1. (a) Explain how the map-reduce framework can be used to join two tables \( R(A, B) \) and \( S(B, C) \). Here \( R(A, B) \) is a table with two attributes \( A \) and \( B \). Similarly \( S(B, C) \) is a table with two attributes \( B \) and \( C \). The tables \( R \) and \( S \) are joined on the attribute \( B \) [10 pts].

This is straight from the textbook.

**Map Function:** For each tuple \((a, b)\) of \( R \), produce the key-pair \((b, (R, a))\). For each tuple \((b, c)\) of \( S \), produce the key-value pair \((b, (S, c))\).

**Reduce Function:** Each key value \( b \) will be associated with a list of pairs that are either of the form \((R, a)\) or \((S, c)\). Construct all pairs consisting of one with the first component \( R \) and the other with the first component \( S \). For example \((R, a)\) and \((S, c)\) will result in \((a, b, c)\).
(b) Suppose we have four documents shown below

<table>
<thead>
<tr>
<th>d1</th>
<th>apple orange pair banana</th>
</tr>
</thead>
<tbody>
<tr>
<td>d2</td>
<td>apple pear melon orange</td>
</tr>
<tr>
<td>d3</td>
<td>pear orange peach</td>
</tr>
<tr>
<td>d4</td>
<td>cherry blueberry apple orange</td>
</tr>
</tbody>
</table>

Explain how the map-reduce framework can be used to count the number of time each word appears in the four documents. Use three mappers and two reducers. [10 pts]

Assign \( d_1 \) and \( d_2 \) to Map1; \( d_3 \) to Map2 and \( d_4 \) to Map3.

Map1 will produce the key-value pairs: \((\text{apple}, 1), (\text{orange}, 1), (\text{pair}, 1), (\text{banana}, 1), (\text{apple}, 1), (\text{pear}, 1), (\text{melon}, 1), (\text{orange}, 1)\).

Map2 will produce key-values \((\text{pear}, 1), (\text{orange}, 1), (\text{peach}, 1)\).

Map3 will produce key-values \((\text{cherry}, 1), (\text{blueberry}, 1), (\text{apple}, 1)(\text{orange}, 1)\).

Now we need to assign the pairs to reducers. The only constraint is that the same keys must end in one reducer.

Reducer1: \((\text{apple}, [1, 1]), (\text{banana}, 1), (\text{blueberry}, 1), (\text{cherry}, 1))\).

Reducer2: \((\text{orange}, [1, 1], ...\)

The reducers then output \((\text{apple}, 2), ..(\text{orange}, 2)\).
(c) Khadoop is a cloud-based service provider. Khadoop claims that their data center has one million computing nodes. Each of these nodes is built using precision hardware and the probability of a single node failing at a given time is \(10^{-6}\). Khadoop then goes on to claim that while they will use many mappers in a map-reduce task they will only use one, but very powerful reducer. The numbers of mappers to reducers (in their data center) follow a ratio 10:1. What is the probability that Khadoop’s map-reduce task will fail at a given time. Assume all nodes are continuously running map-reduce tasks. [10 pts]

Answer: Let \(p\) be the probability of failure of a node. Then \((1 - p)\) is the probability that the node will not fail. \((1 - p)^n\) is the probability that \(n\) will not fail. Thus the probability that at least one node will fail is

\[
(1 - (1 - p)^n)
\]

Now, \(p = 10^{-6}\) and \(n = 100000\).
Thus

\[
(1 - (1 - p)^n) = 0.0952
\]
2. (a) Suppose the probabilities that the ASX stock increases today is 0.60, that it increases tomorrow is 0.60, and that it increases both days is 0.30. What is the probability that it does not increase on either day. [10 pts]

Let $A$ be the event that it increases today. Let $B$ be the event that it increases tomorrow.

$P(A) = 0.6 \; ; \; P(B) = 0.6 \; ; \; P(AB) = 0.3$

Thus the probability it increases on at least one day is

$P(A \cup B) = P(A) + P(B) - P(AB)$.

Thus the probability that it does not increase on either day is

$1 - P(A \cup B) = 1 - (0.6 + 0.6 - 0.3)) = 0.1$

(b) Suppose a credit-card company wants to create a model to predict whether a customer will default. After some investigation they decided that they will only use two attributes: Age and Income. They also decided to use the Naive Bayes classifier to build the model. Discuss the pros and cons of this decision. You can create a small data set to justify your answer. [10 pts]

They are different ways of answering the question. But the key insight is that Age and Income are likely to be correlated.

