Introduction to Data Mining and Statistical Machine Learning

Rebecca C. Steorts, Duke University

STA 325, Module 0

Agenda

- Why Machine Learning?
- Motivational Ideas
- Expectations for this Class
- Topics Covered
- Questions?

What is Machine Learning?

"Statistics is the science of learning from data. Machine Learning (ML) is the science of learning from data. These fields are identical in intent although they differ in their history, conventions, emphasis and culture."

- Larry Wasserman, Rise of the Machines

Machine Learning versus Statistics

- Machine learning and statistics have much to learn from each other.
- In this course, we will focus on machine learning, whereas in other courses, you will learn the fundamentals of statistics.
- To be successful, you need both insights and perspectives. (As a follow up class, take Cynthia Rudin's machine learning course).

Motivations

Let's talk about a very small number of motivational problems in machine learning.

Think on your own about how you might try and tackle them if you had the data right now.

Apple music database

Suppose you have an Apple database with 2 million songs on it. How would you identify the number of unique songs?

I and Love and You	5:01	The Avett Brothers	I and Love and You	Rock
January Wedding	3:48	The Avett Brothers	I and Love and You	Rock
Head Full of Doubt / Road Full o...	4:48	The Avett Brothers	I and Love and You	Rock
And It Spread	4:07	The Avett Brothers	I and Love and You	Rock
The Perfect Space	4:31	The Avett Brothers	I and Love and You	Rock
Ten Thousand Words	5:36	The Avett Brothers	I and Love and You	Rock
Kick Drum Heart	2:54	The Avett Brothers	I and Love and You	Rock
Laundry Room	4:51	The Avett Brothers	I and Love and You	Rock
Ill With Want	4:05	The Avett Brothers	I and Love and You	Rock
Tin Man	3:08	The Avett Brothers	I and Love and You	Rock
Slight Figure of Speech	2:22	The Avett Brothers	I and Love and You	Rock
It Goes On and On	2:57	The Avett Brothers	I and Love and You	Rock
Incomplete and Insecure	2:36	The Avett Brothers	I and Love and You	Rock
I and Love and You	5:01	The Avett Brothers	I and Love and You...	Rock

- What is the main issue here?

Apple music database

Suppose you have an Apple database with 2 million songs on it. How would you identify the number of unique songs?

I and Love and You	5:01	The Avett Brothers	I and Love and You	Rock
January Wedding	3:48	The Avett Brothers	I and Love and You	Rock
Head Full of Doubt / Road Full o...	4:48	The Avett Brothers	I and Love and You	Rock
And It Spread	4:07	The Avett Brothers	I and Love and You	Rock
The Perfect Space	4:31	The Avett Brothers	I and Love and You	Rock
Ten Thousand Words	5:36	The Avett Brothers	I and Love and You	Rock
Kick Drum Heart	2:54	The Avett Brothers	I and Love and You	Rock
Laundry Room	4:51	The Avett Brothers	I and Love and You	Rock
Ill With Want	4:05	The Avett Brothers	I and Love and You	Rock
Tin Man	3:08	The Avett Brothers	I and Love and You	Rock
Slight Figure of Speech	2:22	The Avett Brothers	I and Love and You	Rock
It Goes On and On	2:57	The Avett Brothers	I and Love and You	Rock
Incomplete and Insecure	2:36	The Avett Brothers	I and Love and You	Rock
I and Love and You	5:01	The Avett Brothers	I and Love and You...	Rock

- How might you solve it without knowing any machine learning?
- ML buzzwords: record linkage, entity resolution, de-deduplication (types of clustering)

Electronic health records

Goal: identify patient risk for certain illness (kidney failure).

- Suppose there are 2 million patients.
- Suppose for now we ignore any time component.

1. Suppose you have access to Duke health records (patients) for vitals and labs.
2. Suppose you also have access to patient notes.

