DSC 140A - Homework 06

Due: Wednesday, February 21

Problem 1.

Let X be a continuous random variable, and let Y be a binary random variable. Suppose the class conditional densities are know to be:

$$p_X(x \mid Y = 0) = \begin{cases} 1/5, & \text{if } 0 \le x \le 2 \\ 1/3, & \text{if } 2 < x \le 3 \\ 1/15, & \text{if } 3 < x \le 7 \end{cases}$$

$$p_X(x \mid Y = 1) = \begin{cases} 1/6, & \text{if } 0 \le x \le 1 \\ 1/8, & \text{if } 1 < x \le 5 \\ 1/6, & \text{if } 5 < x \le 7 \end{cases}$$

Suppose also that $\mathbb{P}(Y = 0) = 0.4$ and $\mathbb{P}(Y = 1) = 0.6$.

For what values of $x \in [0, 7]$ will the Bayes classifier predict $y = 1$?

Solution: To find out what the bayes classifier wil predict for $x \in [0, 7]$, we can use the bayes classifier rule:

$$p_X(x|Y = 0)\mathbb{P}(Y = 0) > p_X(x|Y = 1)\mathbb{P}(Y = 1)$$

We will split the intervals to based of the conditional densities given and solve for the values of x that satisfy the inequality.

Interval	$p_X(x\|Y = 0)\mathbb{P}(Y = 0)$	$p_X(x\|Y = 1)\mathbb{P}(Y = 1)$	Predicts
$0 \le x \le 1$	$\frac{1}{5} \cdot 0.4 = \frac{2}{25}$	$\frac{1}{6} \cdot 0.6 = \frac{1}{10}$	Predict $y = 1$
$1 < x \le 2$	$\frac{1}{5} \cdot 0.4 = \frac{2}{25}$	$\frac{1}{8} \cdot 0.6 = \frac{3}{40}$	Predict $y = 0$
$2 < x \le 3$	$\frac{1}{3} \cdot 0.4 = \frac{2}{15}$	$\frac{1}{8} \cdot 0.6 = \frac{3}{40}$	Predict $y = 0$
$3 < x \le 5$	$\frac{1}{15} \cdot 0.4 = \frac{2}{75}$	$\frac{1}{8} \cdot 0.6 = \frac{3}{40}$	Predict $y = 1$
$5 < x \le 7$	$\frac{1}{15} \cdot 0.4 = \frac{2}{75}$	$\frac{1}{6} \cdot 0.6 = \frac{1}{10}$	Predict $y = 1$

So looking at the table, the Bayes classifier will predict $y = 1$ for $x \in [0, 1] \cup [3, 7]$.

Problem 2.

Let X be a continuous random variable, and let Y be a random class label (1 or 0). Recall that the Gaussian probability density function (pdf) is given by:

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

Suppose $p_X(x \mid Y = 1)$ is a Gaussian pdf with $\mu = 2$ and $\sigma = 3$ and that $p_X(x \mid Y = 0)$ is also Gaussian with $\mu = 5$ and $\sigma = 3$. Suppose, too, that $\mathbb{P}(Y = 1) = \mathbb{P}(Y = 0) = \frac{1}{2}$.

Recall that the Bayes error is the probability that the Bayes classifier makes an incorrect prediction. What is the Bayes error for this distribution? Show your work.

Hint: you'll want some way to compute the area under a Gaussian. You can use the tables that appear in the back of your statistics book, or you can use something like `scipy.stats.norm.cdf`. We'll let you read the documentation to see how to use it, but it may be helpful to remember that if F is the cumulative density function for a distribution with density f, then $\int_a^b f(x)\,dx = F(b) - F(a)$.

Solution: Using the information we havev above, we have that the equations for the joint probability of X and Y are:

$$p(x, y) = p(x|y)p(y) \begin{cases} \mathcal{N}(x \,|\, 2, 3^2) \cdot \frac{1}{2} & \text{if } y = 1 \\ \mathcal{N}(x \,|\, 5, 3^2) \cdot \frac{1}{2} & \text{if } y = 0 \end{cases}$$

