School of Computing and Information Systems
The University of Melbourne
COMP90049 Introduction to Machine Learning (Semester 2, 2020)
Workshop: Week 10

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

2. 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.

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

   (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.

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

      (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.)

3. 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.

      (ii) Starting with seeds A and F.

   (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.