Consider the following data set

<table>
<thead>
<tr>
<th>Age</th>
<th>Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>Young</td>
<td>Small</td>
</tr>
<tr>
<td>Young</td>
<td>Medium</td>
</tr>
<tr>
<td>MiddleAge</td>
<td>Medium</td>
</tr>
<tr>
<td>MiddleAge</td>
<td>Large</td>
</tr>
<tr>
<td>Old</td>
<td>Medium</td>
</tr>
<tr>
<td>Old</td>
<td>Small</td>
</tr>
</tbody>
</table>

Then the Naive Bayes Classifier will assume that

$P(Age = Young, Income = Small|CC = D)) = P(Age = Young|CC = D)$
Now the advantage of this decomposition is that we have more data to estimate $P(Age = \text{Young}|CC = D)$ compared to $P(Age = \text{Young}, Income = \text{Small}|CC = D)$. Thus our estimates are more reliable. The disadvantage is that we will end up giving less weight to the feature combination $(Age = \text{Young}, Income = \text{Small})$ thus our classifier will be less accurate.
(c) A bank in Australia wants to buy a software to monitor Internet traffic that flows through its network. Two software vendors A and B are trying to convince the bank to buy their software. Vendor A claims that their software is extremely accurate as it scores every incoming packet and tests have shown that the probability of correctly identifying a malicious (bad) packet is 0.99. Similarly the probability of correctly identifying a non-malicious packet is 0.98. Vendor B on the other hand claims that the probability of identifying a malicious packet (bad) packet is 0.98 and correctly identifying a non-malicious packet is 0.99. The bank has hired you as a consultant to help them decide whether to go with vendor A or B. While planning for the meeting you looked up the web site of a global network provider where it was claimed that the one percent of all network traffic is now malicious. What different scenarios will you lay out for the bank management to help them arrive at a decision. Should the bank be seeking more information from the vendors? [10 pts]

Let $X$ be the event the packet is malicious.
Let $t_A$ be event that vendor A identifies a packet as malicious.
Let $t_B$ be the event that vendor B identifies a packet as malicious.

Then we are given:
$P(X) = 0.01$
$P(t_A|X) = 0.99$
$P(t_A^c|X^c) = 0.98$

$P(t_B|X) = 0.98$
$P(t_B^c|X^c) = 0.99$

What we want is

$$P(X|t_A) = \frac{P(t_A|X)P(X)}{P(t_A|X)P(X) + P(t_A^c|X^c)P(X^c)}$$

$$= \frac{0.99 \times 0.01}{0.99 \times 0.01 + 0.98 \times 0.91}$$

$$= 0.333$$

Similarly

$$P(X|t_B) = \frac{P(t_B|X)P(X)}{P(t_B|X)P(X) + P(t_B^c|X^c)P(X^c)}$$

$$= \frac{0.98 \times 0.01}{0.98 \times 0.01 + 0.99 \times 0.01}$$

$$= 0.4975$$
The key thing to ask a vendor is whether they give more importance to false positives or false negatives.
3. Australia’s per-capita income is now one of the highest in the world. This has prompted many global retailers to open up shop in Australia. One such global retailer, based in Hong Kong, hired an analytics company XYZ, based in Sydney, to commission a survey. XYZ decides to survey 10,000 people between the ages of 20 and 60 and then put them into segments and then estimate the buying capacity of each segment. What are some of the attributes that XYZ should measure in the survey? How would you use the k-means algorithm to come up with a quantitative answer to characterize the buying power of these segments. What are some of the limitations of using k-means and what strategies can you employ to overcome these limitations. Give the complete pseudo-code of the k-means algorithm. [20 pts]

One part is open ended. The other is just writing down the k-means algorithm.

Some good attributes to use are Age, Income, SurplusIncome, Location (Urban, Rural) etc. Now if we apply a clustering algorithm with k clusters then for each cluster we can compute the average income. We can also try and understand (from the cluster) the surplus income and use that to quantify the buying power by counting the number of people in a cluster, dividing by 10K and then multiplying the population of Australia.
4. Two algorithms $A$ and $B$ produce the following scores on a binary classification task. Plot the ROC curves of $A$ and $B$ on the same graph. Which is more accurate algorithm. [20 pts]

<table>
<thead>
<tr>
<th></th>
<th>0.9</th>
<th>0.8</th>
<th>0.75</th>
<th>0.7</th>
<th>0.6</th>
<th>0.3</th>
<th>0.2</th>
<th>0.15</th>
<th>0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Labels</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>A(0.4)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>A(0.6)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>B(0.4)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>B(0.6)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
</tbody>
</table>

True Positive Rate for algorithm A at 0.4 = 1
False Positive Rate for algorithm A at 0.4 = 0.2

True Positive Rate for algorithm A at 0.6 = 1
False Positive Rate for algorithm A at 0.6 = 0.2

True Positive Rate for algorithm B at 0.4 = 1
False Positive Rate for algorithm B at 0.4 = 0.4

True Positive Rate for algorithm B at 0.6 = 1
False Positive Rate for algorithm B at 0.6 = 0.2

Now even though the ROC curve is identical, algorithm $A$ is better as $B$ has a higher false positive rate at 0.4.