STA130H1S TUT0109
W8: Classification Trees

Mar 8, 2019

(Materials used in this presentation are provided by the UofT Statistical Sciences Department)
Special thanks to Vivian Ngo for the slides and Kahoot!

Email: sonia.chhay@mail.utoronto.ca
Website: soniachhay.github.io

Overview

- **Reminder**: Information about final project is posted on Quercus
  - Forming teams
  - Starting next week, will give some designated tutorial time to work together on project and ask questions
- Material, vocabulary, homework discussion
- Group work and presentations
- **Next week**:
  - Let me know who the group members for your final project will be
  - Groups are 3-4 people max., *must* all be from same tutorial
  - Start forming your groups today!
  - **Note:** Cannot switch groups, you'll have until next Friday to decide who you're working with

Material and Vocabulary Review

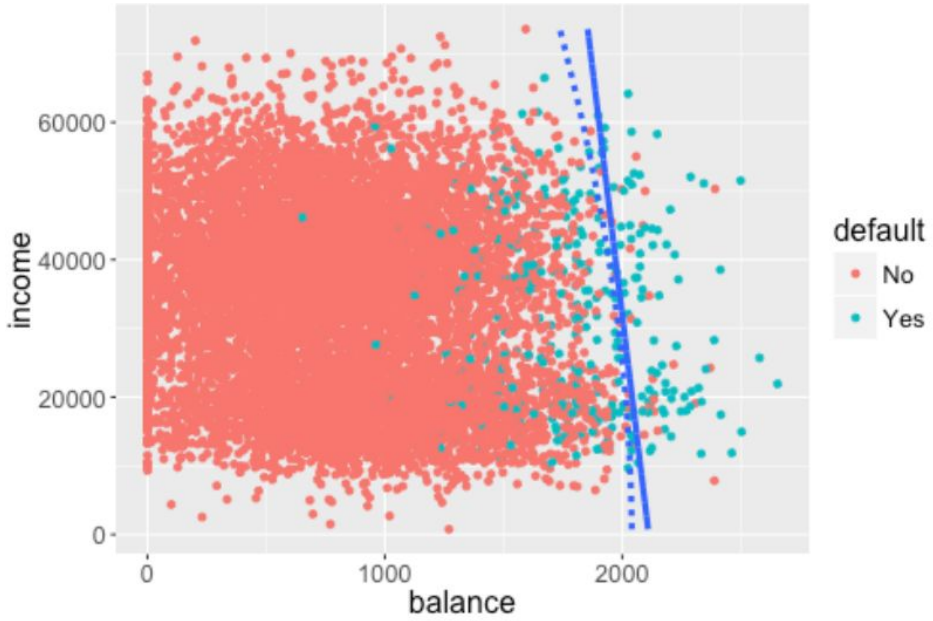
- Supervised learning
- Unsupervised learning
- Classification
- Receiver Operating Characteristic (ROC) curve
- Prediction
- Predictor(s)
- Covariate(s)
- Independent variable(s)
- Dependent variable(s)
- Input(s)
- Output(s)
- Confusion matrix
- Category
- Tree
- Terminal node
- Stopping rule
- Threshold
- True positive (sensitivity)
- True negative (specificity)
- False positive
- False negative
- Accuracy
- Classifier
- Node(s)
- Terminal node
- Binary
- Split(ting)

Supervised and Unsupervised Learning

- Supervised learning: “Teach the model” and then have it predict future instances
  - Model is trained on labeled dataset so it can predict the outcomes
- Unsupervised learning: Let the model “discover” information by itself
  - Only have input data; corresponding response values are not given
  - Less controllable environment

* This course will only focus on statistical methods for supervised learning

Classification: Credit Defaulters



- Solid line is a decision boundary
- How good is the classifier?

