School of Computing and Information Systems
The University of Melbourne
COMP30027 Machine Learning (Semester 1, 2021)

Tutorial: Week 9

1. What is the difference between "model bias" and "model variance"?

   (i). Why is a high bias, low variance classifier undesirable?

   (ii). Why is a low bias, high variance classifier (usually) undesirable?

2. Describe how holdout, and cross-validation can help reduce overfitting?

3. Why *ensembling* reduces model variance?

4. Imagine you are given a dataset from the university's library, and your job is to build a classification model that classify students based on the list of books that they borrowed.

   The dataset includes the list of books available in the library (columns) and the students who borrowed them (rows), and the ranking for each item (ranking value is between 0–5, 0 if the book was not borrowed and 1–5 indicates the student's interest). The metadata for the books (e.g., titles) are not readily available to us, we just have the book IDs (e.g., Book #i). The dataset also includes the students' field of study (in total there are 10 fields), which can be used for the classification task. Answer the following questions, considering that there are 500,000 students and 100,000 books in this dataset.

Student ID	Book #1	···	Book #100,000	Label (Field of Study)
Student # 1	3	···	2	Computer science
Student # 2	5	···	0	Biology
⋮	⋮	⋮	⋮	⋮
Student # 500,000	1	···	4	Mathematic

   (i). Consider the following supervised machine learning methods, and for each one, explain why it would be appropriate or inappropriate to use for this problem:

      i. Naïve Bayes
      ii. k-NN
      iii. Decision Tree

   (ii). Would "feature selection" be useful here? Explain why, by referring to a single machine learning method.

   (iii). Explain how you would evaluate the effectiveness of your system: you should briefly describe an evaluation strategy and an evaluation metric that are suitable for this data. What might be an example of a baseline?