Problem 1

Intrusion Detection System

A computer intrusion detection system (IDS) is designed to provide an alarm whenever an intrusion (e.g., unauthorized access) into a computer system is being attempted. A probabilistic evaluation of a system with two independent operating intrusion detection systems (a double IDS) was published in the Journal of Research of the National Institute of Standards and Technology (Nov/Dec 2003).

Consider a double IDS with System A and System B. If there is an intruder, System A sounds an alarm with probability 0.9, and System B sounds an alarm with probability 0.95. If there is no intruder, System A sounds an alarm with probability 0.2, and System B sounds an alarm with probability 0.1. Assume that Systems A and B operate independently.

Let A be the event that System A sounds an alarm. Let B be the event that System B sounds an alarm. Let I be the event that there is in fact an intruder.

(a) Define the relevant events, and formally express the four probabilities given in the example. (e.g., $P(A|I) = x.xx$)

The relevant events are:

- System A sounds the alarm when there is an intruder: $P(A|I)$.
- System A sounds the alarm when there is no intruder: $P(A|I^c)$.
- System B sounds the alarm when there is an intruder: $P(B|I)$.
- System B sounds the alarm when there is no intruder: $P(B|I^c)$.

The probabilites of each of these events is:

- $P(A|I) = 0.9$
- $P(A|I^c) = 0.95$
- $P(B|I) = 0.2$
- $P(B|I^c) = 0.1$

(b) If there is an intruder, what is the probability that both systems sound an alarm?

Because A and B are independent events the probability of A and B occurring is given by

$$P(A \cap B) = P(A) \times P(B). \tag{1}$$

Plugging in the events and probabilities listed above for when there is an intruder gives

$$P(A \cap B|I) = P(A|I) \times P(B|I) = 0.9 \times 0.95 = 0.855.$$

(c) If there is no intruder, what is the probability that both systems sound an alarm?

Again using Eq. (1), but this time using the events and probabilities for when there is not an intruder (I^c). In this case the probability of both systems sounding an alarm is

$$P(A \cap B|I^c) = P(A|I^c) \times P(B|I^c) = 0.2 \times 0.1 = 0.02.$$

(d) Given an intruder, what is the probability that at least one of the systems sound an alarm?

Here we want to the know the probability of the union of A and B. To do that we use the additive rule, which is

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{2}$$

Plugging in the given probabilities for when an intruder is present and the calculated probability from part (b) gives

$$P(A \cup B|I) = P(A|I) + P(B|I) - P(A \cap B|I) = 0.9 + 0.95 - 0.855 = 0.995.$$

(e) Assume that the probability of an intruder is 0.4. Also continue to assume that both systems operate independently. If both systems sound an alarm, what is the probability that an intruder is detected?

Since I is the event that there is an intruder that means that $P(I) = 0.4$, and the probability there isn't an intruder is $P(I^c) = 0.6$. Written mathematically the probability that an intruder is detected given that both systems sound an alarm is $P(I|A \cap B)$. To solve this we will need Bayes's rule, which is

$$P(Z_i|Y) = \frac{P(Z_i)P(Y|Z_i)}{\sum_{j=1}^{k} P(Z_j)P(Y|Z_j)} \tag{3}$$

for a sample space S, where Y is an observed event and the events Z_j form a partition of S. For simplification, let the event C be the event that both systems sound the alarm, $C = A \cap B$. plugging in I and C into Eq. (3) for Z and Y respectively, and expanding the summation gives

$$P(I|C) = \frac{P(I)P(C|I)}{P(I)P(C|I) + P(I^c)P(C|I^c)}.$$

From parts (b) and (c) we know that $P(C|I) = P(A \cap B|I) = 0.855$ and $P(C|I^c) = P(A \cap B|I^c) = 0.02$. The probability that there is an intruder if both systems sound the alarm is

$$P(I|A \cap B) = P(I|C) = \frac{0.4 \times 0.855}{(0.4 \times 0.855) + (0.6 \times 0.02)} = 0.966.$$

Problem 2

Lie Detector Test

A new type of lie detector, called the Computerized Voice Stress Analyzer (CVSA) has been developed. The manufacturer claims that the CVSA is 98% accurate — that is, it correctly determines if a suspect is lying or not 98% of the time — and unlike a polygraph machine, it will not be thrown off by drugs and medical factors. However, laboratory studies by the DoD found that the CVSA had an accuracy rate of 49.8% — slightly less than pure chance. Suppose the CVSA is used to test the veracity of four suspects. Assume the suspects' responses are independent.

(a) If the manufacturer's claim is true, what is the probability that the CVSA will correctly determine the veracity of all four suspects?

There are two possible outcomes for each individual event. Let C be when the CVSA is correct and W be when the CVSA is incorrect. For each suspect the probability of the CVSA producing a correct result is $P(C) = 0.98$. Since the suspects' responses are independent, the probability of the CVSA being correct all four times is the individual probabilities multiplied together. In this case it is

$$P(\{C, C, C, C\}) = (P(C))^4 = (0.98)^4 = 0.92$$

(b) If the manufacturer's claim is true, what is the probability that the CVSA will yield an incorrect result for at least one of the four suspects?

