School of Computing and Information Systems
The University of Melbourne
COMP90049 Introduction to Machine Learning (Semester 1, 2021)
Solutions: Week 10

1. Let's revisit the logic behind the voting method of classifier combination (used in Bagging, Random Forests, and Boosting to some extent). We are assuming that *the errors between the two classifiers are uncorrelated*

   (a) First, let's assume our three independent classifiers both have an error rate of e = 0.4, calculated over 1000 instances with binary labels (500 A and 500 B).

   (i) Build the confusion matrices for these classifiers, based on the assumptions above.

   All our systems have the error rate of 0.4. It means that in 40% of the times our system makes a mistake and in 60% of the time it makes a correct prediction. So, from 500 instances with label A, 300 instances will be (correctly) predicted with the label A, and 200 instances will be labelled (incorrectly) as B.

   Now since we are doing an ensembled method, we can assume that from the 300 instances that system 1 labelled as A (and 200 instances labelled as B), again 60% will be labelled (correctly) as A and 40% (incorrectly) as B, and so on.

   Table below contains all the counts for our three-classifier ensemble.

Actual class	#	P1	# for Sys1	P2	# for Sys 2	P3	# for Sys 3
A	500	A	(500 x 0.6=) 300	A	(300 x 0.6=) 180	A	(180*0.6=) 108
						B	(180*0.4=) 72
				B	(300 x 0.4=) 120	A	(120*0.6=) 72
						B	(120*0.4=) 48
		B	(500 x 0.4=) 200	A	(200 x 0.6=) 120	A	(120*0.6=) 72
						B	(120*0.4=) 48
				B	(200 x 0.4=) 80	A	(80*0.6=) 48
						B	(80*0.4=) 32
B	500	A	(500 x 0.4=) 200	A	(200 x 0.4=) 80	A	(80*0.4=) 32
						B	(80*0.6=) 48
				B	(200 x 0.6=) 120	A	(120*0.4=) 48
						B	(120*0.6=) 72
		B	(500 x 0.6=) 300	A	(300 x 0.4=) 120	A	(120*0.4=) 48
						B	(120*0.6=) 72
				B	(300 x 0.6=) 180	A	(180*0.4=) 72
						B	(180*0.6=) 108

   (ii) Using that the majority voting, what the expected error rate of the voting ensemble?

   Since we have 3 systems, using a majority voting is very easy. For all the predicted labels we check the results of the 3 system and if 2 out of 3 votes for one class we chose that class as the label for the whole system. The results for the voting system are demonstrated in the following table, where the incorrect predictions are highlighted.

Actual class	#	Pred1	#	Pred2	#	Pred3	#	Majority Vote
A	500	A	300	A	180	A	108	Maj(A,A,A)= A
						B	72	Maj(A,A,B)= A
				B	120	A	72	Maj(A,B,A)= A
						B	48	Maj(A,B,B)= B
		B	200	A	120	A	72	Maj(B,A,A)= A
						B	48	Maj(B,A,B)= B
				B	80	A	48	Maj(B,B,A)= B
						B	32	Maj(B,B,B)= B
B	500	A	200	A	80	A	32	Maj(A,A,A)= A
						B	48	Maj(A,A,B)= A
				B	120	A	48	Maj(A,B,A)= A
						B	72	Maj(A,B,B)= B
		B	300	A	120	A	48	Maj(B,A,A)= A
						B	72	Maj(B,A,B)= B
				B	180	A	72	Maj(B,B,A)= B
						B	108	Maj(B,B,B)= B

From this table we can identify that the total count of incorrectly classified instances is:

$$error = 48 + 48 + 48 + 32 + 32 + 48 + 48 + 48 = 352$$

Therefore, the error rate of the final system (the ensemble of three learners with error rate of 0.4) is $\frac{352}{1000} = 35.2\% = 0.352$. It is better that the error rate of each system individually. It is mostly because using three system has allowed us to disambiguate the instances where each system cannot classify correctly. In other words, the voting system helps the learners to correct each other's mistake.

This relies on the assumption of errors being uncorrelated: if the errors were perfectly correlated, we would see no improvement; if the errors were mostly correlated, we would see only a little improvement.

