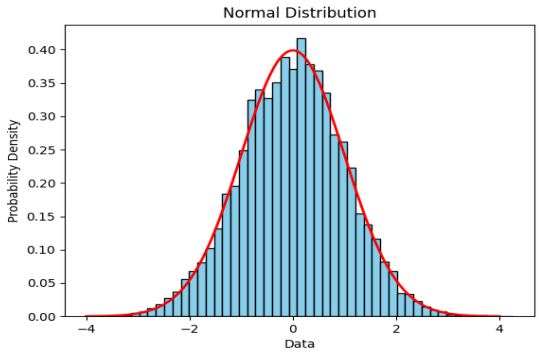
1) What is the probability of having a good date? bad date?
2) What is the probability that you would talk about food and have a good date?
3) What is the probability of you talking about food or travel?
4) What is the probability of you talking about travel **or** having a good date?
5) What is the probability of you having a bad date given you talk about your ex the whole time?
6) What is the probability of you having a good date given that you talk about the weather?
   - Solve this problem with conditional probability.
   - Solve this problem using Bayes' Rule.
   - Note: The solution should be identical.
7) Which topic should you talk about?

Question 2. (30 pts) Go to the course website and find the 3 csv file

`SAT.csv`
`Lunch_wait_time.csv`
`Student_age.csv`

For each file

1) Identify the probability distribution $p(x)$ that describes each distribution.
2) Generate the histogram, and super-impose the probability distribution on top of it.
   - Remember that the histogram describes the probability distribution of the data.
   - If your $p(x)$ is good, the equation should match the histogram really closely.
   - Visually, it should look something like the image below.



3) Use 2 separate approaches to calculate the probability that someone scored higher than 1200 on SAT
   (a) First use the counting approach where $p(x \geq 1200) = \frac{\text{num of people above 1200}}{\text{out of all the people}}$
   (b) Using $p(x)$ and numpy automatic integration to calculate $\int_{1200}^{1600} p(x)\, dx$
   (c) How different are the results?
4) Calculate the probability you have to wait more than 5 min.
   (a) First use the counting approach where $p(x \geq 5) = \frac{\text{num of people waited longer than 5}}{\text{out of all the people}}$
   (b) Using $p(x)$ and numpy automatic integration to calculate $\int_{5}^{\infty} p(x)\, dx$
   (c) How different are the results?

Question 3. (30pts) Given the following distribution

$$p(x) = \begin{cases} x^2 - x + 1 & \forall\ 0 \leq x \leq b \\ 0 & \text{everywhere else} \end{cases}$$

Note: This expression says that between the interval of 0 and b, the probability distribution is $p(x) = x^2 - x + 1$ and $p(x) = 0$ everywhere else.

Given that the total probability is always 1, it implies that the area between 0 and b must add up to 1, implying that the integral between 0 and b must therefore also add up to 1 where

$$\int_0^b x^2 - x + 1\ dx = 1.$$

1) Use Numpy to find the value of b?
   - Hints on how to solve this problem.
   - If you don't remember how to take integrals, make sure you watch my video on calculus refreshers.
     `https://youtu.be/N62Hdt5TV9g?si=Phynx6U_UeJRPEcS`
   - Note that if you have an expression $x^2 - x - 2 = 0$ (which we know the solution as $(x+1)(x-2) = 0$), we can solve a more complex problem with with numpy.root. Try to understand the following command.
   - ```
     np.roots([1, -1, -2])
     >>>  array([ 2., -1.]
     ```

2) Solve the integral by hand and find the probability $p(0 \leq x \leq 0.3)$.

Question 4. (20pts) Given the following distribution

$$p(x) = \begin{cases} \frac{1}{x+1} & \forall\ 0 \leq x \leq b \\ 0 & \text{everywhere else} \end{cases}$$

Calculate by hand the value of b.
     ```