Sensitivity and Specificity

- Sensitivity: Percentage of cases/proportion of actual positives (Y=1) that are correctly identified
  - E.g. Telling someone who does have cancer that they have it
- Specificity: Percentage of non-cases/proportion of actual negatives (Y=0) that are correctly identified
  - E.g. Telling someone who does not have cancer that they do not have it

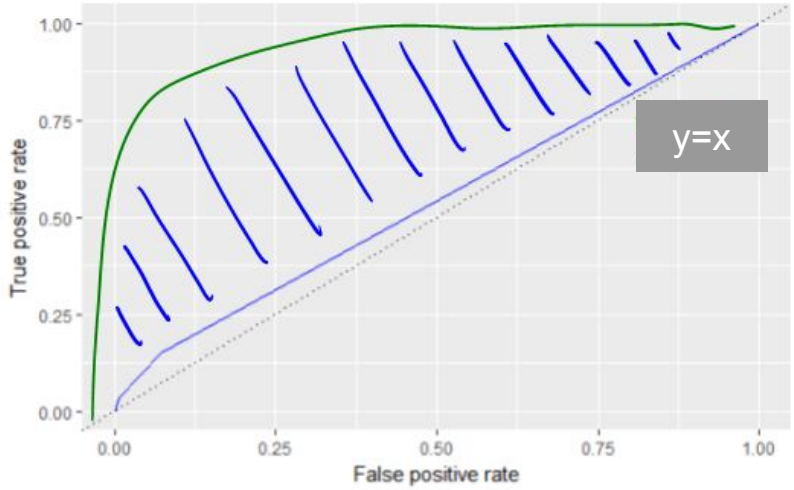
Quantifying the accuracy of a tree

	Actually Positive		Actually negative		Total
Predict positive	a	TP	b	FP	a+b
Predict negative	c	FN	d	TN	c+d
Total	a+c		b+d		N = a+b+c+d

- **True positive (TP) rate:** a/(a+c) **sensitivity**
- **True negative (TN) rate:** d/(b+d) **specificity**
- **False positive (FP) rate:** b/(b+d)
- **False negative (FN) rate:** c/(a+c)
- **Accuracy:** (a+d)/N

Need to decide what "positive" and "negative" mean to calculate these.

ROC curve



- Receiver Operating Characteristic (ROC) curve
- Displays trade-off between sensitivity and specificity for all possible thresholds
  - Plots TP rate (sensitivity) against FP rate (1 - specificity)
- ROC curve that's consistently above line y=x belongs to a better than random classifier
- Good classifier is as close to the top-left corner as possible

Group Discussion

5 questions
(20-25 min)

Discussion 1

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

Discussion 2, 3

- Suppose you developed a classification tree to diagnose whether or not somebody has Disease X. Overall accuracy of your tree was 77%; False-positive rate was 36%; and False-negative rate was 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%.

1. Now suppose that the disease is very serious if untreated. **Explain which classifier you would prefer to use.**

2. Consider the same 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?**

Discussion 4

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.**

Discussion 5

5. 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 to use your classifier to predict which movie Dr. White, your TA, your parents, etc. would like. **How well do you think your classifier will perform in these cases?**