This gives us the 2 equations for both curves for when $y = 1$ and $y = 0$. Then as we know, the bayes errors is given by the area of the region where the two curves overlap. So we get 2 bayes errors, one for when $y = 1$ and one for when $y = 0$. In order to get the area, we first need to find the bayes classifier boundary. This is given by the point where the two curves intersect. We can solve for this by setting the two equations equal to each other and solving for x:

$$\frac{1}{\sqrt{2\pi 3^2}} \exp\left(-\frac{(x-2)^2}{2 \cdot 3^2}\right) \cdot \frac{1}{2} = \frac{1}{\sqrt{2\pi 3^2}} \exp\left(-\frac{(x-5)^2}{2 \cdot 3^2}\right) \cdot \frac{1}{2}$$

Upon inspection of the equation we can see that the $\frac{1}{2}$ and $\frac{1}{\sqrt{2\pi 3^2}}$ terms cancel out due to $\sigma_1 = \sigma_2$ and $\mathbb{P}(Y = 1) = \mathbb{P}(Y = 0)$. Then we can solve for x:

$$\exp\left(-\frac{(x-2)^2}{2 \cdot 3^2}\right) = \exp\left(-\frac{(x-5)^2}{2 \cdot 3^2}\right)$$

We can then take the natural log of both sides and multiply by $-2 \cdot 3^2$ to get:

$$(x-2)^2 = (x-5)^2$$

Then we can solve for x:

$$\begin{aligned} (x-2)^2 &= (x-5)^2 \\ x^2 - 4x + 4 &= x^2 - 10x + 25 \\ 6x &= 21 \\ x &= \frac{21}{6} = 3.5 \end{aligned}$$

Now that we have the boundary x where the two curves intersect, we can find the bayes error by integrating the two curves from $-\infty$ to 3.5 and from 3.5 to ∞. We can use the cumulative density function for the Gaussian distribution to find the area under the curve. We can use the following equation to find the area under the curve using scipy:

$$\text{Error} = 0.3085375387259869$$

The code for the problem is given below:

```python
from scipy.stats import norm

# Parameters
mu1, sigma = 2, 3  # For Y = 1
mu2 = 5            # For Y = 0
x_star = 3.5       # Decision boundary

error_Y1 = 1 - norm.cdf(x_star, mu1, sigma)
# 1 - cdf cause cdf calculates the area from -inf to x_star
# for the lower mean distribution we want the area from x_star to inf

error_Y0 = norm.cdf(x_star, mu2, sigma)

# average the two errors to get the bayes error
bayes_error = 0.5 * (error_Y1 + error_Y0)
print(bayes_error)
```

Problem 3.

The file linked below contains a data set of 150 samples. The first column contains a single continuous feature, X, assumed to have been drawn from an unknown probability density. The second column contains the binary class label Y.

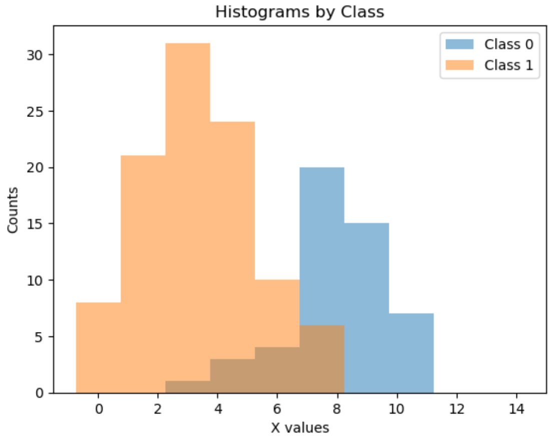
`https://f000.backblazeb2.com/file/jeldridge-data/011-univariate_density_estimation/data.csv`

In this problem, use a histogram estimator with bins $[0, 1.5), [1.5, 3), \ldots, [13.5, 15)$ to estimate the requested probabilities. In each part, show your code and provide your reasoning.

Hint: you may find `np.histogram` useful.

a) Estimate $\mathbb{P}(Y = 1 \mid X = 6.271)$ directly.

Solution: Given the histogram drawn by the code below,



Using the code below as well, we can estimate $\mathbb{P}(Y = 1 \mid X = 6.271)$ directly by counting the number of samples in the bin $[6, 7.5)$ with $Y = 1$ and dividing by the total number of samples in that bin. This givens the estimate probability as 0.7142857142857143

b) Estimate $\mathbb{P}(Y = 1 \mid X = 6.271)$ by estimating all of: 1) the marginal density $p_X(x)$, 2) the class-conditional density $p_X(x \mid Y = 1)$, and 3)the class prior $\mathbb{P}(Y = 1)$, and then applying Bayes' rule.

Solution: All code for the question is below.

