Homework 0: Combinatorics & Basic Probability

UC Irvine CS177: Applications of Probability in Computer Science

REVIEW: NOT COLLECTED OR GRADED

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* bafgkmno*.*

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:

To verify the identity of users who lose their password, a website asks 10 true-false security questions. Let Q equal the number of questions that are answered correctly.

a) A naive hacker attempts to break in by randomly answering the 10 security questions. What is the probability distribution for the number of correctly answered questions Q?

b) Write a Python function that numerically evaluates the probabilities in part (a), and plot these probabilities as a function of Q.

c) A more sophisticated hacker does research on the user whose account he seeks to break into, and gives the correct answer to each security question with probability 0.9. Give an equation for the distribution of the number of correctly answered questions Q, and write a Python function that numerically evaluates and plots these probabilities.

Question 3:

Let X be a continuous random variable representing the time, in hours, that it takes your laptop to backup its data to a server. Suppose that $P(X < 0) = 0$ and $P(X > 4) = 0$, because after 4 hours, the backup times out and fails. For $0 \leq x \leq 4$, X has the probability density function $f_X(x) = cx^{1/2}$.

a) What value of the constant $c > 0$ makes $f_X(x)$ a valid probability density function?

b) What is the cumulative distribution function $F_X(x)$? What is the median of X?

c) What is the expected backup time $E[X]$?

d) What is the standard deviation of X?

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, `trainPatient` and `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. `trainPatient` will be used to "train" your classifier by estimating various probabilities. `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 $P(Y_i \mid X_i) = \frac{P(X_i|Y_i)P(Y_i)}{P(X_i)}$. 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) *We estimate probabilities by counting event frequencies 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 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 classifer 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.*