Write your solutions to the following problems by either typing them up or handwriting them on another piece of paper. Homeworks are due to Gradescope by 11:59pm on the due date. You can use a slip day to extend the deadline by 24 hours.

Homework will be evaluated not only on the correctness of your answers, but on your ability to present your ideas clearly and logically. You should **always explain and justify** your conclusions, using sound reasoning. Your goal should be to convince the reader of your assertions. If a question does not require explanation, it will be explicitly stated.

Homeworks should be written up and turned in by each student individually. You may talk to other students in the class about the problems and discuss solution strategies, but you should not share any written communication and you should not check answers with classmates. You can tell someone how to do a homework problem, but you cannot show them how to do it.

For each problem you submit, you should **cite your sources** by including a list of names of other students with whom you discussed the problem. Instructors do not need to be cited.

This homework will be graded out of 50 points. The point value of each problem or sub-problem is indicated by the number of avocados shown.

Problem 1. Reflection and Feedback Form

Make sure to fill out this Reflection and Feedback Form, linked here for two points on this homework! This form is primarily for your benefit; research shows that reflecting and summarizing knowledge helps you understand and remember it.

Problem 2. Conditional Independence

Given the following three statements:

$$
\begin{array}{llll}
(1) & P(X, Y | E) & = & P(X|E)P(Y|E) \\
(2) & P(X|Y, E) & = & P(X|E) \\
(3) & P(Y|X, E) & = & P(Y|E)
\end{array}
$$

In this problem, you will need to show the three of them are equivalent. You may assume none of the probabilities listed above can be zero.

a) Show that (1) and (2) are equivalent. That is, $(1) \Rightarrow (2)$ and $(2) \Rightarrow (1)$

> **Solution: (1) ⇔ (2):**
> $$P(X, Y|E) = P(X|E)P(Y|E)$$
> $$\Leftrightarrow P(X|Y, E)P(Y|E) = P(X|E)P(Y|E)$$
> $$\overset{P(Y|E)\neq 0}{\Leftrightarrow} P(X|Y, E) = P(X|E)$$

b) Show that (2) and (3) are equivalent. That is, $(2) \Rightarrow (3)$ and $(3) \Rightarrow (2)$

> **Solution: (2) ⇔ (3):**
>
> $$P(X|Y, E) = P(X|E) \Leftrightarrow P(X|Y, E)P(Y|E) = P(X|E)P(Y|E)$$
> $$\Leftrightarrow P(Y|X, E)P(X|E) = P(X|E)P(Y|E)$$
> $$\overset{P(X|E) \neq 0}{\Leftrightarrow} P(Y|X, E) = P(Y|E)$$

c) 😊😊😊 Show that (1) and (3) are equivalent. That is, (1) ⇒ (3) and (3) ⇒ (1)

> **Solution: (1) ⇔ (3):**
> $$P(X, Y|E) = P(X|E)P(Y|E)$$
> $$\Leftrightarrow P(Y|X, E)P(X|E) = P(X|E)P(Y|E)$$
> $$\overset{P(X|E) \neq 0}{\Leftrightarrow} P(Y|X, E) = P(Y|E)$$

Problem 3. Independence, and Markov Property

A Markov Chain is a series of random variables that each one of them only depends on the preceding one. Don't get overwhelmed by its name, let's consider the following simple example:

$$X_1 \to X_2 \to X_3$$

which means X_2 only depends on X_1 and X_3 only depends on X_2. That is, we have $P(X_3|X_1, X_2) = P(X_3|X_2)$

a) 😊😊😊 Show that their joint distribution is

$$P(X_1, X_2, X_3) = P(X_1)P(X_2|X_1)P(X_3|X_2)$$

> **Solution:** Apply the product rule:
>
> $$P(X_1, X_2, X_3) = P(X_1, X_2)P(X_3|X_2, X_1) \qquad = P(X_1)P(X_2|X_1)P(X_3|X_2, X_1)$$
> $$= P(X_1)P(X_2|X_1)P(X_3|X_2)$$

b) 😊😊😊😊 Show the Markov Property that given the present, the future and the past are independent in the Markov chain. That is,

$$P(X_1, X_3|X_2) = P(X_1|X_2)P(X_3|X_2)$$

> **Solution:**
> $$P(X_1, X_3|X_2) = \frac{P(X_1, X_2, X_3)}{P(X_2)}$$
> $$= \frac{P(X_1, X_2)P(X_3|X_1, X_2)}{P(X_2)}$$
> $$= \frac{P(X_2)P(X_1|X_2)P(X_3|X_2)}{P(X_2)}$$
> $$= P(X_1|X_2)P(X_3|X_2)$$

Problem 4. Independence and Conditional Independence

Consider the sample space $S = \{a, b, c, d, e, f, g\}$ with associated probabilities given in the table below.

outcome	a	b	c	d	e	f	g
probability	$\frac{5}{21}$	$\frac{2}{21}$	$\frac{1}{21}$	$\frac{4}{21}$	$\frac{2}{21}$	$\frac{4}{21}$	$\frac{3}{21}$

Let $X = \{d, e\}$ and $Y = \{e, f\}$. Remember to show your work for all calculations.

a) 😊😊😊😊 Are X and Y independent?