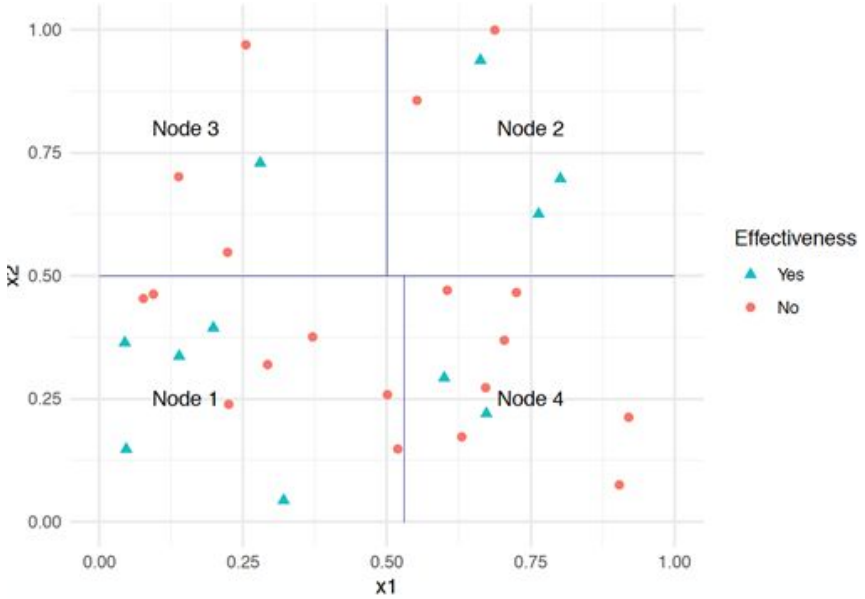
Takeaways

- 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
- You should develop/test your model using datasets that are representative of the population you'd like to apply the classifier to

	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)

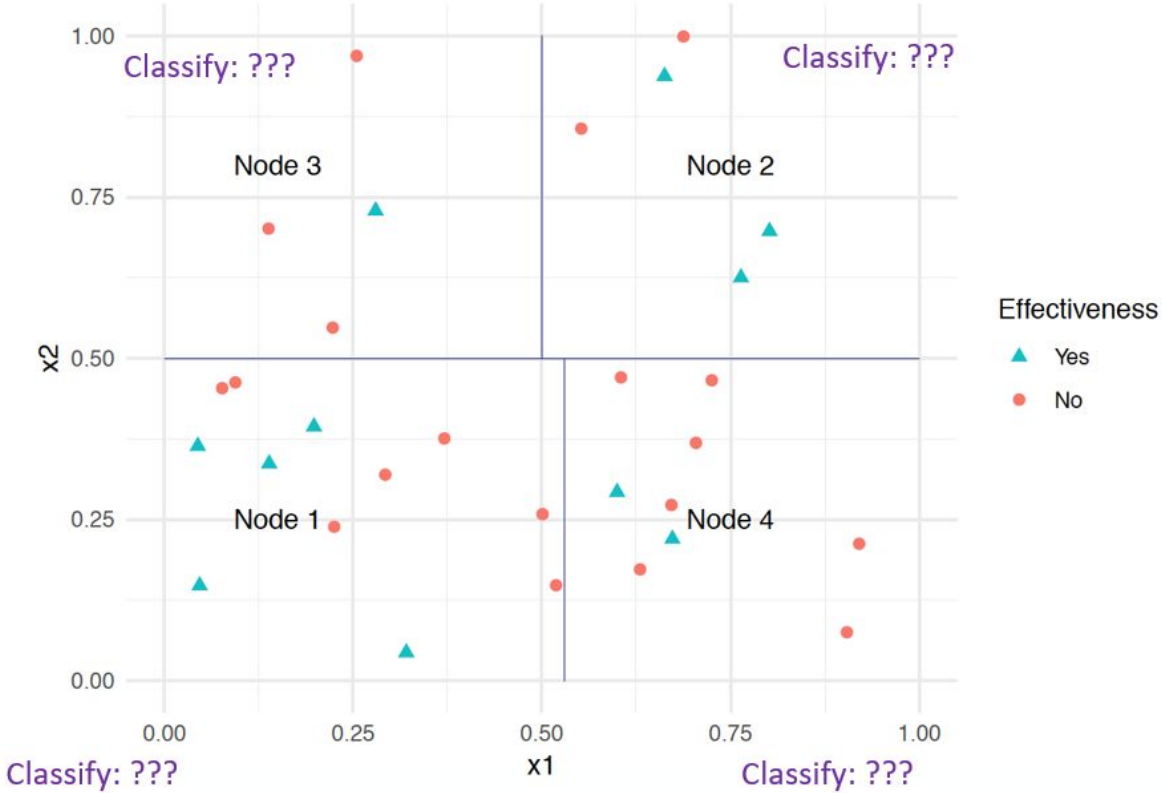
Homework: Question 3

- Data was collected on 30 cancer patients to investigate the effectiveness (Yes/No) of a treatment.
- Two quantitative variables, x_i in (0, 1), i = 1, 2, are considered to be important predictors of effectiveness.
- Suppose that the rectangles labelled as nodes in the scatterplot below represent nodes of a classification tree.



