Homework 6

Due November 7 at 11 pm

Unless stated otherwise, justify any answers you give. You can work in groups, but each student must write their own solution based on their own understanding of the problem.

When uploading your homework to Gradescope you will have to select the relevant pages for each question. Please submit each problem on a separate page (i.e., 1a and 1b can be on the same page but 1 and 2 must be on different pages). We understand that this may be cumbersome but this is the best way for the grading team to grade your homework assignments and provide feedback in a timely manner. Failure to adhere to these guidelines may result in a loss of points. Note that it may take some time to select the pages for your submission. Please plan accordingly. We suggest uploading your assignment at least 30 minutes before the deadline so you will have ample time to select the correct pages for your submission. If you are using LaTeX, consider using the minted or listings packages for typesetting code.

1. (Bayesian coin flip) Let us try out another prior for the Bayesian coin flip problem in the notes. We now model the parameter of the Bernouilli as being uniform between 1/2 and 1.

   (a) Briefly justify the model and compute the probability that the result of the coin flip is heads or tails under this model.

   (b) After the coin flip we update the distribution of the bias of the coin (i.e. the parameter of the Bernouilli that represents the coin flip) by conditioning it on the outcome. Compute the distribution if the outcome is tails and if the outcome is heads. Sketch any distributions you compute and explain why the drawing makes sense.

   (c) You observe 100 coin flips and they all turn out to be tails (i.e. 0). Do you think you should reconsider your prior? If so, why?

2. (Halloween parade) The city of New York hires you to estimate whether it will rain during the Halloween parade. Checking past data you determine that the chance of rain is 20%. You model this using a random variable $\tilde{r}$ with pmf

$$p_{\tilde{r}}(1) = 0.2, \qquad p_{\tilde{r}}(0) = 0.8,$$

   where $\tilde{r} = 1$ means that it rains and $\tilde{r} = 0$ that it doesn't. Your first idea is to be lazy and just use the forecast of a certain website. Analyzing data from previous forecasts, you model this with a random variable $\tilde{w}$ that satisfies

$$\mathrm{P}(\tilde{w} = 1|\tilde{r} = 1) = 0.8, \qquad \mathrm{P}(\tilde{w} = 0|\tilde{r} = 0) = 0.75.$$

   (a) What is the probability that the website is wrong?

   Unsatisfied with the accuracy of the website, you look at the data used for the forecast (they are available online). Surprisingly the relative humidity of the air is not used, so you decide to incorporate it in your prediction in the form of a random variable $\tilde{h}$.

(b) Is it more reasonable to assume that $\tilde{h}$ and $\tilde{w}$ are independent, or that they are conditionally independent given $\tilde{r}$? Explain why.

You assume that $\tilde{h}$ and $\tilde{w}$ are conditionally independent given $\tilde{r}$. More research establishes that conditioned on $\tilde{r} = 1$, $\tilde{h}$ is uniformly distributed between 0.5 and 0.7, whereas conditioned on $\tilde{r} = 0$, $\tilde{h}$ is uniformly distributed between 0.1 and 0.6.

(c) Compute the conditional pmf of $\tilde{r}$ given $\tilde{w}$ and $\tilde{h}$. Use the distribution to determine whether you would predict rain for any possible value of $\tilde{w}$ and $\tilde{h}$.

(d) What is the probability that you make a mistake?

3. (Chad) You find your coworker Chad really annoying. He often works from home, but when he is in the office and you walk by his desk he insists on showing you pictures of his pet iguana. You would like to be able to predict when he is in the office in order to avoid him as much as possible. Another of Chad's annoying habits is to crank up the AC, so you decide to use the temperature in the office to predict his presence. After a month you have gathered the following data.

Chad	61	65	59	61	61	65	61	63	63	59
No Chad	68	70	68	64	64	-	-	-	-	-

Temperature (^{o} F)

(a) You model the temperature using a random variable $\tilde{t}$. Use a kernel density estimate which is rectangular and has width 2 to estimate the conditional pdf of $\tilde{t}$ given the presence or absence of Chad. Sketch the distribution.

(b) We model the presence or not of Chad as a parameter c ($c = 1$ means he is present, $c = 0$ that he is absent). If the temperature is 68°, does a maximum-likelihood estimate of c predict that Chad is in the office?

