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:

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

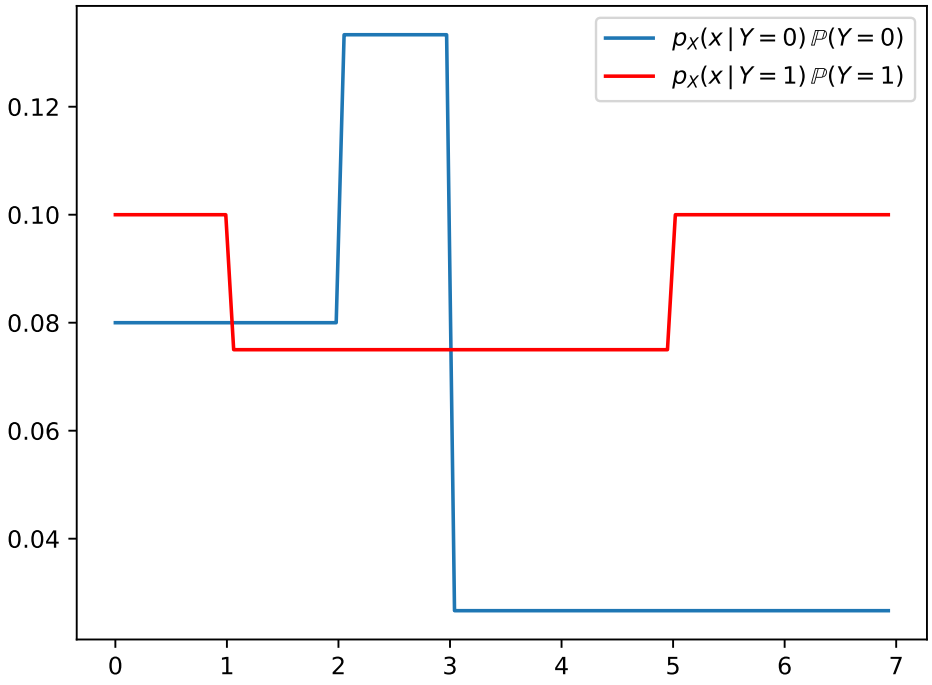
$$p_X(x \mid Y = 1) = \begin{cases} 1/6, & \text{if } 0 \leq x \leq 1 \\ 1/8, & \text{if } 1 < x \leq 5 \\ 1/6, & \text{if } 5 < x \leq 7 \end{cases}$$

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: The Bayes classifier will predict $y = 1$ for all $x \in [0, 1] \cup (3, 7]$.

The Bayes classifier tells us to predict whichever is larger: $p_X(x \mid Y = 1)\mathbb{P}(Y = 1)$ or $p_X(x \mid Y = 0)\mathbb{P}(Y = 0)$. To determine this, we can draw both plots:



We see that the red curve is larger than the blue curve on the intervals $[0, 1]$ and $(3, 7]$, and this is exactly where the Bayes classifier will predict $y = 1$.

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: Because the Gaussians have the same width and the classes are equally probable, the decision boundary is exactly halfway between their means, at $x = 3.5$. Any point to the left of 3.5 is predicted Class 0, and everything to the right is predicted Class 1.

The Bayes error is the probability that a point is misclassified. This can occur in two ways: either the point came from class $Y = 1$, but the prediction is for class 0, or the point came from class $Y = 0$ but was predicted to be class 1. The total probability of an error is the sum of the probabilities of either case occurring.

