Question 1:
A website allows the user to create an 8-character password that consists of lower case letters (a-z) and digits (0-9), allowing repetition.

a) How many valid passwords are there where all characters are letters?

b) How many valid passwords are there where all characters are letters, and are distinct (no repetition)?

c) How many valid passwords are there where all characters are distinct, and alternate between letters and digits? Examples: 1e2t3c4u or a4b6c3d7.

d) How many valid passwords are there where all characters are distinct letters in alphabetical order? For example, abhikmno is allowed, but not bafgkmn.

e) How many valid passwords are there which contain only the letters a and b, and contain each of these letters at least once?

f) How many valid passwords are there which contain only the letters a and b, and contain an equal number of each letter?

g) Suppose a password is randomly generated using only letters from a-e and numbers from 0-4 (inclusive). What is the probability that the password contains at least one letter and at least one digit?

h) A hacker writes a program that can test 1000 different passwords per second. John is bad at making passwords: all of his passwords contain the sequence john, and the rest of the characters are digits. If the hacker knows this, how much time (in seconds) would it take to test all possible passwords and (with certainty) hack John’s account?
Question 2:

Consider a social network that allows accounts to be secured with a 6-digit passcode (any sequence of exactly six digits between 0-9 is valid). Assume the network has $m$ users including you, and that all users choose one of the valid 6-digit passcodes uniformly at random. A user’s passcode is considered safe if no other user has the same passcode.

a) As a function of $m$, what is the probability that your own passcode is safe?

b) How many users must there be for there to be a 50% or greater chance that your own passcode is not safe?

c) As a function of $m$, what is the probability that all users have a safe passcode?

d) How many users must there be for there to be a 50% or greater chance that at least one user’s passcode is not safe?

Question 3:

In this problem, we will say that event $A$ “attracts” event $B$ if $P(B \mid A) > P(B)$, and event $A$ “repels” event $B$ if $P(B \mid A) < P(B)$. For simplicity, you may assume that the probability of all events is greater than zero.

a) Show that if $A$ attracts $B$, then $B$ must also attract $A$.

b) Show that if $A$ neither attracts nor repels $B$, $A$ and $B$ must be independent.

c) Show that if $A$ attracts $B$, then $A$ repels $B^c$.

d) Show that if $A$ attracts $B$, then $P(B \mid A) > P(B \mid A^c)$.

Question 4:

Classifiers are of great use in diagnosing medical conditions based on symptoms. In this problem, you will consider a data set containing information about a set of patients (not a completely random sample of the population) who may or may not have heart disease.

The data is split into two matrices, $\text{trainPatient}$ and $\text{testPatient}$. Each row of these three-column matrices represents a patient. The first column contains the age of the patient in years. The second indicates whether exercise causes them to experience chest pain, with “1” indicating yes and “0” no. The third column is the ground truth of whether they have heart disease, again with “1” indicating yes and “0” no. $\text{trainPatient}$ will be used to “train” your classifier by estimating various probabilities. $\text{testPatient}$ will be used to evaluate classifier performance, but not to estimate probabilities.

To define a probabilistic model of this data, we let $Y_i = D$ if patient $i$ has heart disease, and $Y_i = H$ if patient $i$ does not. To construct a simple Bayesian classifier, we will compute the posterior probability $P(Y_i \mid X_i)$ of the class label given some feature $X_i$. If $P(Y_i = D \mid X_i) > P(Y_i = H \mid X_i)$, we classify patient $i$ as probably having heart disease. Otherwise, we classify them as probably not having heart disease.

By Bayes’ rule, the posterior probability $P(Y_i \mid X_i) = \frac{P(X_i \mid Y_i)P(Y_i)}{P(X_i)}$. Consider the feature $X_i = A_i$, where $A_i = 1$ if patient $i$ has age > 55, and $A_i = 0$ if patient $i$ has age ≤ 55.
a) A simple way to estimate probabilities is by counting how many times each event occurs in the training data. Let $N$ be the total number of patients, $N_D$ the number of patients with heart disease, $N_{DA}$ the number of patients with heart disease over age 55, $N_H$ the number of patients without heart disease, and $N_{HA}$ the number of patients without heart disease over age 55. We then set

$$P(Y_i = D) = \frac{N_D}{N}, \quad P(Y_i = H) = \frac{N_H}{N},$$

$$P(X_i = 1 \mid Y_i = D) = \frac{N_{DA}}{N_D}, \quad P(X_i = 0 \mid Y_i = D) = 1 - P(X_i = 1 \mid Y_i = D) = \frac{N_D - N_{DA}}{N_D},$$

$$P(X_i = 1 \mid Y_i = H) = \frac{N_{HA}}{N_H}, \quad P(X_i = 0 \mid Y_i = H) = 1 - P(X_i = 1 \mid Y_i = H) = \frac{N_H - N_{HA}}{N_H}.$$

Estimate these probabilities from the data in trainPatient, and report their values.

b) Test the Bayesian classifier based on feature $A_i$ using the data in testPatient. What is the accuracy (percentage of correctly classified patients) of this classifier on that data?

c) Consider now the feature $X_i = E_i$, where $E_i = 1$ if exercise causes patient $i$ to experience chest pain, and $E_i = 0$ if it does not. Estimate and report conditional probabilities for this new feature as in part (a). What is the accuracy of a Bayesian classifier based on feature $E_i$ on the data in testPatient?

d) Consider the pair of features $X_i = \{A_i, E_i\}$. This pair of features can take on 4 values: \{0, 0\}, \{0, 1\}, \{1, 0\}, or \{1, 1\} (we do not assume that $A_i$ and $E_i$ are independent). By counting as in part (a), compute and report the probabilities of these four events given $Y_i = D$ and given $Y_i = H$. What is the accuracy of a Bayesian classifier based on feature $X_i$ on the data in testPatient? Compare the accuracy of this classifier to those from parts (b) and (c), and give an intuitive explanation for what you observe.