Probability and Statistics For Data Science HW 6

gjd9961

November 2021

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.

   We use the parameter θ to represent the probability of a coin flip being heads. Because our uncle is a tricketer, and we assume that he's stacking the odds in his favor, we assume that the coin must be weighted in his favore. Therefore, it is reasonable to assume that the odds the coin flip being heads is greater than $\frac{1}{2}$, and since probabilities only go from 0 to 1, then the range of θ is $\frac{1}{2} \leq \theta \leq 1$. Now lets compute the probability of a heads, and the probability of a tails, assuming theta is uniformly distributed along its range.

   $$
   \begin{aligned}
   P(\tilde{x} = heads) &= \int_{.5}^{1} f_\theta(\theta) \times P(heads|\theta)d\theta \\
   P(\tilde{x} = heads) &= \int_{.5}^{1} 2\theta \\
   P(\tilde{x} = heads) &= 1 - .25 = .75 \\
   &\text{And} \\
   P(\tilde{x} = Tails) &= 1 - .75 = .25
   \end{aligned} \tag{1}
   $$

   We know that $f_\theta(\theta)$ is equal to the density at that point, which is equal to 2, since our θ is uniformly distributed along $.5 \leq \theta \leq 1$. Likewise, the probability of flipping a heads given a θ is just θ.

   (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.

   In this problem, we will update our beliefs about the fairness of the coin given some assumptions about the coin and data collected on flips. In this way, we condition on the prior events to calculate a posterior probability. We will need to calculate two posteriors, one for theta if we flip a heads, (which should assign higher probability to greater degrees of favor to landing on heads), and one for theta if we flip

a tails, (one which should assign higher probability to the coin being closer to fair). We do so as follows:

$$f(\theta|heads) = \frac{f_\theta(\theta) \times P(heads|\theta)}{P_{\tilde{x}}(heads)} = \frac{2\theta}{\frac{3}{4}} = \frac{8 \times \theta}{3} \text{ for } \frac{1}{2} \leq \theta \leq 1 \text{ otherwise } 0$$

And

$$f(\theta|tails) = \frac{f_\theta(\theta) \times P(tails|\theta)}{P_{\tilde{x}}(tails)} = \frac{2(1-\theta)}{\frac{1}{4}} = 8 - 8 \times \theta \text{ for } \frac{1}{2} \leq \theta \leq 1 \text{ otherwise } 0$$
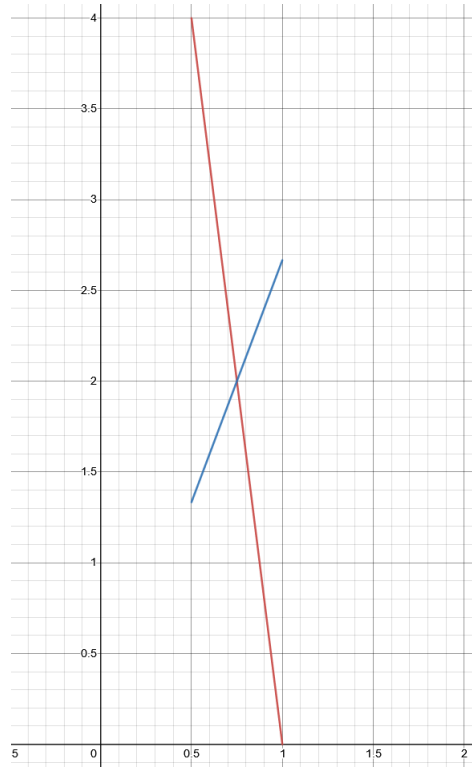
$$(2)$$

Figure 1: Posteriors of Theta Given Heads (Blue) or Tails (Red)
We can see that the flip verifies our assumptions about how the distribution of should change: the area under the red line is higher for $\theta \leq .75$ and the opposite is true for the blue line.

(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?

