STA130H1S – Fall 2022

Module 9 Tutorial Handout

Today's agenda (5 min):

- Q&A/vocabulary list
- Group Discussion
- Writing prompt

This Week's Vocab (15-20 min) :

- Classification
- Prediction
- Predictor(s)
- Covariate(s)
- Independent variable(s)
- Dependent variable(s)
- Input(s)
- Output(s)
- Training set/sample
- Validation
- Testing set/sample (or test set)
- Fitting a model
- Confusion matrix
- Category
- Tree
- Terminal node (or leaf node)
- Stopping rule
- Threshold
- True positive (sensitivity)
- True negative (specificity)
- False positive
- False negative
- Accuracy
- Classifier
- Node(s)
- Binary
- Split(ting)

Discussion (20 min) :

- What is a "good" split? (A "good" split is one that makes its child node as pure as possible (i.e. homogeneous with respect to the response) - A node is pure if it contains only observations from one class; A node is impure if it contains an equal mix of all the classes)
- How do we make the "best" split? (We come up with a measure of how much more pure the two child nodes would be (compared to the parent node), ΔI. When we want to split a node, we look at: 1. each

potential predictor variable; 2. each possible split for each variable and calculate ΔI. The "best" split is the one with the biggest decrease in impurity ΔI.)

- When to stop splitting? (There are two competing goals when building a classification tree: 1. We want each terminal (leaf) node to be as pure as possible; 2. We don't want a tree that is too complex. A simple "stop splitting" rule is to set a threshold $\beta > 0$.If none of these possible splits for a node makes the tree at least β units more pure, we don't split any further.)
- Explain what a confusion matrix is and how each cell is calculated.
- Discuss:
  - Suppose you developed a classification tree to diagnosis whether or not somebody has Disease X, which is a very serious and life-threatening illness if left untreated. The overall accuracy of your tree was 77%; false-positive rate was 32%; and false-negative rate was 7.9%. (Both types of error are bad (and we try to minimize them!) but in some contexts, one type may be worse than the other. It is often more interesting to look at these error rates separately than just to look at the overall accuracy.)
  - Suppose that your colleague also created a classifier for the same purpose. Its overall accuracy is 81%; false-positive rate is 6.4%; and false-negative rate is 39%. Explain which of these two classifiers you would prefer to use to diagnosis Disease X.
  - Consider the same 2 classifiers for Disease X, but now suppose the treatment is very expensive and has many bad side effects; e.g. people taking the treatment tend to get very sick, similar to chemotherapy. In this case which classifier would you prefer?