

CIS 490 Machine Learning

Lecture 11

Instructor: (Julia) Hua Fang

1

Last time

2

Supervised Learning

CART: Classification and Regression Trees (Decision Trees, nonlinear)

- **Regression Trees:** You are expected to know
 - ❖ When to use
 - ❖ Tree building process: Top-down and greedy approach (or recursive binary splitting)
 - ❖ Pruning: Algorithm, Cross validation for tuning parameter
- Note: for regression, check **SSR/SSE/RSS, MSE or RMSE**
- **Classification trees:** You are expected to know
 - ❖ When to use
 - ❖ Tree building process: Top-down and greedy approach (or recursive binary splitting)
 - ❖ Classification error, Gini, Deviance.

Adapted from Jeff Howbert, Greg Shakhnarovich, Patrick Breheny, M. Magdon-Ismael, Patrick Breheny, Jeff Schneider

2

3

Answers to Quiz in LS10

1. If Y is a binary variable, what classification method would you consider? **Logistic regression**
2. If Y has more than 2 categories, e.g, Y0= no risk; Y1= mild risk; Y3 = high risk, what classification method would you consider? **Multinomial Logistic**
3. Logistic regression models the logit-transformed probability as a linear relationship with the predictors (T/F)
4. For the s-curve associated with one attribute logistic regression model, the intercept controls slope of rise (T/F)
5. The steeper the S-curve, the better the classifier (T/F)

3

4

Quick Question:

CART is for regression or classification?

- a. regression
- b. classification
- c. Both

4

5

Quiz: CART/Decision Trees

Works for both **categorical** and **continuous** input (X) and output (Y) variables.

Q1: When to use regression trees?

- a. X is continuous; b. X is categorical;
- c. Y is continuous; d. Y is categorical

Q2: When to use classification trees?

- a. X is continuous; b. X is categorical;
- c. Y is continuous; d. Y is categorical

Idea: Split the population or sample into two or more **homogeneous** sets (or **sub-populations**) based on most significant **splitter** (or **differentiator**) in input variables (X).

5

6

Outline

- R: Running CART in R studio.
- Introduction to Exam I and Project II Classification.

6

7

R: Running CART in R studio.

R: Run regression trees in Rstudio using Hitter data

Run classification trees in Rstudio using Heart Disease data

Let's go through the instruction file
"R_CART_S22.docx" posted with LS11 slides at myCourses.

7

8

Exam I: March 15

- Exam I instruction posted at myCourses
- 100 points
- Bring your calculator; No smart phones are allowed. Inter-personal exchange is NOT allowed in any format.
- Exam Place: **Online** at myCourses;
- Date and Time: March 15, 2:00pm ~ 3:15pm; 1:45pm start to download the exam paper and upload your exam paper by 3:45pm.
- Print your name on the first page after reading exam rules.

8

9

Sectional Project II: (65 points) A report attached with code and output, and Presentation Slides Submission **due March 21**; in-class **presentation, March 22**

Instruction:

Required: Apply **logistic regression** and CART-**Classification Tree** for **one** of three data sets listed below (you pick).

- **Heart disease data:**

<https://www.statlearning.com/s/Heart.csv>

- **default of credit card clients Data Set:**

<https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients>

- **Dogs vs. Cats**

<https://www.kaggle.com/c/dogs-vs-cats/data>

9

10

Instruction: continued

Refer to Lecture slides, Reading assignments, and R Instruction Files for logistic regression and tree models, complete the following:

1. Identify Y and X for this dataset and (Read carefully about the document related to each dataset: what Y is? What X are?). Please **name** Y and X (don't call them x1, x2...etc.);
2. Describe your logistic and classification tree model in detail within the context of your selected dataset: e.g., the pruning algorithm for your tree model within the context of your dataset.
3. Interpret your output within the context of your selected dataset.

10

11

Instruction continued:

4. Generate confusion table for each model, logistics and classification tree, and use a summary table to compare all your models.
5. Compute and compare the classification accuracy and error, sensitivity, specificity, PPV and NPV across your compared models, and interpret them with the context of your selected dataset
6. Generate ROC and compute AUC for each model: you may consider drawing (or comparing) all classifiers on one ROC graph.
7. Generate s-curve for Y against one attribute (you can pick any one attribute), and interpret your findings.
8. Cite references.

