STA 130 Tutorial 7

David Veitch

University of Toronto

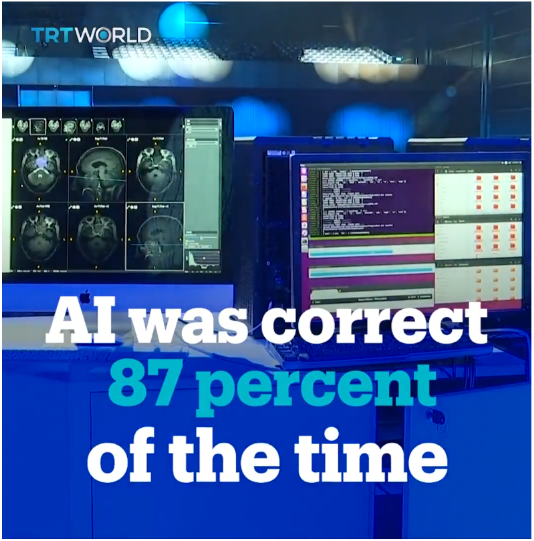
daveveitch.github.io

Mar 8, 2019

Agenda

1. Classification Real World Example
2. Project Groups
3. Vocabulary
4. Review Problems
5. Presentations
   - Making a Good Presentation
   - Presentation Topics

Classification Real World Example



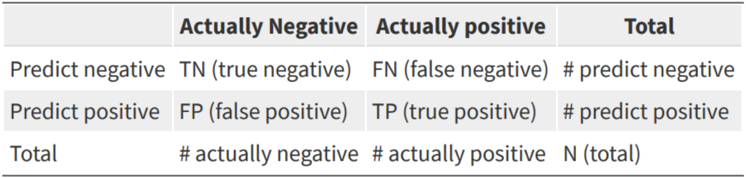
https://www.youtube.com/watch?v=_kLPyDmUUwU:

Project Groups

- **Final Project:** Statistical analysis of real data from www.riipen.com.
- Groups 3-4 students from this tutorial.
- Make them yourselves, tell me next time.

Vocabulary

- **Pure Node:** Contains observations from only one class
- **Impure Node:** Contains an equal mix of all classes
- **Evaluating the accuracy of a tree**:

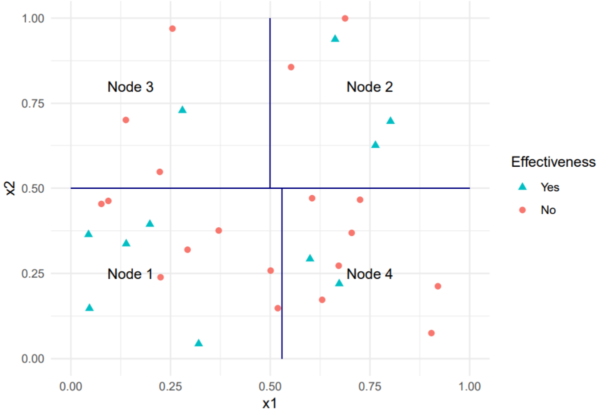


	Actually Negative	**Actually positive**	**Total**
Predict negative	TN (true negative)	FN (false negative)	# predict negative
Predict positive	FP (false positive)	TP (true positive)	# predict positive
Total	# actually negative	# actually positive	N (total)

Vocabulary

Review Problem 1

Create the confusion matrix for this classifier:



Practice Problem 2

Name three reasons that Apple may want to develop a classification tree?

Practice Problem 3

Suppose you developed a classification tree to diagnosis whether or not somebody has Disease X. Overall accuracy of your tree was 77%; False-positive rate was 36%; and False-negative rate: 6.7%. Now suppose that your colleague also created a classifier for the same purpose. Its overall accuracy is 81%; False-positive rate is 15%; and False-negative rate is 21%. Now suppose the disease is very serious if untreated. Explain which classifier you would prefer to use.

Practice Problem 4

Consider the same 2 classifiers for Disease X, but now suppose the treatment for the disease is very expensive and has many bad side effects; e.g. people taking the treatment tend to get very sick while on the treatment, similar to chemotherapy. In this case which classifier would you prefer?

Practice Problem 5

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.

Practice Problem 6

Suppose 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 these cases?

Presentations!

The 4 C's

- Calm (before and during)
- Confident (in yourself and your material)
- Clear (and coherent)
- Concise (but provide enough detail so its coherent, above)

Content Tips

- What is the main message you want to get across?
- Create an (organized) outline of your presentation
- Define terms early
- Make clear transitions between parts of your presentation
- Make your data/figures meaningful
- Summarize

Delivery

- Be confident, make eye contact and avoid reading
- Avoid filler words ummm, like, you know
- Speak slowly and its ok to pause (and breathe!)
- Remember to enunciate all the parts of each word
- Practice! Practice! Practice!

Presentation Topics

1. Refer to HW Q4. Explain how to make a ROC curve and the type of information it provides. Based on the ROC curves provided, describe the accuracy of each of the 3 trees. Does this fit your expectations based on the description of how each classifier identified spam emails?
2. Refer to HW Q2. Explain what a confusion matrix is and how each cell is calculated. Using the calculated confusion matrix answer the following questions: What percentage of disease positive people who were classified as disease positive were actually disease positive according to cutpoint A? According to cutpoint B? What is another term used to describe the percentage you calculated above?
3. Refer to HW Q1B. Summarize the classification tree from part (b); 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. Do you think there may be other important factors to consider? Explain.