DSC 140A - Discussion 04

Problem 1.

Suppose the following data are observed:

x	y
2.1	1
4.7	1
2.3	0
0.8	0
1.3	1
5.2	1
7.4	0
9.4	1
3.9	1

You may assume that the x-values were drawn from a continuous distribution, and the y-values represent the label of each point.

To estimate all probabilities below, use a histogram estimator with 5 equally-sized bins spanning the interval from 0 to 10. The bins should include their starting point and exclude their ending point.

a) What is the estimated density $p_X(x)$ at $x = 3$?

Solution: Let's first create our histogram for the marginal density of X.

To do that we will first count up the number of x's that fall within each bin.

bins	count
$[0, 2)$	2
$[2, 4)$	3
$[4, 6)$	2
$[6, 8)$	1
$[8, 10)$	1

Now to estimate the density within each bin, we can divide the count (i.e. number of data points within each bin) by the total number of data points times the width of the bin. In this case we have 9 total data points and each bin has a width of 2. Therefore we can divide each of the counts above by $9 \times 2 = 18$. So our histogram would have the following counts.

bins	count
$[0, 2)$	$\frac{1}{9}$
$[2, 4)$	$\frac{1}{6}$
$[4, 6)$	$\frac{1}{9}$
$[6, 8)$	$\frac{1}{18}$
$[8, 10)$	$\frac{1}{18}$

So to estimated density at $x = 3$, we can see that $x \in [2, 4)$ so $p_X(3) = \frac{1}{6}$.

b) What is the estimated probability that a new point x is in the interval $[3, 4]$?

Solution:

$$\mathbb{P}(x \in [3,4]) = \int_3^4 p_X(x)dx = 1 \times \frac{1}{6} = \frac{1}{6}$$

. Still using our histogram in part 1, we can see that the area of the histogram between 3 and 4 is $1 \times \frac{1}{6}$ which corresponds to the probability that x falls between 3 and 4.

c) What is the estimated conditional density $p(x \mid Y = 1)$ at the point $x = 2$?

Solution:

First we find the x points such that y is 1.

x	y
2.1	1
4.7	1
1.3	1
5.2	1
9.4	1
3.9	1

Then, we do the same thing as above in part 1. We'll count the values in each bin for this data.

bins	count
$[0, 2)$	1
$[2, 4)$	2
$[4, 6)$	2
$[6, 8)$	0
$[8, 10)$	1

Now the total number of points is 6 with the same bin widths of 2 so every point is now divided by 12.

Our density histogram now would have these values.

bins	count
$[0, 2)$	$\frac{1}{12}$
$[2, 4)$	$\frac{1}{6}$
$[4, 6)$	$\frac{1}{6}$
$[6, 8)$	0
$[8, 10)$	$\frac{1}{12}$

To calculate the estimated density $p(x \mid Y = 1)$ at $x = 2$ we have $p(2| \, Y = 1) = \frac{1}{6}$

d) Using the Bayes classification rule, what is the predicted label y of a new point $x = 2.5$?

Solution:

Bayes classification rule is shown below.

$$\mathbb{P}(Y = y|X = x) = \frac{p(x|Y = y)\mathbb{P}(Y = y)}{p_X(x)}$$

First let's find

$$\mathbb{P}(Y=1|X=2.5)=\frac{p(2.5|Y=1)\mathbb{P}(Y=1)}{p_X(2.5)}$$

There are 6 points where $y=1$ and in total 9 points. So $\mathbb{P}(Y=1)=\frac{2}{3}$.
We have $p(2.5|Y=1)=\frac{1}{6}$ from the histogram in part 2 and $p_X(2.5)=\frac{1}{6}$ from the histogram in part 1. Now we can plug it in to get

$$\mathbb{P}(Y=1|X=2.5)=\frac{\frac{1}{6}\times\frac{2}{3}}{\frac{1}{6}}=\frac{2}{3}$$

As $\frac{2}{3}>\frac{1}{3}$ we will predict the label for $x=2.5$ as $y=1$.

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 independent points $x_1,\dots,x_n$ is drawn from a Rayleigh distribution with unknown parameter σ. Show that the log-likelihood of σ given this data is:

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

Solution:

$$\begin{aligned}
L(\sigma|x_1,\dots,x_n)&=\sum_{i=1}^{n}\log(p(x_i|\sigma))\\
&=\sum_{i=1}^{n}\log(\frac{x_i}{\sigma^2}e^{-x_i^2/(2\sigma^2)})\\
&=\sum_{i=1}^{n}(\log(\frac{x_i}{\sigma^2})+\log(e^{-x_i^2/(2\sigma^2)}))\\
&=\sum_{i=1}^{n}(\log(x_i)+\log(\frac{1}{\sigma^2})-\frac{x_i^2}{2\sigma^2})\\
&=n\log\frac{1}{\sigma^2}+\sum_{i=1}^{n}\log x_i-\frac{1}{2\sigma^2}\sum_{i=1}^{n}x_i^2
\end{aligned}$$