- What is the predicted class of each node?
- What proportion of observations in each node is correctly classified?
- Create a confusion matrix, and calculate the overall accuracy.

Homework: Question 3



Node	Predicted class	Proportion correctly classified
1		
2		
3		
4		

Confusion matrix:

Predicted	Yes	No
Yes		
No		

Overall Accuracy:

Oral Presentations

THE 4 C'S: Calm; Confident; Clear; Concise

Tips for giving a great oral presentation: Content

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

Tips for giving a great oral presentation: Delivery

1. Be confident, make eye contact and avoid reading
2. Avoid filler words – "ummm", "like", "you know"
3. Speak slowly and it's ok to pause (and breathe!)
4. Remember to enunciate all the parts of each word
5. Practice! Practice! Practice!

Group Presentations

- Prepare a 5 minute oral presentation based on the following topics (next slide)
- When not presenting:
  - One person from each group evaluate other students, upload rubric to Quercus
    - Be sure to include the names of the group you're evaluating
  - Write down any questions you have
  - Presentation rubric on Quercus and Github

Oral Presentations

Topic one:

- Refer to Question 4 from the homework.
- 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?

Topic two:

- Refer to Question 2 from the homework.
- 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?

Topic three:

- Refer to Question 1b.
- 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.

	4 (Excellent)	3 (Good)	2 (Adequate)	1 (Poor)
Context	The context and connection to the problem are clear.	Some context was provided and all variables/concepts were mentioned. Some aspects were not clear.	Very little context was provided and only some variables/ concepts were mentioned.	No context and mentioning of any variables/ concepts covering in this week's materials.
Structure	Well organized, follows a logical structure.	The organization follows some logical structure.	Some structure but difficult to follow.	There is no structure, very difficult to follow.
Conclusion	There is a clear central idea and the conclusion is correct.	A central idea or conclusion is present. The conclusion might be incorrect.	The central idea or conclusion is weak and not supported.	The central idea or conclusion is missing. Incorrect conclusion.
Transitions	The progression is logical. Effective use of transitions.	The progression is controlled. The use of transitions is mostly meaningful.	Minor disruptions in flow and weak transitions.	Weak progression and lack of transitions.
Vocabulary	Good use of statistical terms and appropriate choice of words.	Use of statistical terms and phrases mostly correct, demonstrates understanding of concepts.	Some use of statistical terms/ phrases and some understanding of concepts demonstrated.	Inaccurate or incorrect use of statistical terms or phrases and a lack of understanding statistical concepts.
Presentation Skills	Regular eye contact with all parts of the audience. The audience was engaged. The presenter held the audience's attention. Appropriate speaking volume & body language. Good pace.	Somewhat regular eye contact or eye contact with some of the audience The audience was mostly engaged. The presenter mostly spoke at a suitable volume. Spoke too quietly at times. Some fidgeting. Going too fast/slow.	Focused on only one or two members of the audience. Sporadic eye contact. The audience was not engaged. Speaker could be heard by only some of the audience. Body language was distracting.	Minimal (or no) eye contact. The audience was never engaged. The presenter did not speak clearly. Presenter was very difficult to hear.
Preparedness/ Participation	Extremely prepared and rehearsed. The presenter was confident.	Mostly prepared but some dependence on or reading off of notes. The presenter seemed fairly confident.	The presenter was not well prepared. The presenter did not seem confident.	Evident lack of preparation/rehearsal. Complete dependence on notes.

Reminder: Final Project

- Next week, you need to let me know who is in your group
- Each group can have 3-4 people max.