(b) Now consider three classifiers, first with $e_1 = 0.1$, the second and third with $e_2 = e_3 = 0.2$.

(i) Build the confusion matrices.

Similar to part A we can build a combined confusion matrix for all three systems, where the first system is 90% makes correct prediction (and therefore makes incorrect prediction for 10% of the instances). And the two next systems make correct predictions only for 80% of the instances (and so each system makes 20% incorrect predictions).

The following table contains all the counts for the all different combination of the systems in this ensemble.

A	500	A	(500 x 0.9=) 450	A	(450 x 0.8=) 360	A	288
						B	72
				B	(450 x 0.2=) 90	A	81
						B	18
		B	(500 x 0.1=)	A	(50x 0.8=)	A	36

			50		40	B	8
				B	(50 x 0.2=) 10	A	8
						B	2
B	500	A	(500 x 0.1=) 50	A	(50 x 0.2=) 10	A	2
						B	8
				B	(50 x 0.8=) 40	A	8
						B	32
		B	(500 x 0.9=) 450	A	(450 x 0.2=) 90	A	18
						B	72
				B	(450 x 0.8=) 360	A	72
						B	288

(ii) Using the majority voting, what the expected error rate of the voting ensemble?

The results for the voting system are demonstrated in the following table, where the incorrect predictions are highlighted.

Actual class	#	Pred1	#	Pred2	#	Pred3	#	Majority Vote
A	500	A	450	A	360	A	288	Maj(A,A,A)= A
						B	72	Maj(A,A,B)= A
				B	90	A	81	Maj(A,B,A)= A
						B	18	Maj(A,B,B)= B
		B	50	A	40	A	36	Maj(B,A,A)= A
						B	8	Maj(B,A,B)= B
				B	10	A	8	Maj(B,B,A)= B
						B	2	Maj(B,B,B)= B
B	500	A	50	A	10	A	2	Maj(A,A,A)= A
						B	8	Maj(A,A,B)= A
				B	40	A	8	Maj(A,B,A)= A
						B	32	Maj(A,B,B)= B
		B	450	A	90	A	18	Maj(B,A,A)= A
						B	72	Maj(B,A,B)= B
				B	360	A	72	Maj(B,B,A)= B
						B	288	Maj(B,B,B)= B

From this table we can identify that the total count of incorrectly classified instances is:

$$Error = 18 + 8 + 8 + 2 + 2 + 8 + 8 + 18 = 72$$

Therefore, the error rate of the final system (the ensemble of three learners $e_1 = 0.1$ and $e_2 = e_3 = 0.2$) is now $\frac{72}{1000} = 7.2\% = 0.072$. It is better that the error rate of the best system

(iii) What if we relax our assumption of independent errors? In other words, what will happen if the errors between the systems were very highly correlated instead? (Systems make similar mistakes.)

Basically, if all the systems make the same predictions; the error rate will be roughly the same as the correlated classifiers, and voting is unlikely to improve the ensemble. Even if two of the classifiers are correlated, and the third is uncorrelated, the two correlated systems will tend to "out-vote" the third system on erroneous instances.

Therefore, if we want to use ensemble method, it's best to use uncorrelated learners (classifiers).

2. 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	?

(a) 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:

(i) 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 C1 = (4, 0, 1, 1) and cluster 2 C2 = (1, 2, 1, 7).

We now calculate the distance for each instance ("training" and "test" are equivalent in this context) to the centroids of each cluster:

$d(A, C_1) = |4-4| + |0-0| + |1-1| + |1-1| = 0$
$d(A, C_2) = |4-1| + |0-2| + |1-1| + |1-7| = 11$
$d(B, C_1) = |5-4| + |0-0| + |5-1| + |2-1| = 6$
$d(B, C_2) = |5-1| + |0-2| + |5-1| + |2-7| = 15$
$d(C, C_1) = |2-4| + |5-0| + |0-1| + |0-1| = 9$
$d(C, C_2) = |2-1| + |5-2| + |0-1| + |0-7| = 12$
$d(D, C_1) = |1-4| + |2-0| + |1-1| + |7-1| = 11$
$d(D, C_2) = |1-1| + |2-2| + |1-1| + |7-7| = 0$
$d(E, C_1) = |2-4| + |0-0| + |3-1| + |1-1| = 4$
$d(E, C_2) = |2-1| + |0-2| + |3-1| + |1-7| = 11$
$d(F, C1) = |1-4| + |0-0| + |1-1| + |0-1| = 4$
$d(F, C2) = |1-1| + |0-2| + |1-1| + |0-7| = 9$

