School of Computing and Information Systems
The University of Melbourne
COMP90049 Knowledge Technologies (Semester 2, 2017)
Workshop exercises: Week 7

1. What is data mining/machine learning? What makes this a knowledge task?
   - Data mining: extracting implicit, previously unknown, potentially useful information from data
   - Machine learning: algorithms for acquiring structual descriptions from examples (special case of above?)
   - Knowledge task: the information/descriptions we produce are unknown and useful to humans

2. What is the difference between supervised and unsupervised machine learning? Give examples of some supervised and unsupervised techniques.
   - Generally speaking, supervised techniques in machine learning start from exemplars — labelled with classes — in a set of training data, and use these to classify unknown instances in a set of test data.
   - Unsupervised methods are not based on a set of labelled training data: they are often broken down into **weakly** unsupervised methods, where the class set is known, but the system does not have access to labelled training data; and **strongly** unsupervised methods, where even the class set is unknown.
   - For example, Naive Bayes, Support Vector Machines, Decision Trees, and k-Nearest Neighbour are all examples of supervised systems.
   - Clustering (e.g. k-means, Expectation Maximisation) is an example of an unsupervised methodology.

3. In the context of (supervised) machine learning:
   (a) What is an instance?
      - An instance is a single exemplar from the data, consisting of a bundle of (possibly unknown) attribute values (feature values) and a class value, mapping on to the concept that we wish to predict.

   (b) What is an attribute? What different kinds of attribute are there?
      - An attribute is a single measurement of some aspect of an instance, for example, the frequency of some event related to this instance, or the label of some meaningful category.
      - Attributes are usually classified as either nominal, ordinal, or continuous.

   (c) What is a class?
      - A class is the thing (usually attribute) we want to learn. It may be nominal ("classification") or continuous ("regression").

   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	?

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

   (a) Starting with 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:

$$\begin{aligned}
d(A, C_1) &= \mid 4-4 \mid + \mid 0-0 \mid + \mid 1-1 \mid + \mid 1-1 \mid \\
&= 0 \\
d(A, C_2) &= \mid 4-1 \mid + \mid 0-2 \mid + \mid 1-1 \mid + \mid 1-7 \mid \\
&= 11 \\
d(B, C_1) &= \mid 5-4 \mid + \mid 0-0 \mid + \mid 5-1 \mid + \mid 2-1 \mid \\
&= 6 \\
d(B, C_2) &= \mid 5-1 \mid + \mid 0-2 \mid + \mid 5-1 \mid + \mid 2-7 \mid \\
&= 15 \\
d(C, C_1) &= \mid 2-4 \mid + \mid 5-0 \mid + \mid 0-1 \mid + \mid 0-1 \mid \\
&= 9 \\
d(C, C_2) &= \mid 2-1 \mid + \mid 5-2 \mid + \mid 0-1 \mid + \mid 0-7 \mid \\
&= 12 \\
d(D, C_1) &= \mid 1-4 \mid + \mid 2-0 \mid + \mid 1-1 \mid + \mid 7-1 \mid \\
&= 11 \\
d(D, C_2) &= \mid 1-1 \mid + \mid 2-2 \mid + \mid 1-1 \mid + \mid 7-7 \mid \\
&= 0 \\
d(E, C_1) &= \mid 2-4 \mid + \mid 0-0 \mid + \mid 3-1 \mid + \mid 1-1 \mid \\
&= 4 \\
d(E, C_2) &= \mid 2-1 \mid + \mid 0-2 \mid + \mid 3-1 \mid + \mid 1-7 \mid \\
&= 11 \\
d(F, C_1) &= \mid 1-4 \mid + \mid 0-0 \mid + \mid 1-1 \mid + \mid 0-1 \mid \\
&= 4 \\
d(F, C_2) &= \mid 1-1 \mid + \mid 0-2 \mid + \mid 1-1 \mid + \mid 0-7 \mid \\
&= 9
\end{aligned}$$

   - 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) Starting with 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:

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

$$
\begin{aligned}
d(E, C_1) &= |\, 2-4 \,| + |\, 0-0 \,| + |\, 3-1 \,| + |\, 1-1 \,| \\
&= 4 \\
d(E, C_2) &= |\, 2-1 \,| + |\, 0-0 \,| + |\, 3-1 \,| + |\, 1-0 \,| \\
&= 4 \\
d(F, C_1) &= |\, 1-4 \,| + |\, 0-0 \,| + |\, 1-1 \,| + |\, 0-1 \,| \\
&= 4 \\
d(F, C_2) &= |\, 1-1 \,| + |\, 0-0 \,| + |\, 1-1 \,| + |\, 0-0 \,| \\
&= 0
\end{aligned}
$$

- 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:

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

- 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:

$$\begin{aligned}
d(A, C_1) &\approx \mid 4 - 3.67 \mid + \mid 0 - 0 \mid + \mid 1 - 3 \mid + \mid 1 - 1.33 \mid \\
&\approx 2.67 \\
d(A, C_2) &\approx \mid 4 - 1.33 \mid + \mid 0 - 2.33 \mid + \mid 1 - 0.67 \mid + \mid 1 - 2.33 \mid \\
&\approx 6.67 \\
d(B, C_1) &\approx \mid 5 - 3.67 \mid + \mid 0 - 0 \mid + \mid 5 - 3 \mid + \mid 2 - 1.33 \mid \\
&\approx 4 \\
d(B, C_2) &\approx \mid 5 - 1.33 \mid + \mid 0 - 2.33 \mid + \mid 5 - 0.67 \mid + \mid 2 - 2.33 \mid \\
&\approx 10.67 \\
d(C, C_1) &\approx \mid 2 - 3.67 \mid + \mid 5 - 0 \mid + \mid 0 - 3 \mid + \mid 0 - 1.33 \mid \\
&\approx 11 \\
d(C, C_2) &\approx \mid 2 - 1.33 \mid + \mid 5 - 2.33 \mid + \mid 0 - 0.67 \mid + \mid 0 - 2.33 \mid \\
&\approx 6.33 \\
d(D, C_1) &\approx \mid 1 - 3.67 \mid + \mid 2 - 0 \mid + \mid 1 - 3 \mid + \mid 7 - 1.33 \mid \\
&\approx 12.33 \\
d(D, C_2) &\approx \mid 1 - 1.33 \mid + \mid 2 - 2.33 \mid + \mid 1 - 0.67 \mid + \mid 7 - 2.33 \mid \\
&\approx 5.67 \\
d(E, C_1) &\approx \mid 2 - 3.67 \mid + \mid 0 - 0 \mid + \mid 3 - 3 \mid + \mid 1 - 1.33 \mid \\
&\approx 2 \\
d(E, C_2) &\approx \mid 2 - 1.33 \mid + \mid 0 - 2.33 \mid + \mid 3 - 0.67 \mid + \mid 1 - 2.33 \mid \\
&\approx 6.67 \\
d(F, C_1) &\approx \mid 1 - 3.67 \mid + \mid 0 - 0 \mid + \mid 1 - 3 \mid + \mid 0 - 1.33 \mid \\
&\approx 6 \\
d(F, C_2) &\approx \mid 1 - 1.33 \mid + \mid 0 - 2.33 \mid + \mid 1 - 0.67 \mid + \mid 0 - 2.33 \mid \\
&\approx 5.33
\end{aligned}$$

- 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).

5. Perform **agglomerative clustering** of the above dataset (excluding the *id* and LABEL attributes), using the Euclidean distance and calculating the **group average** as the cluster centroid. Do you expect to observe a different dendrogram if we were instead using the cosine similarity?

   - We begin by finding the pairwise similarities — or distances, in this case, between each instance. I'm going to skip the Euclidean distance calculations (you can work through them as an exercise) and go straight to the proximity matrix:
   - We can immediately observe (without simplifying the square roots) that the most similar instances (with the smallest distance) are E and F.

	A	B	C	D	E	F
A	-	$\sqrt{18}$	$\sqrt{31}$	$\sqrt{49}$	$\sqrt{8}$	$\sqrt{10}$
B	$\sqrt{18}$	-	$\sqrt{63}$	$\sqrt{61}$	$\sqrt{14}$	$\sqrt{36}$
C	$\sqrt{31}$	$\sqrt{63}$	-	$\sqrt{60}$	$\sqrt{35}$	$\sqrt{27}$
D	$\sqrt{49}$	$\sqrt{61}$	$\sqrt{60}$	-	$\sqrt{45}$	$\sqrt{53}$
E	$\sqrt{8}$	$\sqrt{14}$	$\sqrt{35}$	$\sqrt{45}$	-	$\sqrt{6}$
F	$\sqrt{10}$	$\sqrt{36}$	$\sqrt{27}$	$\sqrt{53}$	$\sqrt{6}$	-

- We will then form a new cluster EF, for which we calculate the centroid: $\langle 1.5, 0, 2, 0.5 \rangle$, and then we must calculate the distances to this new cluster[1]:

	A	B	C	D	EF
A	-	$\sqrt{18}$	$\sqrt{31}$	$\sqrt{49}$	$\sqrt{7.5}$
B	$\sqrt{18}$	-	$\sqrt{63}$	$\sqrt{61}$	$\sqrt{23.5}$
C	$\sqrt{31}$	$\sqrt{63}$	-	$\sqrt{60}$	$\sqrt{29.5}$
D	$\sqrt{49}$	$\sqrt{61}$	$\sqrt{60}$	-	$\sqrt{47.5}$
EF	$\sqrt{7.5}$	$\sqrt{23.5}$	$\sqrt{29.5}$	$\sqrt{47.5}$	-

- The closest distance now is A with the new cluster EF; the resulting cluster AEF has the centroid $\langle \frac{7}{3}, 0, \frac{5}{3}, \frac{2}{3} \rangle$

	AEF	B	C	D
AEF	-	$\sqrt{20}$	$\sqrt{28.3}$	$\sqrt{46.3}$
B	$\sqrt{20}$	-	$\sqrt{63}$	$\sqrt{61}$
C	$\sqrt{28.3}$	$\sqrt{63}$	-	$\sqrt{60}$
D	$\sqrt{46.3}$	$\sqrt{61}$	$\sqrt{60}$	-

- Now B gets clustered with AEF; ABEF has the centroid $\langle 3, 0, 2.5, 1 \rangle$

	ABEF	C	D
ABEF	-	$\sqrt{33.25}$	$\sqrt{46.25}$
C	$\sqrt{33.25}$	-	$\sqrt{60}$
D	$\sqrt{46.25}$	$\sqrt{60}$	-

- All that is left now is to assign C to ABEF; there is no need to calculate the centroid any more, as there are only two clusters (ABCEF and D) remaining.
- Hence, we have here the agglomerate clustering E-F, A, B, C, D. This is a "traditional" dendrogram, but generally we expect a "non-traditional dendrogram" to result from this process.

[1] There are other ways of performing this step, for example, **single link**: using the shortest distance out of the ones calculated above to the points in this cluster, so that the distance from A to EF is $\min(\sqrt{8}, \sqrt{10}) = \sqrt{8}$