(c) Now we take a Bayesian approach and model the presence or absence of Chad using a random variable $\tilde{c}$ which is equal to 1 if he is there and 0 if he is not. Estimate the pmf of $\tilde{c}$ from the data.

(d) If the temperature is 64°, use the posterior distribution of $\tilde{c}$ to predict whether Chad is in the office.

(e) What problem do we run into if the temperature is 57°? Explain how using parametric estimation may alleviate this problem.

4. (Heart-disease detection) A hospital is interested in developing a system for automatic heart-disease detection[1]. Your task is to use the data in the *heart_disease_data.npz*[2] and

[1] A patient is deemed to suffer from heart disease if at least one of his or her major vessels is 50% narrower than it should be.

[2] The data in this problem, which was compiled from five hospitals in Hungary, Switzerland and the United States, is available at http://archive.ics.uci.edu/ml/datasets/Heart+Disease.

complete the script[3] to detect heart disease in patients. You model heart disease as a random variable $\tilde{h}$ that indicates whether the patient suffers from heart disease or not:

$$\tilde{h} = \begin{cases} 0 & \text{if patient does not suffer from heart disease,} \\ 1 & \text{if patient suffers from heart disease.} \end{cases}$$

The available data contain the patient's sex, the type of chest pain experienced by the patient and the cholesterol of the patient. We model these quantities as the random variables $\tilde{s}$, $\tilde{c}$ and $\tilde{x}$ respectively, where

$$\tilde{s} = \begin{cases} 0 & \text{if patient is female,} \\ 1 & \text{if patient is male,} \end{cases}$$

$$\tilde{c} = \begin{cases} 0 & \text{if the pain is typical angina,} \\ 1 & \text{if the pain is atypical angina,} \\ 2 & \text{for other types of chest pain,} \\ 3 & \text{if there is no chest pain,} \end{cases}$$

and $\tilde{x}$ is a continuous random variable.

(a) Derive the MAP estimate of $\tilde{h}$ given $\tilde{s}$ and $\tilde{c}$ as a function of the pmf of $\tilde{h}$ ($p_{\tilde{h}}$) and the conditional pmfs $p_{\tilde{s}|\tilde{h}}$ and $p_{\tilde{c}|\tilde{h}}$. The MAP estimate is defined as the mode of the posterior distribution. Assume that if we know whether a patient is suffering from heart disease, the sex of the patient and the type of chest pain experienced by the patient are conditionally independent.

(b) Complete the corresponding part of the script to estimate the necessary probability mass functions from the data. The training data consists of 218 patients and is provided in the arrays *data["heart_disease"]*, *data["sex"]* and *data["chest_pain"]*. Apply the MAP decision rule you derived in part (a) to predict whether a group of 50 other patients, whose information is stored in the vectors *data["sex_test"]* and *data["chest_pain_test"]*, suffer from heart disease. Calculate the error rate (i.e. the proportion of predictions that are incorrect) by comparing your results to *data["heart_disease_test"]*, which indicates whether the patients suffer from heart disease or not.

(c) Derive a MAP estimate of $\tilde{h}$ given $\tilde{s}$, $\tilde{c}$ and $\tilde{x}$ that only depends on the pmf of $\tilde{h}$ $p_{\tilde{h}}$, the conditional pmfs $p_{\tilde{s}|\tilde{h}}(s|h)$ and $p_{\tilde{c}|\tilde{h}}$ and the conditional pdf $f_{\tilde{x}|\tilde{h}}$, assuming that if we know whether a patient is suffering from heart disease, the sex, type of chest pain and cholesterol level of the patient are all independent.

(d) You decide to model the cholesterol level of a patient conditioned on whether he or she suffers from heart disease as a Gaussian random variable. For both cases, complete the corresponding part of the script to obtain the ML estimates of the conditional distributions from the data in *cholesterol* and compare the estimated pdf to the histogram of the data.

[3]The script is avaiable at https://github.com/cfgranda/prob_stats_for_data_science/blob/main/modeling_discrete_continous_data/heart_diseases_EXERCISE.ipynb

(e) Complete the corresponding part of the script to apply your MAP decision rule incorporating the cholesterol data and compute the new error rate (using the cholesterol rates of the 50 new patients, stored in *data["cholesterol_test"]*.

(f) We have made some conditional independence assumptions that do not necessarily hold. Another option would have been to estimate the joint distribution of all the random variables from the data. Is this a good idea?