> **Solution:** No, X and Y are not independent because
>
> $$P(X) = \frac{6}{21}$$
> $$P(Y) = \frac{6}{21}$$
> $$P(X \cap Y) = P(\{e\}) = \frac{2}{21},$$
>
> so $P(X) \cdot P(Y) \neq P(X \cap Y)$.
>
> More intuitively, we can see that these events won't be independent because the proportion of the sample space taken up by X is $P(X) = \frac{6}{21} = \frac{2}{7}$, whereas the proportion of Y taken up by X is $P(X|Y) = \frac{2/21}{6/21} = \frac{1}{3}$.

b) 😊😊😊😊😊 Determine if X and Y are conditionally independent given each of the following events Z.

1. $Z = \{a, b, d, e, f, g\}$
2. $Z = \{a, d, e, f, g\}$
3. $Z = \{d, e, f, g\}$

> **Solution:** For X and Y to be conditionally independent given Z, we need to have
>
> $$P(X|Z) \cdot P(Y|Z) = P((X \cap Y)|Z)$$
>
> Our job is to check if this condition holds, for each of three sets Z. By substituting the definition of conditional probability for all three probabilities, the condition we are checking is
>
> $$\frac{P(X \cap Z)}{P(Z)} \cdot \frac{P(Y \cap Z)}{P(Z)} = \frac{P(X \cap Y \cap Z)}{P(Z)}$$
>
> Since we have to look at 3 different values of Z, we need to compute the above 6 probabilities, 3 times. That's a perfectly valid way of approaching the problem.
>
> But, that's a lot of work! But, there's an observation that makes the process much simpler.
>
> Recall that $X = \{d, e\}$ and $Y = \{e, f\}$. All three options for Z include e, f, and g. This means that $X \cap Z = \{d, e\}$, $Y \cap Z = \{e, f\}$, and $X \cap Y \cap Z = \{e\}$, for all three sets Z.
>
> This means that $P(X \cap Z) = \frac{6}{21}$, $P(Y \cap Z) = \frac{6}{21}$, and $P(X \cap Y \cap Z) = \frac{2}{21}$ are all the same, in each of the three subparts (note that this is only true due to our specific choices of Z; if we had $Z = \{a, b, d, g\}$ this would no longer be true). The only thing that's different is $P(Z)$. Our job is now just to verify if

$$\frac{\frac{6}{21}}{P(Z)} \cdot \frac{\frac{6}{21}}{P(Z)} = \frac{\frac{2}{21}}{P(Z)}$$

While we can compute $P(Z)$ directly in each case and plug it into the equation, we can actually just solve the above equation for $P(Z)$. Once we do that, all we need to do in each of the three cases is check if $P(Z)$ is equal to this result; if it is equal, then X and Y are conditionally independent given Z, and if it is not equal, then X and Y are not conditionally independent given Z.

Solving the above equation gives $P(Z) = \frac{18}{21}$. This means for all three options we just need to determine if $P(Z) = \frac{18}{21}$.

1. When $Z = \{a, b, d, e, f, g\}$, $P(Z) = \frac{20}{21}$. Thus, X and Y are not conditionally independent given this Z.

2. When $Z = \{a, d, e, f, g\}$, $P(Z) = \frac{18}{21}$. Thus, X and Y are conditionally independent given this Z.

3. When $Z = \{d, e, f, g\}$, $P(Z) = \frac{15}{21}$. Thus, X and Y are not conditionally independent given this Z.

Problem 5. Independence and Complements

Let E and F be two events in a sample space S, with $0 < P(F) < 1$.

a) 😊😊😊😊 If $P(E|F) = P(E|\overline{F})$, must it be true that E and F are independent? Provide a proof of independence, or give a counterexample by specifying a sample space S and two dependent events E and F that satisfy the given conditions.

Solution: Yes, it must be true that E and F are independent in this case.

