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

Tutorial exercises: 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.

   (b) Using seeds A and F.

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

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

   (a) Explain the significance of the "E" step, and the "M" step.

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

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

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

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

supervised learning: given label in train set.

unsupervised learning: not given. learn the feature of train instance, and classify

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.

   (b) Using seeds A and F.

2.(a) 2 clusters.

seed A: $C_1 = \langle 4, 0, 1, 1 \rangle$

seed D: $C_2 = \langle 1, 2, 1, 7 \rangle$

Calculate distance:

$d(A, C_1) = 0$

$d(A, C_2) = |4-1| + |0-2| + |1-1| + |1-7| = 11$

⋮

assign each instance to closest cluster.

cluster 1: A B C E F

cluster 2: D

Update center:

$C_1 = \langle \frac{4+5+2+2+1}{5}, \dots \rangle = \langle 2.8, 1, 2, 0.8 \rangle$

$C_2 = \langle 1, 2, 1, 7 \rangle$

Again! calculate

⋮

Cluster 1: A B C E F

Cluster 2: D

⇒ stable, stop

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

Soft k-means : (probabilistic)

"Soft max" function

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

z_{ij}: cluster, attribute

For Instance A :

Seed A, F

$$z_{1A} = \frac{e^{-0}}{e^{-0}+e^{-4}} = 0.982$$

$$z_{2A} = \frac{e^{-4}}{e^{-0}+e^{-4}} = 0.018$$

$\vdots$ $\rightarrow$ $z_{1E} = z_{2E} = 0.5$ doesn't matter

Update centroids.

$$C_1^{(1)} = \frac{0.982\,A + 0.982\,B + 0.119\,C + \cdots}{0.982 + 0.982 + 0.119 + \cdots}.$$

$$= \frac{1}{2.72}\left[0.982 \cdot \langle 4,0,1,1\rangle + 0.982 \cdot \langle 5,0,5,2\rangle + \cdots\right]$$

$$= \langle 3.75, 0.30, 2.77, 1.5\rangle.$$

$$C_2^{(1)} = \langle 1.46, 1.88, 1.06, 2.05\rangle$$

after several iterations :

	$d(1,j)$	$d(2,j)$	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

we are quite confident with most instance

until convergence.

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

   (a) Explain the significance of the "E" step, and the "M" step.

Logic : Basically we randomly guess, and progressively improve our guess by evaluating the expected likelihood.

Which is same as clustering, random chose seed and iterate update the centroids by distance.

(a) E (Expectation) : assign weighted label to training data, and calculated expected likelihood.

M (Maximization) : re-estimate the parameter based on these labels.

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

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

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

Semi-supervised learning: we have only a small number of labelled instance, and many unlabelled instances. Typically, this means we don't have enough data to build a reliable model, but potentially we can put labelled data to build a better classifier than purely un-supervised.

(a) Self learning:

1. train model using labelled data
2. use learner to predict unlabelled data
3. put most confident instance to train set
4. Repeat, until all instance labelled, or no new instance confidently

(b) active learning: the learner is able to choose a small number of instances to be labelled by human,

The idea is many instances are easy to classify, only a few instances are hard to classify, but would be easier if we have more trainning data.