Yes we definitely should reconsider our priors. If we performed MLE given our data, by tweaking θ in the range we defined ($.5 \leq \theta \leq 1$), the most likely θ would be $\frac{1}{2}$ and the probability of the event occurring would be $(\frac{1}{2})^{100}$ which is incredibly unlikely. Our assumptions are too rigid, and because we defined our range for θ we will never have non-zero probability of the coin having parameter $\theta < .5$, which means we should definitely reconsider our assumptions.

2

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

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

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

   We can use the conditional probabilities and the law of total probability to calculate when the website is wrong. If we can calculate the event that the website is right using $P(A) \cap P(B) = P(B|A)P(A)$, we can then take the complement to find when the website is wrong. Here is the calculation for website is right, it has two parts, when it predicts rain and it rains, and when it predicts no rain and it doesn't rain:

$$P(\tilde{w} = 1 \cap \tilde{r} = 1) = P(\tilde{w} = 1|\tilde{r} = 1) \times P(\tilde{r} = 1) = .8 \times .2 = .16$$

$$P(\tilde{w} = 0 \cap \tilde{r} = 0) = P(\tilde{w} = 0|\tilde{r} = 0) \times P(\tilde{r} = 0) = .75 \times .8 = .6$$

$$P(Website = Correct) = P(\tilde{w} = 0 \cap \tilde{r} = 0) + P(\tilde{w} = 1 \cap \tilde{r} = 1) = .76$$

$$P(Website = Incorrect) = 1 - (P(\tilde{w} = 0 \cap \tilde{r} = 0) + P(\tilde{w} = 1 \cap \tilde{r} = 1)) = 1 - .76 = .24$$

   The probability that the website is incorrect is .24%.

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.

   Its more reasonable to assume that $\tilde{h}$ and $\tilde{w}$ are conditionally independent given $\tilde{r}$. The website has proved to be a strong predictor of rain, and if humidity is dependent on rain, then if the website gives us information about rain it consequently does about humidity as well. However, if we know that it is already reaining, there is no new information that can be obtained by reading the forecast on the website. So, $\tilde{h}$ and $\tilde{w}$ are not independent, and $\tilde{h}$ and $\tilde{w}$ are conditionally independent given $\tilde{r}$.

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.

3

(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 $\tilde{h}$.

Here we can use the chain rule for discrete and continuous variables, and we can take advantage of the fact that h is 0 between .1 and .5, and if h is between .6 and .7 h is 1. So really, all we need to figure out is what occurs in the range .5 to .6 for $\tilde{w} = 1$ and $\tilde{w} = 0$. We only need to calculate for $\tilde{x} = 1$. We can start with this formula:

$$p_{\tilde{r}|\tilde{w},\tilde{h}}(r|w,h) = \frac{f(h|w,r) \times p(w|r) \times p(r)}{p(w)f(h|w)}$$

$$p_{\tilde{r}|\tilde{w},\tilde{h}}(r|w,h) = \frac{f(h|w,r) \times p(w|r) \times p(r)}{\sum_{r=0}^{1} f(h|r)p(w|r)p(r)}$$

(3)

And now extend it to our situations:

$$p_{\tilde{r}|\tilde{w},\tilde{h}}(1|1,h) = \frac{5 \times .8 \times .2}{(2 \times .25 \times .8) + (5 \times .8 \times .2)} = \frac{2}{3}$$

$$p_{\tilde{r}|\tilde{w},\tilde{h}}(1|0,h) = \frac{5 \times .2 \times .2}{(2 \times .75 \times .8) + (5 \times .2 \times .2)} = .143$$

Now we can assemble what we know into a conditional pmf:

$$p_{\tilde{r}|\tilde{w},\tilde{h}}(1|w,h)) = \begin{cases} 0 & \text{if } .1 \le h \le .5 \\ \frac{2}{3} & \text{if } .5 \le h \le .6 \text{ and } w = 1 \\ .143 & \text{if } .5 \le h \le .6 \text{ and } w = 0 \\ 1 & \text{if } .6 \le h \le .7 \end{cases}$$

(4)

