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

Workshop: Week 11

1. When do we use semi-supervised learning? What is self-training?

2. What is the logic behind active learning, and what are some methods to choose instances for the oracle?

3. One of the strategies for Query sampling was query-by-committee (QBC), where a suite of classifiers is trained over a fixed training set, and the instance that results in the highest disagreement amongst the classifiers, is selected for querying. Using the equation below, which captures vote entropy, determine the instance that our active learner would select first.

$$x_{VE}^{*} = \underset{x}{\text{argmax}}(-\sum_{y_i} \frac{V(y_i)}{C} log_2 \frac{V(y_i)}{C})$$

Respectively y_i, $V(y_i)$, $and\ C$ are the possible labels, the number of "votes" that a label receives from the classifiers, and the total number of classifiers.

	Instance 1			Instance 2			Instance 3		
classifier	y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3
C_1	0.2	0.7	0.1	0.2	0.7	0.1	0.6	0.1	0.3
C_2	0.1	0.3	0.6	0.2	0.6	0.2	0.21	0.21	0.58
C_3	0.8	0.1	0.1	0.05	0.9	0.05	0.75	0.01	0.24
C_4	0.3	0.5	0.2	0.1	0.8	0.1	0.1	0.28	0.62

4. Given the following univariate dataset, calculate a statistical model based on the assumption that your data is coming from a normal distribution. Determine whether the instance x=1.2 is anomalous or not if we use the boxplot test?

X = {2, 2.5, 2.6, 3, 3.1, 3.2, 3.4, 3.7, 4, 4.1,4.8}

5. Given the following univariate dataset, determine the outlier score for instances (x=0.5) and (x=4) using the following strategies:

Dataset = {1,1.05, 1.1, 1.15, 1.2, 1.21, 1.3, 1.4, 1.45, 1.5, 4.55, 5.6, 6.8, 7.58, 8.6, 9.7, 10.3, 11.4, 12.3,13.5}

(a) Inverse Relative density using 2-NN (Manhattan distance)

(b) Distance to 2nd nearest neighbor (Manhattan distance)