DSC 140A - Homework 07

Due: Wednesday, May 22

Instructions: Write your solutions to the following problems either by typing them or handwriting them on another piece of paper. Show your work or provide justification unless otherwise noted. If you write code to solve a problem, include the code by copy/pasting or as a screenshot. You may use `numpy`, `pandas`, `matplotlib` (or another plotting library), and any standard library module, but no other third-party libraries unless specified. Submit homeworks via Gradescope by 11:59 PM.

Problem 1.

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.

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.

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

Problem 2.

The Rayleigh distribution has pdf:

$$p(x) = \frac{x}{\sigma^2} e^{-x^2/(2\sigma^2)},$$

where σ is a parameter.

Suppose a data set of points $x_1, \ldots, x_n$ is drawn from a Rayleigh distribution with unknown parameter σ. It was shown in discussion section that the log-likelihood of σ given this data is:

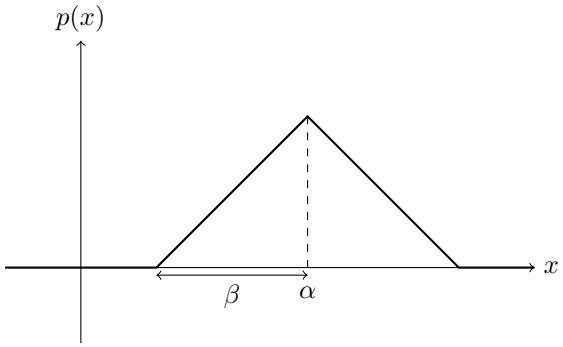
$$\tilde{L}(\sigma | x_1, \ldots, x_n) = n \ln \frac{1}{\sigma^2} + \sum_{i=1}^{n} \ln x_i - \frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2$$

Show that the maximum likelihood estimate of σ is:

$$\sigma_{\text{MLE}} = \sqrt{\frac{1}{2n} \sum_{i=1}^{n} x_i^2}.$$

Problem 3.

Justin's triangle is a parametric density function p that looks like a triangle. It has two parameters, α and β, which control the location and width of the triangle, respectively. A plot of the pdf is shown below:



a) Show that $p(\alpha) = \frac{1}{\beta}$.

b) Write down the formula for the density function $p(x)$.

c) Let $\mathcal{X} = \{2, 3, 5, 7, 8\}$ be a data set of 5 points. Note that this data set is symmetric around the middle point, 5.

It can be shown that, in this case, the maximum likelihood estimate for α is 5.

What is the maximum likelihood estimate for β?

Problem 4.

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 the first problem on this homework, you used a histogram estimator to apply the Bayes classification rule. In this problem, you will instead estimate the class-conditional densities by fitting Gaussians using the method of maximum likelihood. In each part, show your code and provide your reasoning.

Note: you should *not* use `sklearn`, `scipy`, or any other library to fit the Gaussian densities. You should instead calculate the maximum likelihood estimates directly (using `numpy` or `pandas` to do this is fine).

a) Estimate the class-conditional densities $p_X(x \mid Y = 0)$ and $p_X(x \mid Y = 1)$ with Gaussians using the method of maximum likelihood and report the estimated parameters.

b) Let $\tilde{p}_X(x \mid Y = 0)$ and $\tilde{p}_X(x \mid Y = 1)$ be the estimated class-conditional densities from the previous part, and let $\tilde{\mathbb{P}}(Y = 1)$ be the estimate for $\mathbb{P}(Y = 1)$.

Plot $\tilde{p}_X(x \mid Y = 0) \cdot \tilde{\mathbb{P}}(Y = 0)$ and $\tilde{p}_X(x \mid Y = 1) \cdot \tilde{\mathbb{P}}(Y = 1)$ on the same axis. Label your plot so that the grader can tell which Gaussian corresponds to which class. Remember to show your code.

c) Suppose a new point is observed at $x = 6.271$. What class does the Bayes classifier predict for x when the estimated densities and probabilities are used? Remember to show your code, and determine the predicted class through calculation (and not by visual inspection of the plot from the previous part).

d) The `scipy.optimize.fsolve` function can be used to find the *roots* of a function f; that is, the places where function $f(x) = 0$. Use `fsolve` to find the decision boundary for the classifier you trained above. Show your code.

Note that there may actually be multiple decision boundaries at the extremes, but there is one clear decision boundary in the middle of the data, as should be evident in your plot; find that one.