What we now have is a pmf which we can use as a classifier. If the probability of rain given the humidity and forecast is greater than $1/2$, then we predict it will rain. In this case, we will predict it will rain if $.5 \le h \le .6$ and $w = 1$ and we will also predict rain if $.6 \le h \le .7$. Otherwise, we predict no rain (if $.1 \le h \le .5$ or $.5 \le h \le .6$ and $w = 0$)

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

Because we are given that it cannot rain if the humidity if is in between .1 and .5, and it certainly will be raining if the humidity is between .6 and .7, we only have two scenarios to consider when talking about prediction error. The first scenario is when we predict rain ($.5 \le h \le .6$ and $w = 1$) but it actually doesn't rain, and the second is when we predict no rain, when ($.5 \le h \le .6$ and $w = 0$), but it actually rains. The combined probability of these two events will be our chance of error.

$$P(\text{Bad Prediction}) = P(r = 0, w = 1, .5 \le h \le .6) + P(r = 1, w = 0, .5 \le h \le .6)$$

$$P(\text{Bad Prediction}) = \int_{.5}^{.6} 2 \times .8 \times .25 \, dh + \int_{.5}^{.6} 5 \times .2 \times .2 \, dh$$

$$P(\text{Bad Prediction}) = .06$$

(5)

4

3. (Chad)

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

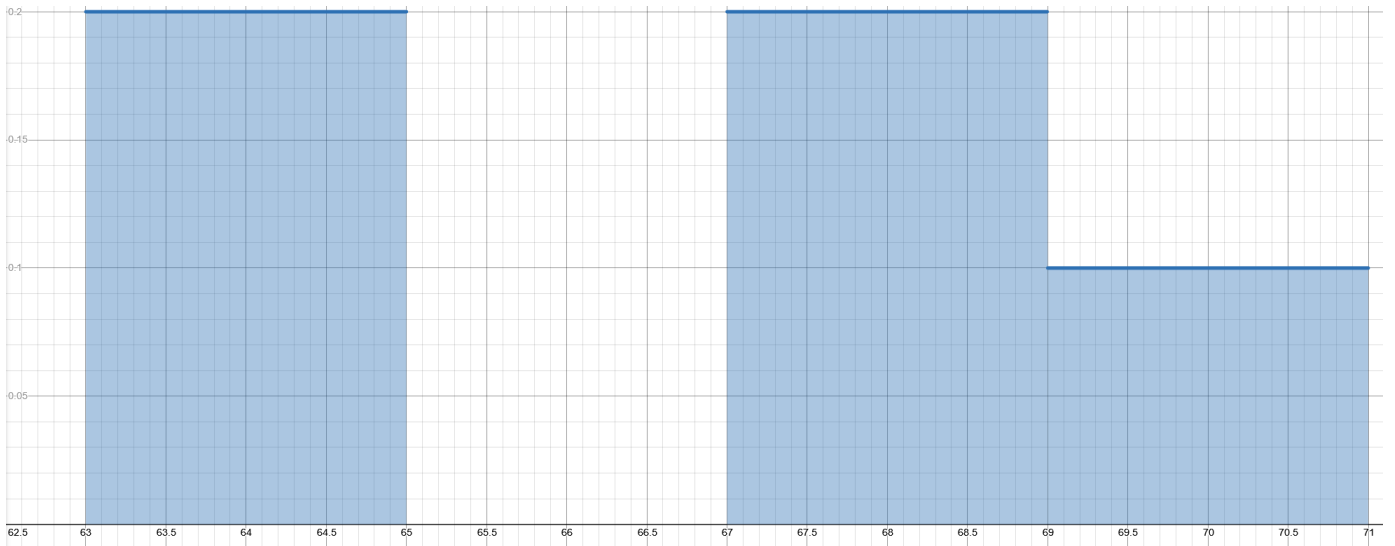Figure 2: Distribution of Temperature Given Chad Being in the Office

Figure 3: Distribution of Temperature Given No Chad the Office

(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?

If the temperature is 68 degrees we conclude that he is not at the office. This is because we have no data points of chad being at the office when the temperature is 68

degrees, but we do have observed data of Chad being gone when temperature is 68 degrees. If we use the kernel density estimates we calculated, $P(68 \text{ degrees} \mid \text{No Chad}) = .2$ and $P(68 \text{ degrees} \mid \text{Chad}) = 0$.