We now assign each instance to the cluster with the smallest (Manhattan) distance to the cluster's centroid: for A, this is C1 because 0 < 11, for B, this is C1 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:

$$C1 = \left(\frac{4+5+2++2+1}{5}, \frac{0+0+5+0+0}{5}, \frac{1+5+0+3+1}{5}, \frac{1+2+0+1+0}{5}\right) = (2.8,1,2,0.8)$$

For cluster 2, we're just taking the average of a single value, so obviously the centroid is just (1, 2, 1, 7).

Now, we re-calculate the distances of each instance to each centroid:

d(A, C1) = | 4 – 2.8 | + | 0 – 1 | + | 1 – 2 | + | 1 – 0.8 | = 3.4

d(B, C1) = | 5 – 2.8 | + | 0 – 1 | + | 5 – 2 | + | 2 – 0.8 | = 7.4

d(C, C1) = | 2 – 2.8 | + | 5 – 1 | + | 0 – 2 | + | 0 – 0.8 | = 7.6

d(D, C1) = | 1 – 2.8 | + | 2 – 1 | + | 1 – 2 | + | 7 – 0.8 | = 10

d(E, C1) = | 2 – 2.8 | + | 0 – 1 | + | 3 – 2 | + | 1 – 0.8 | = 3

d(F, C1) = | 1 – 2.8 | + | 0 – 1 | + | 1 – 2 | + | 0 – 0.8 | = 4.6

(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 centroids 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}.

(ii) Starting with seeds `A` and `F`.

This time, the initial centroids are C1 = (4, 0, 1, 1) and C2 = (1, 0, 1, 0).

We calculate the (Manhattan) distances of each instance to each centroid:

$d(A, C_1)$ = | 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

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:

$$C1 = \left(\frac{4+5}{2}, \frac{0+0}{2}, \frac{1+5}{2}, \frac{1+2}{2}\right) = (4.5,0,3,1.5)$$

$$C2 = \left(\frac{2+1+2+1}{4}, \frac{5+2+0+0}{4}, \frac{0+1+3+1}{4}, \frac{0+7+1+0}{4}\right) = (1.5,1.75,1.25,2)$$

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

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

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:

$$C1 = \left(\frac{4+5+2}{3}, \frac{0+0+0}{3}, \frac{1+5+3}{3}, \frac{1+2+1}{3}\right) \approx (3.67,0,3,1.33)$$

$$C2 = \left(\frac{2+1+1}{3}, \frac{5+2+0}{3}, \frac{0+1+1}{3}, \frac{0+7+0}{3}\right) \approx (1.33,2.33,0.67,2.33)$$

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

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

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).

(b) 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.

We begin by finding the pairwise similarities — or distances, in this case, between each instance. I'm going to skip the Euclidian 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: (1.5, 0,2,0.5), and then we must calculate the distances to this new cluster[1].

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

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

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

Now B gets clustered with AEF; ABEF has the centroid (3, 0, 2.5, 1).

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

All that is left now is to assign C to ABEF; there is no need to calculate the centroid anymore, as there are only two clusters (ABCEF and D) remaining.

Hence, we have here the agglomerate clustering E-F, A, B, C, D.

3. Explain the two main concepts that we use to measures the goodness of a clustering structure without external information.

The two main concepts that we check in unsupervised clustering evaluation are the clusters *cohesion* and *separability*. We want the members of each clusters to be as integrated and close to each other as possible meanwhile we want the clusters to be as separate and independent as possible from other clusters.

[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 $(\sqrt{8}, \sqrt{10}) = \sqrt{8}$