1. Consider using $k$-means on the unsupervised 1D dataset $\{x\} = \{1, 3, 5, 10, 12\}$ to create $k = 2$ clusters. Suppose the two initial randomly-chosen cluster centroids are $c_1 = 3$ and $c_2 = 5$.

(a) What are the centroids after the first iteration of $k$-means?

\[ c_1 = \quad \text{and} \quad c_2 = \quad \]

(b) What are the centroids after the second iteration?

\[ c_1 = \quad \text{and} \quad c_2 = \quad \]

2. For each situation, indicate which hyperparameter search strategy, $G =$ grid search or $R =$ random search, is more likely to be successful. Suppose computation time is limited.

(a) $\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\:\\]
3. Consider the use of bagging applied to classification decision trees of depth 1 (one decision node and two leaf nodes per tree). A training data set, on the left, consists of \( \{(x, y)\} = \{(x, y)\} \) because \( x \) has only one feature, \( x \). It is followed by \( B = 3 \) bootstrap resamples created by sampling with replacement from the training data.

<table>
<thead>
<tr>
<th>Training data</th>
<th>Resample #1</th>
<th>Resample #2</th>
<th>Resample #3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x )</td>
<td>( y )</td>
<td>( x )</td>
<td>( y )</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

Consider making a prediction for \( x = 2 \).

(a) What prediction is made by the tree trained on Resample #1? \( \hat{y} = \) _____________

(b) What prediction is made by the tree trained on Resample #2? \( \hat{y} = \) _____________

(c) What prediction is made by the tree trained on Resample #3? \( \hat{y} = \) _____________

(d) What prediction is made by this bagging classifier? \( \hat{y} = \) _____________
4. Here is a graph of 1D data \( \{x_i\} = \{x_i\} = \{1, 2, 4\} \) and corresponding Gaussian curves \( \{f_{\mu=x_i, \sigma=b}(x)\} \) made with bandwidth \( b = 0.25 \).

(a) Supposing the data were randomly sampled from some population, use kernel density estimation to estimate the population's probability density \( f(x) \) at \( x = 1 \).
Based on the plot, the estimate is \( \hat{f}_{b=0.25}(1) \approx \) __________.

(b) Estimate the density at \( x = 1.5 \).
Based on the plot, the estimate is \( \hat{f}_{b=0.25}(1.5) \approx \) __________.

(c) On the figure above, draw the estimated density function over the interval \([0, 6]\).
5. Consider the following questions about model assessment.

(a) Consider a classifier trained on examples \((x, y)\) in the first two columns of the table below that makes the predictions on training data in the third column.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>(\hat{y})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0, 0)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(1, 4)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>(3, -2)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>(3, 0)</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Complete the corresponding confusion matrix:

<table>
<thead>
<tr>
<th>actual</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(b) The classifier is evaluated on unseen test data yielding this confusion matrix:

<table>
<thead>
<tr>
<th>actual</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>

What is the precision on the test data?

(c) What is the recall on the test data?

(d) What is the accuracy on the test data?

(e) For a classifier that is randomly guessing with \(P(\hat{y} = 1) = \frac{1}{3}\), what is the AUC?

(f) For a classifier with TPR = 1 and FPR = 0, what is the AUC?

(g) For each situation, indicate whether P = precision or R = recall should be optimized:

i. _____ A bank is doing fraud detection where a fraudulent transaction ("positive") that is missed is expensive but a valid transaction labeled fraudulent is inexpensive.

ii. _____ A doctor is screening patients for a disease in which an ill patient ("positive") infects others and dies if the disease is not diagnosed.

iii. _____ A marketing campaign invests considerable expense in a prospective customer when it classifies that customer as likely to make a purchase ("positive").
6. Consider a one-vs.-rest SVM classifier trained on the following data depicted by circles, squares, and triangles:

(a) On the graph above, draw the three binary classifiers required by this method.

(b) How does this classifier classify the point indicated by “+”?

- circle
- square
- triangle

(c) Which category is ranked second by this classifier’s decision method for the “+”?

- circle
- square
- triangle
7. Here is a graph of the data set \( \{(x_i, y_i)\} = \{(x_i, y_i)\} = \{(1, 3), (2, 2), (4, 4)\} \) (here each \( x_i \) is a 1D \( x_i \)) along with corresponding Gaussian curves \( \{f_{\mu=x_i,\sigma=b} (x)\} \) made with bandwidth \( b = 0.25 \):

(a) Use kernel regression to estimate \( y = f(x) \) for \( x = 1 \).

Based on the plot, the estimate is \( \hat{y} \approx \) ____________.

(b) Estimate \( y = f(x) \) for \( x = 1.5 \).

Based on the plot, the estimate is \( \hat{y} \approx \) ____________.

(c) On the figure above, draw the estimated regression function over the interval \([0, 6]\).
8. The next two questions are about principal component analysis (PCA).

(a) Consider the following code and its output:

```python
rng = np.random.default_rng(seed=0)
(n_rows, n_cols) = (10, 4)
X = rng.normal(loc=0, scale=1, size=n_rows*n_cols).reshape((n_rows, n_cols))
pca = PCA(n_components=n_cols, random_state=0)
pca.fit(X=X)
with np.printoptions(precision=3):
    print(f'pca.components_=
{pca.components_}')
    print(f'pca.explained_variance_={pca.explained_variance_}')
    print(f'pca.explained_variance_ratio_={pca.explained_variance_ratio_}')
    print(f'pca.noise_variance_={pca.noise_variance_}')
    print(f'pca.mean_={pca.mean_}')
    print(f'pca.singular_values_={pca.singular_values_}')
```

Output:

```
pca.components_=
[[-0.219 -0.091 -0.752 -0.615]
 [ 0.854 0.439 -0.085 -0.265]
 [-0.41  0.882 -0.138  0.184]
 [-0.232  0.142  0.639 -0.72 ]]
```

What is the minimum number of principal components we must retain to account for 90% of the variability in the data? ___________

(b) Suppose PCA is run on the data in the plot. Draw arrows on the plot representing the first two principal components. (There is more than one correct answer.)