(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.

If we take this approach we can estimate our parameters in the following way:

$$P(\text{Chad}) = \frac{\text{Observations of Chad}}{\text{Total Observations}} = \frac{10}{15} = \frac{2}{3}$$

$$P(\text{No Chad}) = \frac{\text{Observations of No Chad}}{\text{Total Observations}} = \frac{5}{15} = \frac{1}{3}$$

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

We can use Bayes rule to predict the to predict if chad is in the office given its $64°$.

$$P(\text{Chad}|64°) = \frac{P(\text{Chad}) \times P(64°|\text{Chad})}{P(64°)}$$

$$P(\text{Chad}|64°) = \frac{P(\text{Chad}) \times P(64°|\text{Chad})}{P(\text{Chad})P(64|\text{Chad}) + P(\text{No Chad})P(64|\text{No Chad})}$$

$$P(\text{Chad}|64°) = \frac{\frac{2}{3} \times \frac{1}{10}}{(\frac{1}{3} \times \frac{2}{10}) + (\frac{2}{3} \times \frac{1}{10})} = \frac{1}{2} \tag{6}$$

And

$$P(\text{No Chad}|64°) = 1 - P(\text{Chad}|64°) = \frac{1}{2}$$

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

The problem is if we have a temperature is $57°$ we have no observations or KDE density around that point, so we have no way of predicting whether or not Chad is in the office. Parametric models circumvent this, as they have a set distribution that real world data molds. If we used something like a Gaussian distribution, then there would be real, non negative, probability of Chad being in the office or not in the office at any given temperature. This is one way that parametric modeling can help model phenomena when you do not have much observational data.

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

---

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.

Since $\tilde{s}$ and $\tilde{c}$ are conditionally independent given $\tilde{h}$, we can express the probability of $\tilde{h}$ given $\tilde{c}, \tilde{s}$ as:

$$p(h|s,c) = \frac{P(h)P(s|h)P(c|h)}{P(s,c)}$$

Now we can compute the MAP estimate of h given s, c

$$MAP(h) = \begin{cases} 0 & \text{if } P(1)P(s|1)P(c|1) < P(0)P(s|0)P(c|0) \\ 1 & \text{Otherwise} \end{cases}$$

(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.

---

[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

(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.

Since $\tilde{s}$ and $\tilde{c}$ and are conditionally independent given $\tilde{h}$, we can express the probability of $\tilde{h}$ given $\tilde{c}, \tilde{s}$ and as:

$$p(h|s,c,x) = \frac{P(h)P(s|h)P(c|h)f(x|h)}{P(s,c)f(x|s,c)}$$

Now we can compute the MAP estimate of h given s, c, x

$$MAP(h) = \begin{cases} 0 & if \ P(1)P(s|1)P(c|1)f(x|1) < P(0)P(s|0)P(c|0)f(x|0) \\ 1 & Otherwise \end{cases}$$

(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.

(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?

We honestly don't have much data, with only 218 observations estimating a joint distribution for all possible combinations of features would be intractable. I personally don't think that it would be advisable to estimate all of the joint distributions. When it comes to real world modelling, we must introduce some assumptions to fight off the curse of dimentionality. Otherwise, we would never be able to model anything. It's better to have models that are wrong, but useful, rather than no models at all.

In [28]:

```python
import numpy as np
import matplotlib.pyplot as plt
import os.path
import pandas as pd
import scipy
# load data
npz = np.load('heart_disease_data.npz')

df = pd.DataFrame.from_dict({item: npz[item] for item in npz.files}, orient='

train = df.iloc[0:50,:]

test = df.iloc[50:,:]

def ind_x_eq_val(x, val):
    return np.where(x==val)[0]

def count_x_eq_val(x, val):
    return len(ind_x_eq_val(x, val))/float(len(x))

def gaussian(x, mu, sig):
    return np.exp(-np.power(x - mu, 2.) / (2 * np.power(sig, 2.))) / sig / np
```