Now estimating using Bayes' rule, we need more information from the data in order to calculate the $\mathbb{P}(Y = 1|X = 6.271)$.

So first we calculate the class prior $\mathbb{P}(Y = 1)$, by just counting the number of samples with $Y = 1$ and dividing by the total number of samples. This gives us $\mathbb{P}(Y = 1) = 0.6666666666666666$.

Next we want to calculate the marginal density $p_X(x)$ given $\mathbb{P}(Y = 1)$. To do so we given all the counts of each bins that are class 1 and divide it by the number of samples with $Y = 1$. This gives us the following counts for each bin. Then since $X = 6.271$ we can get the bin density by indexing into the counts array. Giving us the class conditional density $p_X(x \,|\, Y = 1)$ as 0.06666666666666667.

Finally we estimate, the marginal density $p_X(x)$ by first calculating the counts of each bin and dividing by the total number of samples. Then we can index into the counts array to get the marginal density when $X = 6.271$ as 0.06222222222222222.

Finally using bayes' rule we can calculate the $\mathbb{P}(Y = 1|X = 6.271)$ as,

$$\mathbb{P}(Y = 1|X = 6.271) = \frac{0.0666666 \cdot 0.66666}{0.062222222} = 0.7142857142857142$$

c) For what values of $x \in [0, 15]$ will the Bayes classifier predict $y = 1$?

Solution: Finally, we can estimate the values of x that the bayes classifier will predict $y = 1$ by calculating an additional piece of information which is the class conditional density $p_X(x \,|\, Y = 0)$. Then we can compare the class conditionals to each other, and if $p_X(x \,|\, Y = 1) > p_X(x \,|\, Y = 0)$ then we predict $y = 1$. This gives us the following values of x that the bayes classifier will predict $y = 1$,

Bins = [0, 1.5), [1.5, 3), [3 ,4.5), [4.5, 6), [6, 7.5)

Code

```
import numpy as np
import matplotlib.pyplot as plt

data = np.loadtxt('data.csv', delimiter=',')
X = data[:, :-1]
y = data[:, -1]

X_0 = X[y == 0]
X_1 = X[y == 1]

bins = [0, 1.5, 3, 4.5, 6, 7.5, 9, 10.5, 12, 13.5, 15]

hist_0, bin_edges = np.histogram(X_0, bins=bins)
hist_1, _ = np.histogram(X_1, bins=bins)


# Plotting the histograms
# Hist 0
plt.bar(bin_edges[:-1], hist_0, width=1.5, alpha=0.5, label='Class 0')

# Hist 1
plt.bar(bin_edges[:-1], hist_1, width=1.5, alpha=0.5, label='Class 1')

plt.xlabel('X values')
plt.ylabel('Counts')
plt.title('Histograms by Class')
plt.legend()
plt.show()
```

```
# Question a
# Index where X = 6.271
idx = 4

count_Y1 = hist_1[idx]
count_total = hist_0[idx] + hist_1[idx]

prob_Y1_given_X = count_Y1 / count_total

print(prob_Y1_given_X)

# Question b
P_Y1 = np.mean(y == 1)
print("Class 1 prior: {}".format(P_Y1)) # Class 1 prior

# Bin width
bin_width, total_samples = 1.5, len(X)

# Normalized histograms (densities)
density_X = (hist_0 + hist_1) / (total_samples * bin_width)
density_X_given_Y1 = hist_1 / (np.sum(hist_1) * bin_width)

p_X_6271 = density_X[4]
print("p(X=6.271): {}".format(p_X_6271)) # P(X=6.271)

p_X_given_Y1_6271 = density_X_given_Y1[4]
print("P(X=6.271|Y=1): {}".format(p_X_given_Y1_6271)) # P(X=6.271|Y=1)

P_Y1_given_X_6271_Bayes = (p_X_given_Y1_6271 * P_Y1) / p_X_6271
print("P(Y=1|X=6.271) with Bayes: {}".format(P_Y1_given_X_6271_Bayes)) # P(Y=1|X=6.271) with Bayes


# Question c
# Density for all X values given Y = 0
density_X_given_Y0 = hist_0 / (np.sum(hist_0) * bin_width)

# Using the bayes classifier
pred_Y1_idx = density_X_given_Y1 > density_X_given_Y0

pred_bins = bin_edges[:-1][pred_Y1_idx]

print("Predicted bins: {}".format(pred_bins)) # Predicted bins
```