Consider the first case, where the point came from class $Y = 1$ but is predicted to be Class 0. This will occur when the point is to the left of 3.5. What is the probability that a point comes from $Y = 1$ and is to the left of 3.5? It is:

$$\begin{aligned}\mathbb{P}(\text{x} < 3.5 \text{ and } \text{Y} = 1) &= \mathbb{P}(\text{x} < 3.5 \mid Y = 1)\mathbb{P}(Y = 1) \\ &= \mathbb{P}(\text{x} < 3.5 \mid Y = 1) \times 0.5\end{aligned}$$

The probability that $x < 3.5$ given that it comes from Class 1 is computed as the area under the curve of the Gaussian for Class 1's distribution from $-\infty$ to 3.5. This can be computed with `scipy.stats.norm.cdf(3.5, 5, 3)` the result is 0.308. Therefore:

$$\begin{aligned}\mathbb{P}(\text{x} < 3.5 \text{ and } \text{Y} = 1) &= \mathbb{P}(\text{x} < 3.5 \mid Y = 1)\mathbb{P}(Y = 1) \\ &= 0.308 \times 0.5 \\ &= 0.154\end{aligned}$$

Likewise, the probability of the second case is:

$$\begin{aligned}\mathbb{P}(\text{x} > 3.5 \text{ and } \text{Y} = 0) &= \mathbb{P}(\text{x} > 3.5 \mid Y = 0)\mathbb{P}(Y = 0) \\ &= \mathbb{P}(\text{x} > 3.5 \mid Y = 0) \times 0.5\end{aligned}$$

The probability that $x > 3.5$ when drawn from the Gaussian with mean $\mu = 2$ is:

```
1 - scipy.stats.norm.cdf(3.5, 2, 3)
```

This is also 0.308, which could have been recognized from symmetry.

Therefore:

$$\begin{aligned}\mathbb{P}(\text{x} > 3.5 \text{ and } \text{Y} = 0) &= \mathbb{P}(\text{x} > 3.5 \,|\, Y = 0)\mathbb{P}(Y = 0) \\ &= .308 \times 0.5\end{aligned}$$

All together, then, the Bayes error is 0.308.

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 $\mathbb{P}(Y = 1 \,|\, X = 6.271)$ directly.

Solution: To estimate $\mathbb{P}(Y = 1 \,|\, X = 6.271)$ directly, we filter the data to include only the samples that are within the same histogram bin as $x = 6.271$, and then find the fraction of those samples that have $y = 1$.

In code:

```
data = np.loadtxt('data.csv', delimiter=',')

# histogram bin containing 6.271
bin_lower, bin_upper = (6, 7.5)

# split the two columns of data into x and y arrays
x, y = data.T

# filter the data to only include the x values in the bin
y[(bin_lower <= x) & (x < bin_upper)].mean()
```

This gives 0.714, approximately.

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

Solution:

```
# filter the data into two subsets:
# the data from class 1 and the data from class 0
samples_1 = data[data[:, 1] == 1][:, 0]
samples_0 = data[data[:, 1] == 0][:, 0]

bins = np.arange(0, 15.1, 1.5)

# create class-conditional density histograms
p_1 = np.histogram(samples_1, bins=bins, density=True)[0]
```

```
p_0 = np.histogram(samples_0, bins=bins, density=True)[0]

# create a histogram for the marginal density
p = np.histogram(x, bins=bins, density=True)[0]

# our estimate of P(Y = 1)
p_y1 = y.mean()

# the point x = 6.271 falls within the fifth bin (bin at index 4)
# of the histogram. To compute p(x | Y = 1) * P(Y = 1) / p(x), we write:
p_1[4] * p_y1 / p[4]
```

This yields the same answer as part (a): 0.714.

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

Solution: The Bayes classifier predicts $y = 1$ whenever $p_X(x \,|\, Y = 1)\mathbb{P}(Y = 1) > p_X(x \,|\, Y = 0)\mathbb{P}(Y = 0)$. Because we are using a histogram estiamtor, the density estimates are piecewise constant, and the prediction can only change at the bin boundaries. Therefore, it's enough to check the prediction within each bin.

We can do this in one line of code: `p_1 * y.mean() > p_0 * (1 - y.mean())`

The result is:

```
array([ True,  True,  True,  True,  True, False, False, False, False, False])
```

This tells us that the prediction is for $y = 1$ in the first 5 bins, and for $y = 0$ in the remaining bins. The first five bins cover the interval $[0, 7.5)$, and so the Bayes classifier will predict $y = 1$ for $x \in [0, 7.5)$.