STA130H1S – Fall 2022

Week 11 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, or we risk over fitting. A simple rule is to stop splitting when there's no "sufficient improvement" (defined by the `cp` parameter in the `rpart()` function as $Error(T) + cp \times splits \times Error(T_0)$)]

- 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?

Writing prompt (30 min) : The activity this week will focus on two areas, varying your register and describing classification trees. Prior to starting the assignment, it is highly recommended that you review the infographic available here **to be changed**, and also available under Module 8.

Imagine you built a classifier to predict what movie somebone would be most interested in. You used this classifier to recommend movies to your friends and family. They were amazed by how well you know their movie taste and wanted to know more about how you did it. Describe what a classification tree is, which variables you used, using at least 3 of the words from the vocabulary, to one of the audience below:

- Your university friend not in statistics major.
- Your 10-year-old cousin.
- Your parent or guardian with little background in maths.

(Hint: Depending on your audience, their movie preferences and factors that determine their movie preferences will likely differ. You should develop/test your model using datasets that are representative of the population you'd like to apply the classifier to.)

Vocabulary

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

- Node(s)
- Terminal Node
- Binary
- Split(ting)

Some things to keep in mind

- Try to not spend more than 20 minutes on your writing (plus the time to read the article).
- Aim for more than 200 but less than 400 words.
- Use full sentences.
- Grammar is not the main focus of the assessment, but it is important that you communicate in a clear and professional manner (i.e., no slang or emojis should appear).
- Be specific. A good principle when responding to a prompt in STA130 is to assume that your audience is not aware of the subject matter (or in this case has not read the prompt).