Shown below is a data set where the X is a continuous random variable drawn from a density $p(x)$, and Y is a discrete variable taken from a distribution on $\{0, 1\}$.

X	Y
7.2	0
11.3	1
8.0	1
5.1	0
5.6	1
12.3	1
13.1	1
10.9	0
12.0	1
5.0	0

Estimate $p(10.1)$ using a histogram estimators with bins $[0, 2), [2, 4), [4, 6), [6, 8), [8, 10), [10, 12)$ and $[12, 14)$.

Recall that we can estimate densities using histogram estimators as follows,

$$f(x) \text{ within bin } i = \frac{\#\text{data points } \in [a_i, b_i)}{n \times (b_i - a_i)}.$$

Since $10.1 \in [10, 12)$ and there are 2 data points in this bin, $p(10.1) = \frac{2}{10 \times 2} = 0.1$.

Suppose a density estimate $f : \mathbb{R}^3 \to \mathbb{R}^1$ is made using histogram estimators with bins having a length of 2 units, a width of 3 units, and a height of 1 unit.

What is the largest value that $f(\vec{x})$ can possible have? Write your answer as a decimal number with at least 3 digits of precision.

The largest value that $f(\vec{x})$ can have is if all n data points were inside one bin, i.e.

$$f(\vec{x}) = \frac{n}{n \times \text{“bin volume”}} = \frac{1}{2 \times 3 \times 1} = \frac{1}{6} \approx 0.16667.$$

Let $(\vec{x}^{(1)}, y_1), \ldots, (\vec{x}^{(n)}, y_n)$ be a set of n points in a binary classification problem, where $\vec{x}^{(i)} \in \mathbb{R}^2$ and $y_i \in \{0, 1\}$.

Suppose a classifier is trained by estimating the class-conditional densities with histograms using rectangular bins and applying the Bayes classification rule.

True or False: it is always possible to achieve a 100% training accuracy with this classifier by choosing the rectangular bins to be sufficiently small. You may assume that no two points $\vec{x}^{(i)}$ and $\vec{x}^{(j)}$ are identical.

(x) True
( ) False

See the following video explanation.

Suppose $\mathcal{L}(\theta)$ is a likelihood function for a parameter theta, and let $\tilde{\mathcal{L}}(\theta) = \ln \mathcal{L}(\theta)$ be the log likelihood. True or False: if θ^* maximizes $\tilde{\mathcal{L}}$, it also maximizes $\mathcal{L}$.

(X) True
( ) False

Since the natural log, $\ln(x)$, is a monotonically increasing function, i.e. it preserves the property that if $a \leq b$, then $f(a) \leq f(b)$ for all values in its domain ($x \in \mathbb{R}, x > 0$), the value that maximizes $\tilde{\mathcal{L}}(\theta)$ will also maximize $\mathcal{L}(\theta)$ and vice versa.

Consider the data shown below:

X	Y
3	1
1	0
2	1
1	0
-1	0
0	0
4	1

Suppose the density $p_X(x \mid Y = 1)$ is modeled as a Gaussian. What is the maximum likelihood estimate for this Gaussian's σ parameter?

Recall that the MLE formula for σ is

$$\sigma_{\text{MLE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x^{(i)} - \mu_{\text{MLE}})^2}.$$

Consequently, we will need to evaluate μ_{MLE} as well.

$$
\begin{aligned}
\mu_{\text{MLE}} &= \frac{1}{n}\sum_{i=1}^{n} x^{(i)} \\
&= \frac{1}{3}(3+2+4) \\
&= 3 \\
\sigma_{\text{MLE}} &= \sqrt{\frac{1}{3}(0^2+(-1)^2+1^2)} \\
&= \sqrt{\frac{2}{3}} \\
&\approx 0.816
\end{aligned}
$$

Consider a new point, $X = 1$. What label does the Bayes classifier predict if the class densities $p_X(x \mid Y = 1)$ and $p_X(x \mid Y = 0)$ are estimated by fitting Gaussians using the maximum likelihood estimates?

( ) Class 1
(X) Class 0

Suppose we fit each of the class densities using MLE and obtain

$$X|Y = 1 \sim \mathcal{N}\left(3, \frac{2}{3}\right)$$

and

$$X|Y = 0 \sim \mathcal{N}\left(\frac{1}{4}, \frac{11}{16}\right).$$

Then, for a new point, $X = 1$,

$$p_X(x = 1|Y = 1) \approx 0.02433 < 0.31960 \approx p_X(x = 1|Y = 0),$$

so the Bayes classifier shall predict Class 0.

Suppose data points $x_1, \ldots, x_n$ are drawn iid from an arbitrary, unknown distribution with density f.

True or False: it is guaranteed that, given enough data (that is, as $n \to \infty$), a Gaussian fit to the data using the method of maximum likelihood will approximate the underlying density f arbitrarily well.

( ) True
(x) False

The "arbitrary, unknown distribution" that the data are drawn from may not be Gaussian, so fitting such data to a Gaussian would be an inaccurate estimation.

Suppose a discrete random variable X takes on values of either 0 or 1 and has the distribution $\mathbb{P}(X = x) = \theta^x(1-\theta)^{1-x}$ where $\theta \in [0, 1]$ is a parameter.

Given a data set $x_1, \ldots, x_n$, what is the maximum likelihood estimator for θ?

( ) $\sum_{i=1}^n \log x_i$
( ) $\frac{1}{n}\sum_{i=1}^n \log x_i$
( ) $\sum_{i=1}^n x_i$
(x) $\frac{1}{n}\sum_{i=1}^n x_i$
( ) x_i/n

Recall from your probability/statistics courses that this discrete random variable X has the distribution corresponding to the Bernoulli random variable! Perhaps, you have already derived/known that the MLE for the Bernoulli parameter, p, is

$$p_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

If that was not the case, not to fret! Alternatively, one can simply derive the MLE for this distribution using approaches learned in lecture.

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$$

$$\ell(\theta) = \sum_{i=1}^{n} x_i \ln(\theta) + (1-x_i)\ln(1-\theta)$$

$$\frac{\partial \ell}{\partial \theta} = \sum_{i=1}^{n} x_i \left(\frac{1}{\theta}\right) + (1-x_i)\left(\frac{-1}{1-\theta}\right)$$

$$\frac{1}{1-\hat{\theta}}(n - \sum_{i=1}^{n} x_i) = \frac{1}{\hat{\theta}} \sum_{i=1}^{n} x_i$$

$$\theta_{\text{MLE}} = \hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

(Make sure you are familiar with your logarithm properties, basic derivatives, and algebra!)