QUESTION (b)

Non-parametric model Compute empirical pmf, derive the conditional pmf, and estimate the MAP decision by the mode of posterior distribution $p_{\tilde{h}|\tilde{s},\tilde{c}}$. (The MAP estimates should be $\tilde{h} = 0$ or 1)

In [41]: ▶|
```python
# Estimate the pmf of H, i.e. P(H=0) and P(H=1)
P_H = np.zeros(2)

for i in range(2):
    P_H[i] = count_x_eq_val(df['heart_disease'],i)
    if i == 0:
        P_H0 = count_x_eq_val(df['heart_disease'],i)
    else:
        P_H1 = count_x_eq_val(df['heart_disease'],i)

# Estimate the conditional pmf of S given H, i.e. P(S|H=0) and P(S|H=1)
P_S_H0 = np.zeros(2)
P_S_H1 = np.zeros(2)
for ind_S in range(2):
    P_S_H0[ind_S] = len(df[(df['heart_disease'] == 0) & (df['sex']==ind_S)])
    P_S_H1[ind_S] = len(df[(df['heart_disease'] == 1) & (df['sex']==ind_S)])

# Estimate the conditional pmf of C given H, i.e. P(C|H=0) and P(C|H=1)
P_C_H0 = np.zeros(4)
P_C_H1 = np.zeros(4)
for ind_C in range(4):
    P_C_H0[ind_C] = len(df[(df['heart_disease'] == 0) & (df['chest_pain']==in
    P_C_H1[ind_C] = len(df[(df['heart_disease'] == 1) & (df['chest_pain']==in

h0_list = []
for s in range(2):
    for c in range(4):
        top = P_H[0]*P_S_H0[s]*P_C_H0[c]
        h0_list.append(top)

h1_list = []
for s in range(2):
    for c in range(4):
        top = P_H[1]*P_S_H1[s]*P_C_H1[c]
        h1_list.append(top)

map_estimate_S_C = []

for i in range(len(h0_list)):
    if h1_list[i] < h0_list[i]:
        map_estimate_S_C.append(0)
    else:
        map_estimate_S_C.append(1)

errors = 0
index = 0
df = df.iloc[0:50,:]
for s in range(2):
    for c in range(4):
        df_of_interest = df[(df['sex_test']==s) & (df['chest_pain_test']==c)]
        for i in df_of_interest['heart_disease_test']:
            if i != map_estimate_S_C[index]:
                errors += 1
        index += 1

error_rate_S_C = errors / len(df)
```

```
print("Probability of error " + str(error_rate_S_C))#
```

Probability of error 0.18

QUESTION (d)

Maximum likelihood estimates Find the parameters of two normal distributions ($\tilde{x}|\tilde{h} = 1$ and $\tilde{x}|\tilde{h} = 0$) that maximize the likelihood functions.

