School of Computing and Information Systems
The University of Melbourne
COMP90049 Knowledge Technologies (Semester 2, 2018)
Workshop sample solutions: Week 8

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	?

1. 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:

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

$$
\begin{aligned}
d(F, C_1) &= |\,1-4\,| + |\,0-0\,| + |\,1-1\,| + |\,0-1\,| \\
&= 4 \\
d(F, C_2) &= |\,1-1\,| + |\,0-2\,| + |\,1-1\,| + |\,0-7\,| \\
&= 9
\end{aligned}
$$

- 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-calculate the distances of each instance to each centroid:

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

- (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:

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

- 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:

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

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

- 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 re-calculate the distances according to these new centroids:

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

$$
\begin{aligned}
d(E, C_1) &\approx |\, 2 - 3.67 \,| + |\, 0 - 0 \,| + |\, 3 - 3 \,| + |\, 1 - 1.33 \,| \\
&\approx 2 \\
d(E, C_2) &\approx |\, 2 - 1.33 \,| + |\, 0 - 2.33 \,| + |\, 3 - 0.67 \,| + |\, 1 - 2.33 \,| \\
&\approx 6.67 \\
d(F, C_1) &\approx |\, 1 - 3.67 \,| + |\, 0 - 0 \,| + |\, 1 - 3 \,| + |\, 0 - 1.33 \,| \\
&\approx 6 \\
d(F, C_2) &\approx |\, 1 - 1.33 \,| + |\, 0 - 2.33 \,| + |\, 1 - 0.67 \,| + |\, 0 - 2.33 \,| \\
&\approx 5.33
\end{aligned}
$$

- 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).

2. 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:

	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 can immediately observe (without simplifying the square roots) that the most similar instances (with the smallest distance) are E and F.
   - 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 $\langle \frac{7}{3}, 0, \frac{5}{3}, \frac{2}{3} \rangle$ (see overleaf)
   - Now B gets clustered with AEF; ABEF has the centroid $\langle 3, 0, 2.5, 1 \rangle$ (see overleaf)
   - 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] The lectures discuss 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 $\min(\sqrt{8}, \sqrt{10}) = \sqrt{8}$ — the main advantage of these strategies are to save the cost of re-calculating the distances to the new centroids, which constitutes a very large difference in time for a non-trivial dataset.

	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	$\sqrt{46.3}$	$\sqrt{61}$	$\sqrt{60}$	-

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

&

3. What is **overfitting**? What does it mean for a classifier to **generalise**?

   - A classifier "generalises" when it learns the target function well, rather than the (possibly meaningless) specifics of the training set. Overfitting is when the classifier fails to generalise — it builds a model which describes the training data very well, but doesn't describe the test data well.

4. A **confusion matrix** is a summary of the performance of a (supervised) classifier over a set of development ("test") data, by counting the various instances:

		Actual			
		a	b	c	d
Classified	a	10	2	3	1
	b	2	5	3	1
	c	1	3	7	1
	d	3	0	3	5

   (a) Calculate the classification **accuracy** of the system. Find the **error rate** for the system.

   - In this context, Accuracy is defined as the fraction of correctly identified instances, out of all of the instances. In the case of a confusion matrix, the correct instances are the ones enumerated along the main diagonal (classified as a and actually a etc.):

$$\begin{aligned} \textit{Accuracy} &= \frac{\text{\# of correctly identified instances}}{\text{total \# of instances}} \\ &= \frac{10+5+7+5}{10+2+3+1+2+5+3+1+1+3+7+1+3+0+3+5} \\ &= \frac{27}{50} = 54\% \end{aligned}$$

   - Error rate is just the complement of accuracy:

$$\begin{aligned} \textit{Error Rate} &= \frac{\text{\# of incorrectly identified instances}}{\text{total \# of instances}} \\ &= 1 - \textit{Accuracy} \\ &= 1 - \frac{27}{50} = 46\% \end{aligned}$$

   (b) Calculate the **Precision**, **Recall**, **F-score** (where $\beta = 1$), **Sensitivity**, and **Specificity** for class d. (Why can't we do this for the whole system? How can we consider the whole system?)

   - Precision for a given class is defined as the fraction of correctly identified instances of that class, from the times that class was attempted to be classified. We are interested in the true positives (TP) where we attempted to classify an item as an instance of said class (in

this case, d) and it was actually of that class (d): in this case, there are 5 such instances. The false positives (FP) are those items that we attempted to classify as being of class d, but they were actually of some other class: there are $3 + 0 + 3 = 6$ of those.

$$\begin{aligned} \textit{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \\ &= \frac{5}{5 + 3 + 0 + 3} \\ &= \frac{5}{11} \approx 45\% \end{aligned}$$

- Recall for a given class is defined as the fraction of correctly identified instance of that class, from the times that class actually occurred. This time, we are interested in the true positives, and the false negatives (FN): those items that were actually of class d, but we classified as being of some other class; there are $1 + 1 + 1 = 3$ of those.

$$\begin{aligned} \textit{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\ &= \frac{5}{5 + 1 + 1 + 1} \\ &= \frac{5}{8} \approx 62\% \end{aligned}$$

- F-score is a measure which attempts to combine Precision (P) and Recall (R) into a single score. In general, it is calculated as:

$$F_\beta = \frac{(1 + \beta^2)P \cdot R}{(\beta^2 \cdot P) + R}$$

- By far, the most typical formulation is where the parameter β is set to 1: this means that Precision and Recall are equally important to the score, and that the score is a harmonic mean.
- In this case, we have calculated the Precision of class d to be $\frac{5}{11}$ and the Recall to be $\frac{5}{8}$. The F-score where ($\beta = 1$) of class d is then:

$$\begin{aligned} F_{\beta=1} &= \frac{2 \cdot P \cdot R}{P + R} \\ &= \frac{2 \cdot \frac{5}{11} \cdot \frac{5}{8}}{\frac{5}{11} + \frac{5}{8}} \\ &= \frac{50}{95} \approx 53\% \end{aligned}$$

- Sensitivity is defined the same way as Recall: $\frac{TP}{TP+FN}$.
- Specificity is Precision with respect to the negative instances:

$$\begin{aligned} \textit{Specificity} &= \frac{\text{TN}}{\text{TN} + \text{FP}} \\ &= \frac{10 + 2 + 3 + 2 + 5 + 3 + 1 + 3 + 7}{10 + 2 + 3 + 2 + 5 + 3 + 1 + 3 + 7 + 3 + 0 + 3} \\ &= \frac{36}{42} \approx 86\% \end{aligned}$$

5. How is **holdout** evaluation different to **cross-validation** evaluation?

   - In a holdout evaluation strategy, we partition our data into a training set and a test set: we build the model on the former, and evaluate on the latter.
   - In a cross-validation evaluation strategy, we do the same as above, but a number of times, where each iteration uses one partition of the data as a test set and the rest as a training set (and the partition is different each time). The evaluation metric is typically averaged across the various partitions.