1. (5 points) What is the difference between data, information, and knowledge? Use specific examples in your answer.

**Solution:** Data can be numbers, labels, etc without any identifying information associated. An example is the number 5. This could be transformed into information by added information about the number, such as 5 ft. Knowledge is the ability to describe higher level concepts using information. An example could be that the average height of everyone in this room is 5 ft 7 inches.
2. (5 points) Describe the Knowledge Discovery in Databases (KDD) process. Specifically describe the steps of the KDD process and its goals. Your steps may be organized differently than the notes, but the order must be preserved.

**Solution:** I would accept a variety of answers to this question, but they must begin by talking about the pre-processing, normalization, scaling, and filtering of data before actually performing data mining. Then you must discuss how the process of data mining may be interactive, where we modify previous steps of the KDD process. Then it must finish with abstracting the results of data mining into knowledge representation.

3. (5 points) What is the difference between unsupervised and supervised learning?

**Solution:** Supervised learning involves a learning algorithm that uses labeled data to train, or if training is not necessary for the algorithm, it uses the labels in another manner. Unsupervised learning is a class of learning algorithms that does not use labeled data.

4. (5 points) What are the four scales of measurement? If you don’t remember the correct name, I’ll give you partial credit for clear examples.
Solution: Categorical (nominal), ordinal, interval, and ratio

5. (5 points) In clustering, do we want to maximize or minimize within cluster variation?

Solution: Minimize

6. (5 points) In clustering, do we want to maximize or minimize between cluster variation?

Solution: Maximize
7. (5 points) Compare and contrast exclusive clustering and overlapping clustering? General ideas are fine.

Solution: Exclusive clustering divides the samples into disjoint sets (clusters). There is no such restrictions for overlapping clustering. Samples may belong to more than one group.

8. (5 points) For the following data, draw the first and second principal component vectors.
9. (15 points) Cluster the smartphone dataset using k-means cluster (k = 2). The initial centroids are 10% and 20%.

**Solution:**
Cluster 1: US, Canada, France, Germany, Sweden, and UK
Cluster 2: Italy and Spain
10. (15 points) Using the naive Bayes classifier, compute the probability that house number 3 is acceptable. You must first construct the histograms necessary for this calculation (5 points). Then you must compute the probabilities (10 points, showing all of your work).

<table>
<thead>
<tr>
<th>House</th>
<th>Furniture/F</th>
<th># Rooms/R</th>
<th>Kitchen/K</th>
<th>Acceptable/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Not included</td>
<td>3</td>
<td>New</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>Included</td>
<td>3</td>
<td>Old</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>Not included</td>
<td>4</td>
<td>Old</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>Not included</td>
<td>3</td>
<td>Old</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>Included</td>
<td>4</td>
<td>Old</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Solution:** \[ P(A = Yes|\text{Not included, 4, Old}) = 1 \]
11. (5 points) Using the k-nearest-neighbors algorithms and \( k = 3 \), what is the class label of the ? mark.

**Solution:** Lemon