In [42]: ⏭

```python
df = pd.DataFrame.from_dict({item: npz[item] for item in npz.files}, orient='

## Estimate MLE of X given H
no_heart_cond = (df[df['heart_disease']==0])
heart_cond = (df[df['heart_disease']==1])
mean_X_H = np.zeros(2)
std_X_H = np.zeros(2)
mean_X_H[0]= np.mean(np.array(no_heart_cond['cholesterol']))
std_X_H[0] = np.std(np.array(no_heart_cond['cholesterol']))
mean_X_H[1]= np.mean(np.array(heart_cond['cholesterol']))
std_X_H[1] = np.std(np.array(heart_cond['cholesterol']))

n_plot = 100
for i in range(2):
    plt.figure(figsize=(12, 9))
    ax = plt.subplot(111)
    ax.spines["top"].set_visible(False)
    ax.spines["right"].set_visible(False)
    ax.get_xaxis().tick_bottom()
    ax.get_yaxis().tick_left()
    yticks = ax.yaxis.get_major_ticks()
    yticks[0].label1.set_visible(False)
    plt.xticks(fontsize=22)
    plt.yticks(fontsize=22)
    plt.xlabel("Cholesterol", fontsize=22)

    plt.hist(df.loc[df['heart_disease']==i]['cholesterol'],
             bins=60,stacked=True,edgecolor = "none", color="skyblue", densit

    plt.plot(np.linspace(50, 450, n_plot),gaussian(np.linspace(50, 450, n_plo
                         mean_X_H[i], std_X_H[i]), color="darkred", lw=2)
```
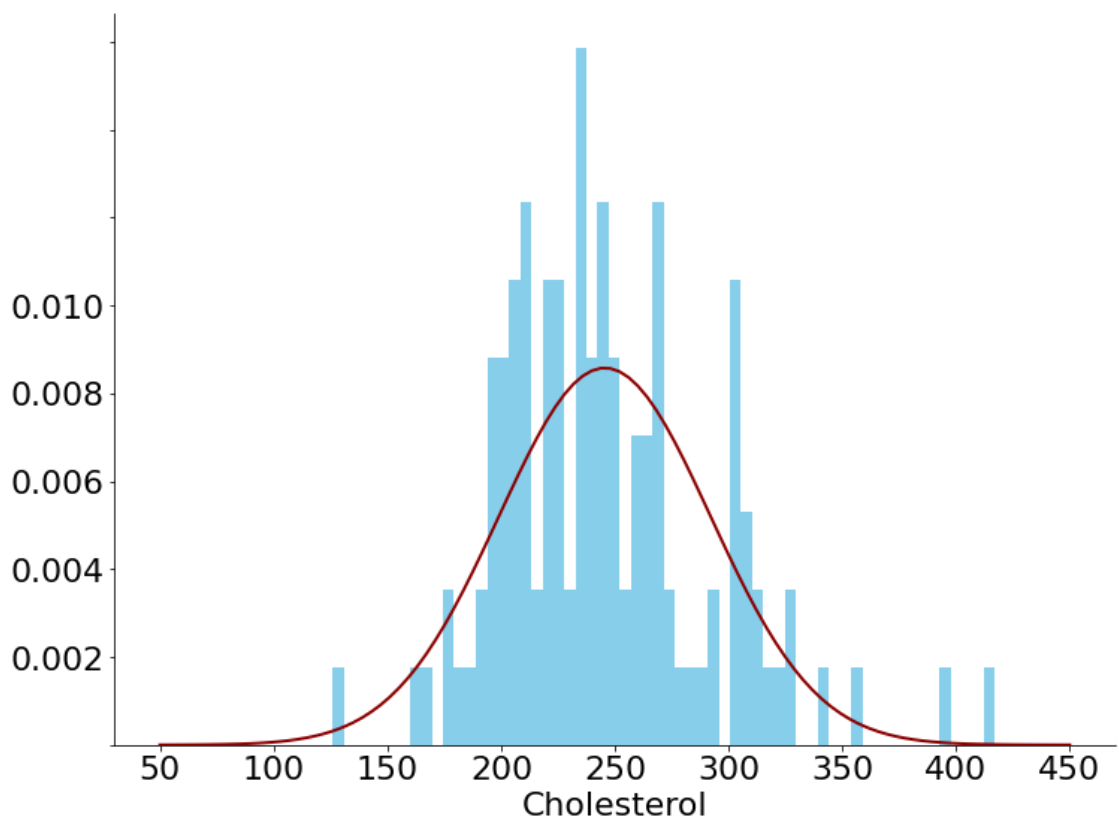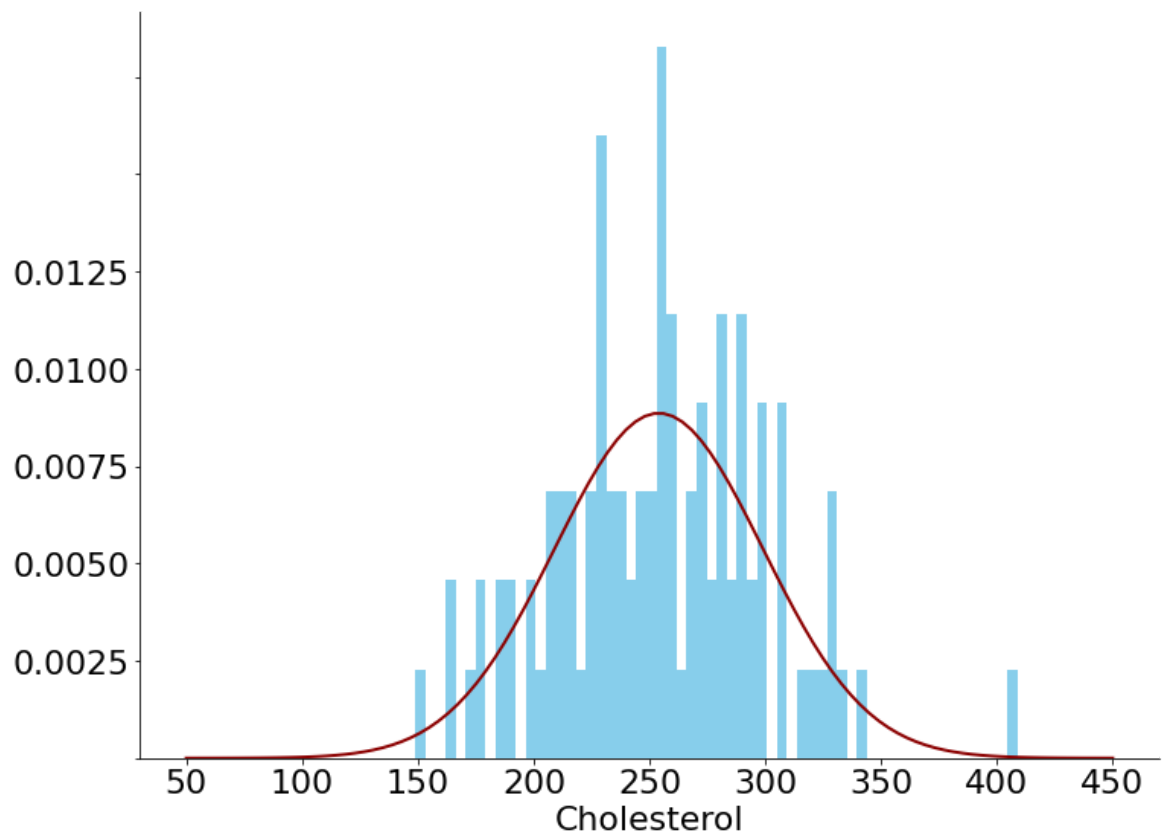
QUESTION(e)