The only set that does not result in the CVSA yielding an incorrect result for at least one suspect is the one given above, where it is correct every time. Since there is a probability of 0.92 for it being correct all four times, the probability that it has at least one incorrect result is 0.08.

(c) Suppose that in a laboratory experiment conducted by the DoD on four suspects, the CVSA yielded incorrect results for two of the suspects. Use this result to make an inference about the true accuracy rate of the new lie detector.

Lets consider the probability of getting two incorrect results with the manufactures accuracy claim. There are six combinations of the CVSA getting 2 correct and 2 incorrect results ($CCWW$, $CWCW$, $CWWC$, $WCCW$, $WCWC$, $WWCC$). The probability of each one of these individually is

$$P(C)^2 * P(W)^2 = (0.98)^2 * (0.02)^2 = 0.00038$$

Because there are 6 different combinations the total probability of getting two incorrect results out of four suspects is $0.00038 \times 6 = 0.0023$.

Now lets consider the same probability if the CVSA accuracy is only 49.8% ($P(C) = 0.498$). In this case the probability of getting two incorrect results is

$$((P(C)^2 * P(W)^2)) \times 6 = ((0.498)^2 * (0.502)^2) \times 6 = 0.37$$

The result of two incorrect results out of four suspects with the DoD's claim of accuracy is around 160 times more likely than with the manufacturer's claim of accuracy. Given this small sample size I think the true accuracy is closer to the DoD's claim (and its not just because I work for them).

Problem 3

Auditing an Accounting System

In auditing a firm's financial statements, an auditor will (1) assess the capability of the firm's accounting system to accumulate, measure, and synthesize transactional data properly and (2) assess the operational effectiveness of the accounting system. In performing the second assessment, the auditor frequently relies on a random sample of actual transactions (Stickney and Weil, Financial Accounting: An Introduction to Concepts, Methods, and Uses, 2002). A particular firm has 5,382 customer accounts that are numbers from 0001 to 5382.

(a) One account is to be selected at random for audit. What is the probability that account number 3,241 is selected?

Since there is an equal chance of selecting any of the account numbers, the probability that account number 3,241 is selected is $\frac{1}{5{,}382}$.

(b) Draw a random sample of 10 accounts, and explain in detail the procedure you used. (Hint: Python can do this)

Here is a random sampling of the accounts: (1362, 5119, 5145, 4367, 4846, 686, 4581, 3129, 779, 3873). To get this I first created a list all the accounts using the built in range function. Then I used NumPy's random.choice function, and set the size to 10, to get the 10 random account numbers. I also made sure there was no replacement to avoid having the same number selected more than once.

(c) Referring to part b, is one sample of size 10 more likely to be chosen than any other? What is the probability that the sample you drew in part b was selected?

No sample of 10 account numbers is more likely to be chosen then any other because there is an equal probability for every account number each time. The order of the account numbers in a sample does not matter so the total number of samples of 10 that can be selected is given by

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}, \tag{4}$$

where $n = 5,382$ and $r = 10$. The result is 5.57×10^{30} combinations. Since they are all equally likey, the probability that the sample from (b) was selected was

$$\frac{1}{5.57 \times 10^{30}} = 1.79 \times 10^{-31}.$$

Problem 4

Fish Contamination

A U.S. Army Corps of Engineers (USACE) study focused on DDT contamination of fish in the Tennessee River in Alabama. Part of that investigation studied how far upstream contaminated fish have migrated. A fish is considered to be contaminated if its measured DDT concentration is greater than 5.0 parts per million (ppm).

(a) Considering only contaminated fish captured from the Tennessee River, the data reveal that 52% of the fish are found 275–300 miles upstream, 39% are found 305-325 miles upstream, and 9% are found 330–350 miles upstream. Use the percentages to estimate the probabilities $P(275\text{–}300)$, $P(305\text{–}325)$, and $P(330\text{–}350)$.

From the given percentages of contaminated fish at the various distances the probabilities are:

$$P(275-300) = 0.52$$
$$P(305-325) = 0.39$$
$$P(330-350) = 0.09$$

(b) Given that a contaminated fish is found a certain distance upstream, the probability that it is a channel catfish (CC) is determined from the data as $P(CC|275\text{–}300) = 0.775$, $P(CC|305\text{–}325) = 0.77$, and $P(CC|330\text{–}350) = 0.86$. If a contaminated channel catfish is captured from the Tennessee River, what is the probability that it was captured 275–300 miles upstream?

In order to find $P(275-300|CC)$ we need to use Bayes's Rule, which is given in Eq. (3). For this problem the distances upstream form the partition of our sample space. So we need the probabilities of a contaminated fish at each distance and the probabilities that a fish caught at each distance is a channel catfish. Writing this out mathematically is

$$P(275-300|CC) = \frac{P(275-300)P(CC|275-300)}{P(275-300)P(CC|275-300)+P(305-325)P(CC|305-325)+P(330-350)P(CC|330-350)}$$

Plugging in the respective probabilities, the probability that the contaminated channel catfish was captured 275-300 miles upstream is

$$P(275-300|CC) = \frac{(0.52)(0.775)}{(0.52)(0.775)+(0.39)(0.77)+(0.09)(0.86)} = 0.516$$