Since $0 < P(F) < 1$, we know that $P(F) \neq 0$ and $P(\overline{F}) \neq 0$. Thus, we can divide by both $P(F)$ and $P(\overline{F})$. As a result, each of the following statements is equivalent.

$$P(E|F) = P(E|\overline{F})$$
$$\frac{P(E \cap F)}{P(F)} = \frac{P(E \cap \overline{F})}{P(\overline{F})}$$
$$\frac{P(E \cap F)}{P(F)} = \frac{P(E \cap \overline{F})}{1 - P(F)}$$
$$P(E \cap F) \cdot (1 - P(F)) = P(E \cap \overline{F}) \cdot P(F)$$
$$P(E \cap F) - P(E \cap F) \cdot P(F) = P(E \cap \overline{F}) \cdot P(F)$$
$$P(E \cap F) = P(E \cap F) \cdot P(F) + P(E \cap \overline{F}) \cdot P(F)$$
$$P(E \cap F) = (P(E \cap F) + P(E \cap \overline{F})) \cdot P(F)$$
$$P(E \cap F) = P(E) \cdot P(F)$$

This is exactly what it means for E and F to be independent.

The last step is a consequence of the Law of Total Probability. Since F and $\overline{F}$ partition the sample space, $P(E \cap F) + P(E \cap \overline{F}) = P(E)$.

An even cleaner solution uses the Law of Total Probability in the first step:

$$P(E) = P(E \cap F) + P(E \cap \overline{F})$$
$$= P(E|F) \cdot P(F) + P(E|\overline{F}) \cdot P(\overline{F})$$
$$= P(E|F) \cdot P(F) + P(E|F) \cdot P(\overline{F})$$
$$= P(E|F)\left(P(F) + P(\overline{F})\right)$$
$$= P(E|F) \cdot 1$$
$$= P(E|F)$$

Since $P(E|F) = P(E)$, E and F are independent.

b) 😊😊😊😊 If $P(E|F) = P(\overline{E}|F)$, must it be true that E and F are independent? Provide a proof of independence, or give a counterexample by specifying a sample space S and two dependent events E and F that satisfy the given conditions.

> **Solution:** No, it need not be the case that E and F are independent.
>
> For example, consider the sample space $S = \{1, 2, 3, 4, 5, 6\}$ where each outcome is equally likely (the results of a die roll, for example).
>
> Let $E = \{1\}$ and $F = \{1, 2\}$. Then $P(E|F) = \frac{1}{2}$. Then $\overline{E} = \{2, 3, 4, 5, 6\}$, so $P(\overline{E}|F) = \frac{1}{2}$. Therefore, $P(E|F) = P(\overline{E}|F)$ as required.
>
> However this E and F are dependent because $P(E) = \frac{1}{6}$ and $P(E|F) = \frac{1}{2}$ so $P(E) \neq P(E|F)$.

Problem 6. Baby Avi

When Avi was a baby, he was a picky eater. When he was offered a food, he'd eat it only sometimes, depending on the type of food. Baby Avi ate

- bananas 95% of the time,
- crackers 60% of the time,
- meat 45% of the time, and
- zucchini 30% of the time.

a) 😊😊😊😊 Baby Avi's grandpa gave him one of the above four foods and Baby Avi ate it all up. You have no idea which of the four foods it was, so assume it was equally likely to be any of them.

Given that Baby Avi ate the food, what's the probability that it was a banana? A cracker? Meat? Zucchini? Show your work.

> **Solution:** Throughout this question, we'll define the following events:
>
> - E: he eats the food.
> - B: the food is banana.
> - C: the food is a cracker.
> - M: the food is meat.
> - Z: the food is zucchini.
>
> We're given the following information:

- The food is equally likely to be any of them, so $P(B) = P(C) = P(M) = P(Z) = 0.25$.
- $P(E|B) = 0.95$, $P(E|C) = 0.6$, $P(E|M) = 0.45$, and $P(E|Z) = 0.3$.

We need to find four probabilities, $P(B|E)$, $P(C|E)$, $P(M|E)$, and $P(Z|E)$. To do so, we will use Bayes rule in each case, which states that, for any food F,

$$P(F|E) = \frac{P(F)P(E|F)}{P(E)}$$

Note that we will use this same formula for all four probabilities, and all four probabilities have $P(E)$ as part of their calculation. Furthermore, since the events B, C, M, and Z form a partition of our sample space (exactly one of these events must be true), we can use the law of total probability to re-write $P(F)$ in terms of the information we're given:

$$P(E) = P(E \cap B) + P(E \cap C) + P(E \cap M) + P(E \cap Z)$$
$$P(E) = P(B)P(E|B) + P(C)P(E|C) + P(M)P(E|M) + P(Z)P(E|Z)$$
$$= 0.25 \cdot 0.95 + 0.25 \cdot 0.6 + 0.25 \cdot 0.45 + 0.25 \cdot 0.3$$
$$= 0.575$$

