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.

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	A
						B	(180*0.4=) 72	A
				B	(300 x 0.4=) 120	A	(120*0.6=) 72	A
						B	(120*0.4=) 48	B
		B	(500 x 0.4=) 200	A	(200 x 0.6=) 120	A	(120*0.6=) 72	A
						B	(120*0.4=) 48	B
				B	(200 x 0.4=) 80	A	(80*0.6=) 48	B
						B	(80*0.4=) 32	B
B	500	A	(500 x 0.4=) 200	A	(200 x 0.4=) 80	A	(80*0.4=) 32	A
						B	(80*0.6=) 48	A
				B	(200 x 0.6=) 120	A	(120*0.4=) 48	A
						B	(120*0.6=) 72	B
		B	(500 x 0.6=) 300	A	(300 x 0.4=) 120	A	(120*0.4=) 48	A
						B	(120*0.6=) 72	B
				B	(300 x 0.6=) 180	A	(180*0.4=) 72	B
						B	(180*0.6=) 108	B

(Same pred as above)

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

For Actual A: $48+48+48+32 = 176$

For Actual B: $32+48+48+48 = 176$

Total error count $= 176+176 = 352$

Expected $ER = \frac{352}{1000} = 0.352 < 0.4$ (individual ER)

learners correct each others' mistake (assume errors uncorrelated)

Alternative:

$$\underbrace{\binom{3}{2}(0.4)^2(0.6)}_{\text{two making mistakes}} + \underbrace{\binom{3}{3}(0.4)^3}_{\text{all wrong}} = 0.352$$

(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.

Actual		Pred 1 ($e_1 = 0.1$)		Pred 2 ($e_2 = 0.2$)		Pred 3 ($e_3 = 0.2$)		
A	500	A	(500 x 0.9=) 450	A	(450 x 0.8=) 360	A	288	A
						B	72	A
				B	(450 x 0.2=) 90	A	81	A
						B	18	B
		B	(500 x 0.1=) 50	A	(50x 0.8=) 40	A	36	A
						B	8	B
				B	(50 x 0.2=) 10	A	8	B
						B	2	B
B	500	A	(500 x 0.1=) 50	A	(50 x 0.2=) 10	A	2	A
						B	8	A
				B	(50 x 0.8=) 40	A	8	A
						B	32	B
		B	(500 x 0.9=) 450	A	(450 x 0.2=) 90	A	18	A
						B	72	B
				B	(450 x 0.8=) 360	A	72	B
						B	288	B

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

Error count $= (18+8+8+2) + (2+8+8+18)$

$= 72$

$ER = \frac{72}{1000} = 0.072 < 0.1$ (sys 1)

Alternative :

$$ER = \underbrace{(0.1)(0.2)^2}_{\text{all wrong}} + \underbrace{(0.9)(0.2)^2}_{\text{S2 \& S3 wrong}} + \underbrace{2 \times (0.1)(0.2)(0.8)}_{\text{S1 \& S2/S3 wrong}}$$

$= 0.072$

(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.)

Highly correlated errors ⇒ voting unlikely to improve the results
(making same mistakes)

E.g. S_1 & S_2 correlated, S_3 uncorrelated

⇒ S_1 & S_2 "out-vote" S_3

If S_1 & S_2 wrong ⇒ ensemble wrong

⇒ Choose uncorrelated classifiers for ensembling.

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.

(a) Seed A, D (K=2)

⇒ two initial centroids:

$C_1 : \langle 4, 0, 1, 1 \rangle$

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

① Calculate distances from each instance to the centroids (of clusters)

$d_M(A, C_1) = 0$

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

$d_M(B, C_1) = |5-4| + |0-0| + |5-1| + |2-1| = \underline{6}$

$d_M(B, C_2) = \underline{15}$

$d_M(C, C_1) = 9$ $d_M(D, C_1) = 11$ $d_M(E, C_1) = 4$

$d_M(C, C_2) = 12$ $d_M(D, C_2) = 0$ $d_M(E, C_2) = 11$

$d_M(F, C_1) = 4$

$d_M(F, C_2) = 9$

② Assign each instance to the nearest cluster

$A \to C_1$ $B \to C_1$ $C \to C_1$ $D \to C_2$

$E \to C_1$ $F \to C_1$

$C_1 : A, B, C, E, F$ $C_2 : D$

③ Update centroids (average the instances in that cluster)

$C_1 = \left\langle \frac{4+5+2+2+1}{5}, \frac{5}{5}, \frac{1+5+3+1}{5}, \frac{1+2+1}{5} \right\rangle$

$= \langle 2.8, 1, 2, 0.8 \rangle$

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	?

$C_2 = \langle 1, 2, 1, 7 \rangle$ (Just "D")

$\Rightarrow$ repeat! (with new C_1 & C_2)

C_2 hasn't change : reuse dist from last iter.

$d_M(A, C_2) = 11$

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

$d_M(B, C_2) = 15$
$d_M(C, C_2) = 12$
$d_M(D, C_2) = 0$
$d_M(E, C_2) = 11$
$d_M(F, C_2) = 9$

C_1: A, B, C, E, F
C_2: D
} final assignment

Same as last iter! => Converged! => Stop!

(ii) Starting with seeds A and F.

Skip.

tie:

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
} randomly pick one

(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.

In workshop slides.

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

① Cohesiveness

Members of each cluster to be integrated and close to each other as possible.

② Separability

Clusters to be separate & independent as possible from the other clusters.