11

12

Submission Instruction

- Submit two files:
 - Written report in Word or PDF format: No page limit; no template for sectional project report; requirement: make it nice and neat (note: Final project does have a template)
 - Include your code and output on last pages.
 - PowerPoint slides for presentation (~8 slides)

Note: Don't submit zipped file.

12

Grading Rubric: total 65 points

13

- A report in Word or PDF Format, including R code, output appended on last pages (total 45 points): No page limit.
Written Report: 30 points (covering 8 items as specified in above instruction);
Coding: 15 points.
 - Slide presentation (20 points): Each group has ~10 min to present and demo your project, so using ~ 8 slides
Suggestions: use graphs/images/tables/flowcharts for presentation, you can use GIF to demo your work.
Summarize and compare results from these two methods.
 - Highlight the specific techniques you applied and learnt from these lectures
 - Highlight the part you are most proud of in this project
- Zoom presentation:** All group members are required to turn on videos when you are presenting, while other groups, please turn off your videos. Please turn on your videos if you have questions raised for the presenting group.

13

14

On your last slide: Please show Summary of your group meet time and duration

In person or Zoom:

Group meet time and duration (e.g., 5pm-7pm, Feb 1st):
Average time in communication and discussion regarding assigned group work (via email or other social media, e.g. What's app.):
Participants (Print and sign your names):

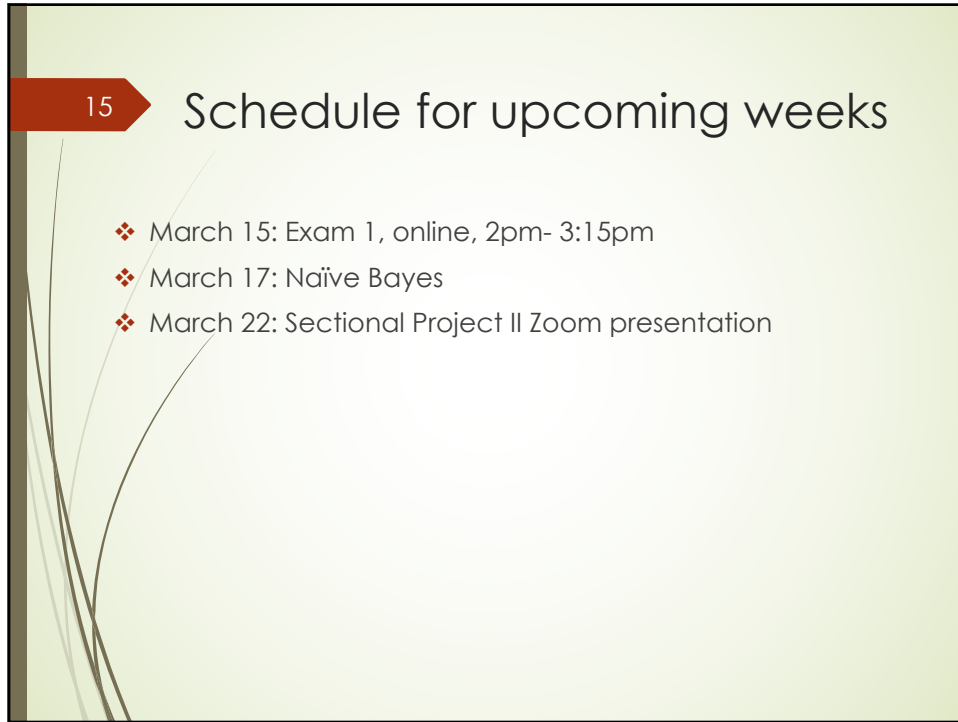
Contribution report:

If your team members contribute equally to this project, please make this statement "Each member contributes equally" on your last page, so that each of you will receive the same score.

If your team members do not contribute equally to this project, please note your team members' names, and mark the percentage of effort each member makes (e.g., Sukumar: 80% then if your group receives a project score of 30, then this member with 80% effort will only get 24).

Participants: Print and sign your names

14



15

Schedule for upcoming weeks

- ❖ March 15: Exam 1, online, 2pm- 3:15pm
- ❖ March 17: Naïve Bayes
- ❖ March 22: Sectional Project II Zoom presentation