Now, we can compute the required probabilities as follows:

$$P(B|E) = \frac{P(B)P(E|B)}{P(E)} = \frac{0.25 \cdot 0.95}{0.575} \approx 0.413$$
$$P(C|E) = \frac{P(C)P(E|C)}{P(E)} = \frac{0.25 \cdot 0.6}{0.575} \approx 0.261$$
$$P(M|E) = \frac{P(M)P(E|M)}{P(E)} = \frac{0.25 \cdot 0.45}{0.575} \approx 0.196$$
$$P(Z|E) = \frac{P(Z)P(E|Z)}{P(E)} = \frac{0.25 \cdot 0.3}{0.575} \approx 0.130$$

As expected, the food is most likely to be banana, since that's the food that Baby Avi likes best.

b) As before, Baby Avi's grandpa gave him one of the above four foods and Baby Avi ate it all up. This time, suppose you know that Baby Avi's grandpa offers

- bananas 15% of the time,
- crackers 40% of the time,
- meat 10% of the time, and
- zucchini 35% of the time.

Given that Baby Avi ate the food, what's the probability that it was a banana? A cracker? Meat? Zucchini? Show your work.

Solution:

We can use the same framework as in the last part, it's just that $P(B)$, $P(C)$, $P(M)$, and $P(Z)$

have all changed. Specifically, $P(B) = 0.15$, $P(C) = 0.4$, $P(M) = 0.1$, and $P(Z) = 0.35$.

$$P(E) = P(E \cap B) + P(E \cap C) + P(E \cap M) + P(E \cap Z)$$
$$P(E) = P(B)P(E|B) + P(C)P(E|C) + P(M)P(E|M) + P(Z)P(E|Z)$$
$$= 0.15 \cdot 0.95 + 0.4 \cdot 0.6 + 0.1 \cdot 0.45 + 0.35 \cdot 0.3$$
$$= 0.5325$$

Then:

$$P(B|E) = \frac{P(B)P(E|B)}{P(E)} = \frac{0.15 \cdot 0.95}{0.5325} \approx 0.268$$

$$P(C|E) = \frac{P(C)P(E|C)}{P(E)} = \frac{0.4 \cdot 0.6}{0.5325} \approx 0.451$$

$$P(M|E) = \frac{P(M)P(E|M)}{P(E)} = \frac{0.1 \cdot 0.45}{0.5325} \approx 0.085$$

$$P(Z|E) = \frac{P(Z)P(E|Z)}{P(E)} = \frac{0.35 \cdot 0.3}{0.5325} \approx 0.197$$

Now it's most likely that the food was a cracker, which makes sense because crackers are the most given food, and they are also a well-liked food.

c) 😊😊😊 Compare your answers to part (a) and part (b) above. Identify which of the four probabilities you computed increased and which decreased, and explain why this makes sense intuitively.

Solution: The probability that the food was a banana or meat has decreased, which makes sense because Baby Avi's grandpa was less likely to feed him those foods.

The probability that the food was a cracker or zucchini has increased, which makes sense since we know Baby Avi's grandpa is especially fond of those foods.

Notice how our prior knowledge about the chance of each food being fed changes our results. Having this knowledge leads to different results than in part (a), when we did not have this knowledge.

Problem 7. Optional Extra Credit Challenge: Classifying Neutrino Signals in Majorana Demonstrator!

We are hosting a optional class-wide competition to see who can make the best predictions! Top 10 predictions can earn extra credit on Midterm 2! Please note that no point(😊) is assigned to this problem.

High-Purity Germanium (HPGe) detector is one of the most sensitive detectors human beings have ever manufactured. It is sensitive in a sense that it measures the energy of elementary particles (electron, photon, etc) very accurately. Because of this, HPGe detectors have a wide range of applications, including the search for neutrinos and dark matters, medical imaging, as well as nuclear non-proliferation.

When a particle comes into HPGe detector, it produces a waveform, or time series data, as shown in the picture below. A time series is a sequence of data points that occur in successive order over some period of time. More formally, we can define time series this way: for each data point, a time series contains n pairs of t_i, a_i where t_i is the ith time sample and a_i is the value at i^{th} time sample. To simplify this problem,

we will not ask you to directly analyze the time series, but we extract certain features from the HPGe time series for you to build model.