MAP decision compute posterior $p_{\tilde{h}|\tilde{s},\tilde{c},\tilde{x}}$ and derive MAP

In [43]:
```python
from scipy import stats
```

In [44]:

```python
# Calculate the MAP estimate
MAP_list = []
for s in range(2):
    for c in range(4):
        for chol in set(df['cholesterol_test']):
            pdf_value_H0 = scipy.stats.norm.pdf(chol, mean_X_H[0], std_X_H[0]
            pdf_value_H1 = scipy.stats.norm.pdf(chol, mean_X_H[1], std_X_H[1]
            var1 = P_H1*P_S_H1[s]*P_C_H1[c]*pdf_value_H1
            var2 = P_H0*P_S_H0[s]*P_C_H0[c]*pdf_value_H0
            if var1 < var2: MAP_list.append(0)
            else: MAP_list.append(1)


MAP_estimate_S_C_X = max(set(MAP_list), key=MAP_list.count)

# Calculate the error rate
list_index = 0
error_count = 0
for s in range(2):
    for c in range(4):
        for chol in set(df['cholesterol_test']):
            sub_df = train[(train['sex_test'] == s) & (train['chest_pain_test
            for idx2 in sub_df['heart_disease_test']:
                if idx2 != MAP_list[list_index]:
                    error_count += 1
            list_index += 1

error_rate_S_C_X = error_count / len(train)
print("Probability of error using cholesterol " + str(error_rate_S_C_X))
```

Probability of error using cholesterol 0.14

In [ ]:

In [ ]:

In [ ]: