School of Computing and Information Systems
The University of Melbourne
COMP30027 MACHINE LEARNING (Semester 1, 2019)

Tutorial sample solutions: Week 8

1. Revise the difference between **supervised** and **unsupervised** machine learning.

   Then, consider the following dataset:

id	*apple*	*ibm*	*lemon*	*sun*	LABEL
A	4	0	1	1	FRUIT
B	5	0	5	2	FRUIT
C	2	5	0	0	COMP
D	1	2	1	7	COMP
E	2	0	3	1	?
F	1	0	1	0	?

2. Treat the problem as an unsupervised machine learning problem (excluding the *id* and LABEL attributes), and calculate the clusters according to (hard) k-**means** with $k = 2$, using the Manhattan distance:

   (a) Using seeds A and D.

   - This is an unsupervised problem, so we ignore (or don't have access to) the LABEL attribute. (We're going to ignore *id* as well, because it obviously isn't a meaningful point of comparison.)
   - We begin by setting the initial centroids for our two clusters, let's say cluster 1 has centroid $C_1 = \langle 4, 0, 1, 1 \rangle$ and cluster 2 $C_2 = \langle 1, 2, 1, 7 \rangle$.
   - We now calculate the distance for each instance ("training" and "test" are equivalent in this context) to the centroids of each cluster:

$$
\begin{aligned}
d(A, C_1) &= |\,4-4\,| + |\,0-0\,| + |\,1-1\,| + |\,1-1\,| \\
&= 0 \\
d(A, C_2) &= |\,4-1\,| + |\,0-2\,| + |\,1-1\,| + |\,1-7\,| \\
&= 11 \\
d(B, C_1) &= |\,5-4\,| + |\,0-0\,| + |\,5-1\,| + |\,2-1\,| \\
&= 6 \\
d(B, C_2) &= |\,5-1\,| + |\,0-2\,| + |\,5-1\,| + |\,2-7\,| \\
&= 15
\end{aligned}
$$

$$
\begin{aligned}
d(C, C_1) &= \mid 2-4 \mid + \mid 5-0 \mid + \mid 0-1 \mid + \mid 0-1 \mid \\
&= 9 \\
d(C, C_2) &= \mid 2-1 \mid + \mid 5-2 \mid + \mid 0-1 \mid + \mid 0-7 \mid \\
&= 12 \\
d(D, C_1) &= \mid 1-4 \mid + \mid 2-0 \mid + \mid 1-1 \mid + \mid 7-1 \mid \\
&= 11 \\
d(D, C_2) &= \mid 1-1 \mid + \mid 2-2 \mid + \mid 1-1 \mid + \mid 7-7 \mid \\
&= 0 \\
d(E, C_1) &= \mid 2-4 \mid + \mid 0-0 \mid + \mid 3-1 \mid + \mid 1-1 \mid \\
&= 4 \\
d(E, C_2) &= \mid 2-1 \mid + \mid 0-2 \mid + \mid 3-1 \mid + \mid 1-7 \mid \\
&= 11 \\
d(F, C_1) &= \mid 1-4 \mid + \mid 0-0 \mid + \mid 1-1 \mid + \mid 0-1 \mid \\
&= 4 \\
d(F, C_2) &= \mid 1-1 \mid + \mid 0-2 \mid + \mid 1-1 \mid + \mid 0-7 \mid \\
&= 9
\end{aligned}
$$

- We now assign each instance to the cluster with the smallest (Manhattan) distance to the cluster's centroid: for `A`, this is C_1 because $0 < 11$, for `B`, this is C_1 because $6 < 15$, and so on. It turns out that `A`, `B`, `C`, `E`, and `F` all get assigned to cluster 1, and `D` is assigned to cluster 2.
- We now update the centroids of the clusters, by calculating the arithmetic mean of the attribute values for the instances in each cluster. For cluster 1, this is:

$$
\begin{aligned}
C_1 &= \langle \frac{4+5+2+2+1}{5}, \frac{0+0+5+0+0}{5}, \frac{1+5+0+3+1}{5}, \frac{1+2+0+1+0}{5} \rangle \\
&= \langle 2.8, 1, 2, 0.8 \rangle
\end{aligned}
$$

- For cluster 2, we're just taking the average of a single value, so obviously the centroid is just $\langle 1, 2, 1, 7 \rangle$.
- Now, we re-calcuate the distances of each instance to each centroid:

$$
\begin{aligned}
d(A, C_1) &= \mid 4-2.8 \mid + \mid 0-1 \mid + \mid 1-2 \mid + \mid 1-0.8 \mid \\
&= 3.4 \\
d(B, C_1) &= \mid 5-2.8 \mid + \mid 0-1 \mid + \mid 5-2 \mid + \mid 2-0.8 \mid \\
&= 7.4 \\
d(C, C_1) &= \mid 2-2.8 \mid + \mid 5-1 \mid + \mid 0-2 \mid + \mid 0-0.8 \mid \\
&= 7.6 \\
d(D, C_1) &= \mid 1-2.8 \mid + \mid 2-1 \mid + \mid 1-2 \mid + \mid 7-0.8 \mid \\
&= 10 \\
d(E, C_1) &= \mid 2-2.8 \mid + \mid 0-1 \mid + \mid 3-2 \mid + \mid 1-0.8 \mid \\
&= 3 \\
d(F, C_1) &= \mid 1-2.8 \mid + \mid 0-1 \mid + \mid 1-2 \mid + \mid 0-0.8 \mid \\
&= 4.6
\end{aligned}
$$

- (Obviously, the distance of each instance to cluster 2 hasn't changed, because the value of the centroid is the same as the previous iteration.)
- Now, we re-assign instances to clusters, according to the smaller (Manhattan) distance: `A` gets assigned to cluster 1 (because $3.4 < 11$), `B` gets assigned to cluster 1 (because $7.4 < 15$), and so on. In all, `A`, `B`, `C`, `E`, and `F` get assigned to cluster 1, and `D` to cluster 2.

- At this point, we observe that the assignments of instances to clusters is the same as the previous iteration, so we stop. (The newly-calculated centriods are going to be the same, so the algorithm has reached equilibrium.)
- The final assignment of instances to clusters here is: cluster 1 {`A`,`B`,`C`,`E`,`F`} and cluster 2 {`D`}.

(b) Using seeds `A` and `F`.

- This time, the initial centroids are $C_1 = \langle 4, 0, 1, 1\rangle$ and $C_2 = \langle 1, 0, 1, 0\rangle$.
- We calculate the (Manhattan) distances of each instance to each centroid:

$$\begin{aligned}
d(A, C_1) &= \mid 4 - 4 \mid + \mid 0 - 0 \mid + \mid 1 - 1 \mid + \mid 1 - 1 \mid \\
&= 0 \\
d(A, C_2) &= \mid 4 - 1 \mid + \mid 0 - 0 \mid + \mid 1 - 1 \mid + \mid 1 - 0 \mid \\
&= 4 \\
d(B, C_1) &= \mid 5 - 4 \mid + \mid 0 - 0 \mid + \mid 5 - 1 \mid + \mid 2 - 1 \mid \\
&= 6 \\
d(B, C_2) &= \mid 5 - 1 \mid + \mid 0 - 0 \mid + \mid 5 - 1 \mid + \mid 2 - 0 \mid \\
&= 10 \\
d(C, C_1) &= \mid 2 - 4 \mid + \mid 5 - 0 \mid + \mid 0 - 1 \mid + \mid 0 - 1 \mid \\
&= 9 \\
d(C, C_2) &= \mid 2 - 1 \mid + \mid 5 - 0 \mid + \mid 0 - 1 \mid + \mid 0 - 0 \mid \\
&= 7 \\
d(D, C_1) &= \mid 1 - 4 \mid + \mid 2 - 0 \mid + \mid 1 - 1 \mid + \mid 7 - 1 \mid \\
&= 11 \\
d(D, C_2) &= \mid 1 - 1 \mid + \mid 2 - 0 \mid + \mid 1 - 1 \mid + \mid 7 - 0 \mid \\
&= 9 \\
d(E, C_1) &= \mid 2 - 4 \mid + \mid 0 - 0 \mid + \mid 3 - 1 \mid + \mid 1 - 1 \mid \\
&= 4 \\
d(E, C_2) &= \mid 2 - 1 \mid + \mid 0 - 0 \mid + \mid 3 - 1 \mid + \mid 1 - 0 \mid \\
&= 4 \\
d(F, C_1) &= \mid 1 - 4 \mid + \mid 0 - 0 \mid + \mid 1 - 1 \mid + \mid 0 - 1 \mid \\
&= 4 \\
d(F, C_2) &= \mid 1 - 1 \mid + \mid 0 - 0 \mid + \mid 1 - 1 \mid + \mid 0 - 0 \mid \\
&= 0
\end{aligned}$$

- Here, `A` is closer to cluster 1's centroid, `B` to cluster 1, `C` to cluster 2, `D` to cluster 2, `F` to cluster 2, and for `E` we have a tie.
- Let's say we randomly break the tie for instance `E` by assigning it to cluster 2. (We'll see what would have happened if we'd assigned `E` to cluster 1 below.) So, cluster 1 is {`A`,`B`} and cluster 2 is {`C`,`D`,`E`,`F`}. We re-calculate the centroids:

$$\begin{aligned}
C_1 &= \langle \frac{4+5}{2}, \frac{0+0}{2}, \frac{1+5}{2}, \frac{1+2}{2} \rangle \\
&= \langle 4.5, 0, 3, 1.5 \rangle \\
C_2 &= \langle \frac{2+1+2+1}{4}, \frac{5+2+0+0}{4}, \frac{0+1+3+1}{4}, \frac{0+7+1+0}{4} \rangle \\
&= \langle 1.5, 1.75, 1.25, 2 \rangle
\end{aligned}$$

- Now, let's re-calculate the distances according to these new centroids:

$$\begin{aligned}
d(A,C_1) &= \mid 4-4.5 \mid + \mid 0-0 \mid + \mid 1-3 \mid + \mid 1-1.5 \mid \\
&= 3 \\
d(A,C_2) &= \mid 4-1.5 \mid + \mid 0-1.75 \mid + \mid 1-1.25 \mid + \mid 1-2 \mid \\
&= 5.5 \\
d(B,C_1) &= \mid 5-4.5 \mid + \mid 0-0 \mid + \mid 5-3 \mid + \mid 2-1.5 \mid \\
&= 3 \\
d(B,C_2) &= \mid 5-1.5 \mid + \mid 0-1.75 \mid + \mid 5-1.25 \mid + \mid 2-2 \mid \\
&= 9 \\
d(C,C_1) &= \mid 2-4.5 \mid + \mid 5-0 \mid + \mid 0-3 \mid + \mid 0-1.5 \mid \\
&= 12 \\
d(C,C_2) &= \mid 2-1.5 \mid + \mid 5-1.75 \mid + \mid 0-1.25 \mid + \mid 0-2 \mid \\
&= 7 \\
d(D,C_1) &= \mid 1-4.5 \mid + \mid 2-0 \mid + \mid 1-3 \mid + \mid 7-1.5 \mid \\
&= 13 \\
d(D,C_2) &= \mid 1-1.5 \mid + \mid 2-1.75 \mid + \mid 1-1.25 \mid + \mid 7-2 \mid \\
&= 6 \\
d(E,C_1) &= \mid 2-4.5 \mid + \mid 0-0 \mid + \mid 3-3 \mid + \mid 1-1.5 \mid \\
&= 3 \\
d(E,C_2) &= \mid 2-1.5 \mid + \mid 0-1.75 \mid + \mid 3-1.25 \mid + \mid 1-2 \mid \\
&= 5 \\
d(F,C_1) &= \mid 1-4.5 \mid + \mid 0-0 \mid + \mid 1-3 \mid + \mid 0-1.5 \mid \\
&= 7 \\
d(F,C_2) &= \mid 1-1.5 \mid + \mid 0-1.75 \mid + \mid 1-1.25 \mid + \mid 0-2 \mid \\
&= 4.5
\end{aligned}$$

- What are the assignments of instances to clusters now? Cluster 1 {`A,B,E`} and cluster 2 {`C,D,F`}. (Note that we're at the same place now that we would have been if we'd randomly broke the tie for instance `E` to cluster 1 earlier.)
- We calculate the new centroids based on these instances:

$$\begin{aligned}
C_1 &= \langle \frac{4+5+2}{3}, \frac{0+0+0}{3}, \frac{1+5+3}{3}, \frac{1+2+1}{3} \rangle \\
&\approx \langle 3.67, 0, 3, 1.33 \rangle \\
C_2 &= \langle \frac{2+1+1}{3}, \frac{5+2+0}{3}, \frac{0+1+1}{3}, \frac{0+7+0}{3} \rangle \\
&\approx \langle 1.33, 2.33, 0.67, 2.33 \rangle
\end{aligned}$$

- We recalculate the distances according to these new centroids:

$$\begin{aligned}
d(A,C_1) &\approx \mid 4-3.67 \mid + \mid 0-0 \mid + \mid 1-3 \mid + \mid 1-1.33 \mid \\
&\approx 2.67 \\
d(A,C_2) &\approx \mid 4-1.33 \mid + \mid 0-2.33 \mid + \mid 1-0.67 \mid + \mid 1-2.33 \mid \\
&\approx 6.67 \\
d(B,C_1) &\approx \mid 5-3.67 \mid + \mid 0-0 \mid + \mid 5-3 \mid + \mid 2-1.33 \mid \\
&\approx 4 \\
d(B,C_2) &\approx \mid 5-1.33 \mid + \mid 0-2.33 \mid + \mid 5-0.67 \mid + \mid 2-2.33 \mid \\
&\approx 10.67
\end{aligned}$$

$$
\begin{aligned}
d(C, C_1) &\approx \mid 2 - 3.67 \mid + \mid 5 - 0 \mid + \mid 0 - 3 \mid + \mid 0 - 1.33 \mid \\
&\approx 11 \\
d(C, C_2) &\approx \mid 2 - 1.33 \mid + \mid 5 - 2.33 \mid + \mid 0 - 0.67 \mid + \mid 0 - 2.33 \mid \\
&\approx 6.33 \\
d(D, C_1) &\approx \mid 1 - 3.67 \mid + \mid 2 - 0 \mid + \mid 1 - 3 \mid + \mid 7 - 1.33 \mid \\
&\approx 12.33 \\
d(D, C_2) &\approx \mid 1 - 1.33 \mid + \mid 2 - 2.33 \mid + \mid 1 - 0.67 \mid + \mid 7 - 2.33 \mid \\
&\approx 5.67 \\
d(E, C_1) &\approx \mid 2 - 3.67 \mid + \mid 0 - 0 \mid + \mid 3 - 3 \mid + \mid 1 - 1.33 \mid \\
&\approx 2 \\
d(E, C_2) &\approx \mid 2 - 1.33 \mid + \mid 0 - 2.33 \mid + \mid 3 - 0.67 \mid + \mid 1 - 2.33 \mid \\
&\approx 6.67 \\
d(F, C_1) &\approx \mid 1 - 3.67 \mid + \mid 0 - 0 \mid + \mid 1 - 3 \mid + \mid 0 - 1.33 \mid \\
&\approx 6 \\
d(F, C_2) &\approx \mid 1 - 1.33 \mid + \mid 0 - 2.33 \mid + \mid 1 - 0.67 \mid + \mid 0 - 2.33 \mid \\
&\approx 5.33
\end{aligned}
$$

- The new assignments of instances to clusters are cluster 1 {`A`, `B`, `E`} and cluster 2 {`C`, `D`, `F`}. This is the same as the last iteration, so we stop (and this is the final assignment of instances to clusters).

3. Repeat the previous question using "soft" k-means, and a "stiffness" $\beta = 1$.

   - Let's use initial centroids as A and F, like the previous question. It's better to use the Euclidean distance here, but I'll use Manhattan to make our lives a little easier.
   - We'll pick up from having calculated the distances of each point to the two initial centroids. In "hard" k-means, we would assign each instance to whichever cluster is closer, but in "soft" k-means, we probabilistically assign each point according to the "softmax" function:

$$
z_{ij} = \frac{e^{-\beta d(i,j)}}{\sum_i e^{-\beta d(i,j)}}
$$

   - For each point, we are essentially normalising, but rather than by the sum of the raw distances, we raise each negated distance to the power of e — this effectively handles the fact that smaller distances mean more similar instances, as well as making a transformation to account for differences in the distances.
   - For instance A (conveniently also the centroid of one cluster), the distances to the two clusters are 0 and 4. The probabilistic assignment is consequently:

$$
\begin{aligned}
z_{1A} &= \frac{e^{-0}}{e^{-0} + e^{-4}} \approx 0.982 \\
z_{2A} &= \frac{e^{-4}}{e^{-0} + e^{-4}} \approx 0.018
\end{aligned}
$$

   - Since the first cluster is much closer, this instance amost entirely placed in that cluster — however, unlike with "hard" k-means, it is slightly in the second cluster, too.

- And so on for the other instances:

$$
\begin{aligned}
z_{1B} &= \frac{e^{-6}}{e^{-6}+e^{-10}} \approx 0.982 \\
z_{2B} &= \frac{e^{-10}}{e^{-6}+e^{-10}} \approx 0.018 \\
z_{1C} &= \frac{e^{-9}}{e^{-9}+e^{-7}} \approx 0.119 \\
z_{2C} &= \frac{e^{-7}}{e^{-9}+e^{-7}} \approx 0.881 \\
z_{1D} &= \frac{e^{-11}}{e^{-11}+e^{-9}} \approx 0.119 \\
z_{2D} &= \frac{e^{-9}}{e^{-11}+e^{-9}} \approx 0.881 \\
z_{1E} &= \frac{e^{-4}}{e^{-4}+e^{-4}} \approx 0.5 \\
z_{2E} &= \frac{e^{-4}}{e^{-4}+e^{-4}} \approx 0.5 \\
z_{1F} &= \frac{e^{-4}}{e^{-4}+e^{-0}} \approx 0.018 \\
z_{2F} &= \frac{e^{-0}}{e^{-4}+e^{-0}} \approx 0.982
\end{aligned}
$$

- Instance E is still tied between the two clusters, but that doesn't create any issues here.
- If you look closely, you can see that the probabilistic assignment doesn't depend on the distance exactly, but rather, the difference between the distances. For example, instances C and D get soft-assigned exactly the same way, even though C is two units closer to both clusters.
- We now update the centroids, just like with "hard" k-means, except this time it is a weighted average:

$$
\begin{aligned}
C_1^{(1)} &: \frac{0.982A + 0.982B + 0.119C + 0.119D + 0.5E + 0.018F}{0.982 + 0.982 + 0.119 + 0.119 + 0.5 + 0.018} \\
&= \frac{1}{2.72}[(0.982)\langle 4,0,1,1\rangle + (0.982)\langle 5,0,5,2\rangle + (0.119)\langle 2,5,0,0\rangle \ldots] \\
&= \frac{1}{2.72}[\langle (0.982)(4) + (0.982)(5) + (0.119)(2) + \ldots, (0.982)(0) + \ldots\rangle] \\
&= \frac{1}{2.72}[\langle 10.21, 0.833, 7.53, 4.28\rangle] \\
&\approx \langle 3.75, 0.307, 2.77, 1.57\rangle \\
C_2^{(1)} &: \frac{0.018A + 0.018B + 0.881C + 0.881D + 0.5E + 0.982F}{0.018 + 0.018 + 0.881 + 0.881 + 0.5 + 0.982} \\
&\approx \langle 1.46, 1.88, 1.06, 2.05\rangle
\end{aligned}
$$

- With our new centroids, we are set to iterate. I'll summarise the distances and corresponding z values in a table, for convenient reading:
- We can see that there are some instances where we have become more confident — like B and D — and others where we appear to be changing our mind — like E and F.

	$d(1,j)$	$d(2,j)$	z_1j	z_2j
A	2.89	5.53	0.933	0.067
B	4.21	9.41	0.995	0.005
C	10.79	6.77	0.018	0.982
D	11.64	5.59	0.002	0.998
E	2.87	5.41	0.927	0.073
F	6.40	4.45	0.124	0.876

- We would re-calculate the centroids now as follows:

$$\begin{aligned} C_1^{(2)} &: \frac{0.933A + 0.995B + 0.018C + 0.002D + 0.927E + 0.124F}{0.933 + 0.995 + 0.018 + 0.002 + 0.927 + 0.124} \\ &\approx \langle 3.58, 0.031, 2.94, 1.29 \rangle \\ C_2^{(2)} &: \frac{0.067A + 0.005B + 0.982C + 0.998D + 0.073E + 0.876F}{0.067 + 0.005 + 0.982 + 0.998 + 0.073 + 0.876} \\ &\approx \langle 1.43, 2.30, 0.729, 2.38 \rangle \end{aligned}$$

- The centroids have only moved a small way (a couple of tenths in most cases), but that causes our predictions to change further:

	$d(1,j)$	$d(2,j)$	z_1j	z_2j
A	2.68	6.52	0.979	0.021
B	4.23	10.52	0.998	0.002
C	10.77	6.38	0.012	0.988
D	12.19	5.62	0.001	0.999
E	1.96	6.52	0.990	0.010
F	5.83	5.38	0.387	0.613

- Already, we can see that we are quite certain about most of the instances, and in the next couple of iterations, instance F will get "pulled into" Cluster 1, at which point the method will gradually converge.
- (If you're curious, the two cluster centroids eventually become approximately $\langle 3, 0, 2.5, 1 \rangle$ and $\langle 1.5, 3.4, 0.5, 3.5 \rangle$ — after about 6 iterations — you might like to compare these to the corresponding "hard" centroids.)

4. What is logic behind the **EM algorithm**, when used for clustering?

   - Basically, that we start with a random (or uniform) guess, and then we progressively improve our guess by evaluating the expected likelihood on the given data.
   - Another way of looking at it: our initial estimate probabilistically defines some labels for the training data; we can then use the counts over these labels to re-estimate the corresponding elements of the model (typically probabilities).

   (a) Explain the significance of the "E" step, and the "M" step.
      - Expectation: assign weighted labels to the training data, and use these to calculate the (log-)likelihood function
      - Maximisation: re-estimate the parameter based on these labels

5. What is **semi–supervised learning**, and when is it desirable?

   - We have a small number of labelled instances, and a large number of unlabelled instances.
   - Typically, this means that we don't have enough data to train a reliable classifier (purely supervised), but we can potentially leverage the labelled instances to build a better classifier than a purely unsupervised method might come up with.

(a) What is **self training**?

- Self training is a method of using a learner to build a training data set as follows:
  - Train the learner on the currently–labelled instances
  - Use the learner to predict the labels of the unlabelled instances
  - Where the learner is very confident, add newly–labelled instances to the training set
  - Repeat until all instances are labelled, or no new instances can be labelled confidently.

(b) What is the logic behind **active learning**, and what are some methods to choose instances for the **oracle**?

- In active learning, the learner is allowed to choose a small number of instances to be labelled (by a human judge).
- The idea here is two–fold: many instances are easy to classify; and a small number of instances are difficult to classify, **but** would be easy(ier) to classify with more training data.
- In some cases, the learner generates its own difficult instances; in others, the instances are selected as the ones which are most difficult to classify in a fixed, unlabelled set.