The MAJORANA DEMONSTRATOR Experiment uses HPGe detector to search for neutrinoless double-beta decay. This kind of decay happens at least a trillion times slower than the age of our universe. Finding it will unravel the fundamental understanding of particle physics. For more information please read this Press Release.

In the supplementary Jupyter notebook (linked), you are given access to a CSV containing training data with information about 3000 elementary particles which deposits their energy in a HPGe detector. Some of them are signal-like, i.e. they exhibit the same shape with neutrinoless double-beta decay, others are noise-like, i.e. they looks different from neutrinoless double-beta decay. This is a labeled dataset where signal-like data has a label of 1 and background-like data has a label of 0. We read this in as a DataFrame where the columns are different parameters. The columns are:

- **tDrift50:** Period from the Start of Rise (t_{SR}) to when the waveform reaches 50% of Max_Amp (t_{50}), can also be written as $t_{50} - t_{SR}$

- **tDrift90:** Period from the Start of Rise (t_{SR}) to when the waveform reaches 90% of Max_Amp (t_{90}), can also be written as $t_{90} - t_{SR}$

- **tDrift100:** Period from the Start of Rise (t_{SR}) to when the waveform reaches Max_Amp, can also be written as $t_{Max_Amp} - t_{SR}$

- **blnoise:** The standard deviation of amplitude values a_i in the green-colored region.

- **tslope:** The slope of the waveform tail.

- **Energy:** The energy of each waveform, i.e. the target of the previous challenge

- **Current_Amplitude:** A new parameter extracted from the waveform, by taking a derivative of the waveform and read out the maximum of the derivative.
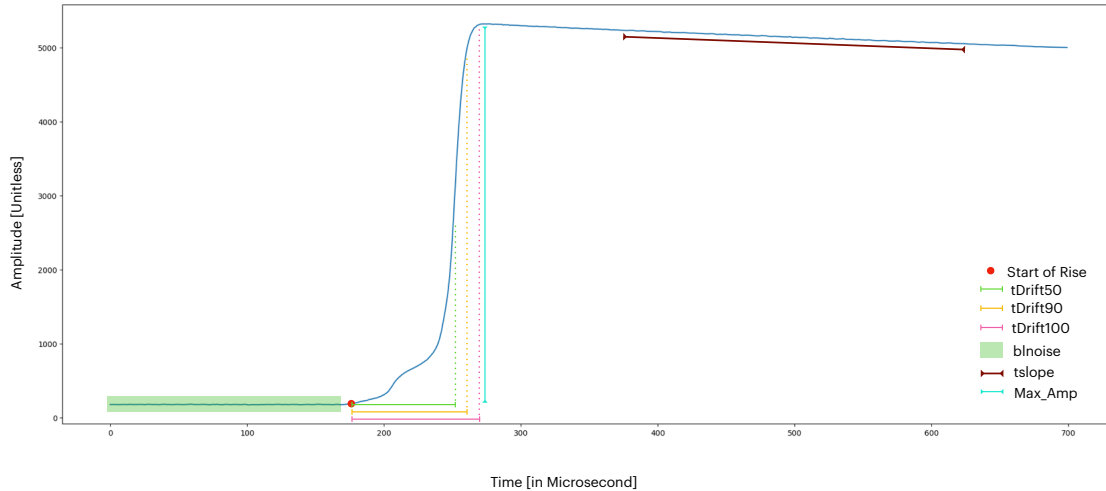
Your task for this problem is to produce a "classification score" for each HPGe detector waveforms. This score does not have to be between 0 and 1, but signal data point (data point with a label of 1) should have a higher score than noise data point (data point with a label of 0). This extra credit opportunity as minimal constraint: you are not allowed to import any additional python package, but you are allowed to use any

model (including but not limited to the Naive Bayes classifier we will discuss in class) and any amount of parameters.

We've provided you with a function `calculate_AUC` to calculate the area under curve (AUC) of the Receiver Operating Characteristic (ROC) curve of your predictions on each waveform in the training data. For more information please read this link. Your job is to fill in the body of the `predict` function. This function should take as input one row of the DataFrame (corresponding to one particular waveform) and return the predicted classification score corresponding to this waveform. How you make this prediction is up to you, subject to the rules above. Feel free to add more cells and functions, and to change the provided `predict` function, but do not change the provided `calculate_AUC` function.

The ten best prediction functions (as determined by the AUC on the hidden test data) will earn some extra credit on the upcoming midterm exam according to the following scheme: for $n \leq 10$, the nth ranked prediction function in the class earns $11 - n$ percentage points as extra credit on Midterm 1. But the total extra creit from both challenges (HW4 Q6 plus this one) is capped at 10%.