STA130 Fall 2019 – T0107

Week 8: Classification

(Materials used in this presentation are provided by the U of T Statistical Sciences Department.

This presentation was prepared by Vivian Ngo.)

Github.com/vivianngo97/STA130-Fall-2019

viv.ngo@mail.utoronto.ca

Agenda

- **Material & Vocabulary review**
- **Discussion questions**
- **Oral presentations**
- **Group work – in poster project groups**
- **Record poster project groups *****
- **In the remaining time, students may work on the poster project**

Material and Vocabulary Review

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

Discussion questions

- 1
- Can you think of any real-life examples where you may want to develop a classification tree?

Discussion questions

- 2
- [From Practice Problem Question 2] Suppose you developed a classification tree to diagnose 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: 7.9%.
- 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.

Discussion questions

- 3
- 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?

Discussion questions

- **Note:** There are many metrics to measure the performance of your classifier and the context for which you are using the classifier matters: accuracy vs false negatives vs false positives.

Discussion questions

- 4
- Suppose you developed a classification tree only to later discover that the values for one of your covariates is missing for a number of observations. Can you use the classification tree you built to make a prediction for these individuals? Explain. [See Practice Problem 1e for example]

Discussion questions

- 5. Imagine you were interested in making a classifier to predict what movie somebody would be most interested in. To do this, you first gathered data from a sample of your closest friends. You validated and tested your classifier using different subsets of this data. Now you wish you use your classifier to predict which movie Dr. Moon/ White, your TA, your parents, etc. would like. How well do you think your classifier will perform in each of these cases?

Oral presentations

- ***Topic one:***
  - Explain how to make an ROC curve and the type of information it provides.
  - Based on the ROC curves you created for Practice Problem 4c, describe the accuracy of each of the two trees.
  - Does this fit your expectations based on the description of each classifier?
  - Which ROC curve would you prefer to classify your spam mail?
- ***Topic two:***
  - Explain what a confusion matrix is and how each cell is calculated.
  - Using the confusion matrix you calculated in question 1d to answer the following questions: What percentage of countries with "good life expectancy" that were classified as having such actually had "good" life expectancy according to the majority rules cutpoint (i.e., 50%) based on each of the two classifiers?
  - What are other terms used to describe the percentages you calculated above?
  - How do the two classifiers compare? Does this fit your expectations based on the description of each classifier?
- ***Topic three:***
  - Summarize the classification tree from Practice Problem 1b. Make sure to include *at least* the following points: how the splits on each variable were selected, how a new observation would be predicted by this classification tree. In part c, you considered more factors. Do you think there may be other important factors to consider? Explain how including these might impact the accuracy of your tree.

Poster project groups

Group #	Member 1 (First Last Names)	Member 2	Member 3	Member 4
1				
2				
3				
4				
5				