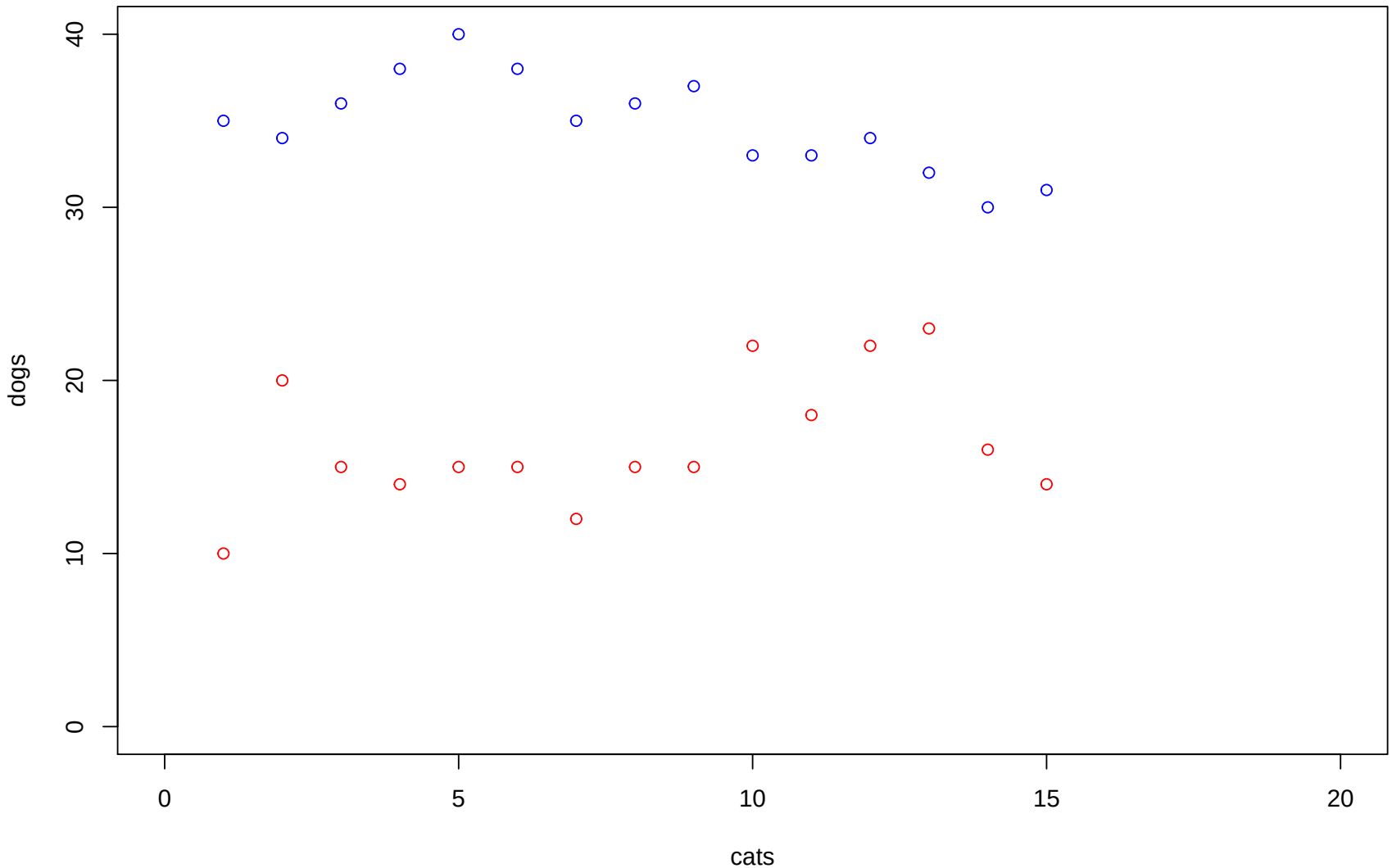
How would you identify at risk patients in real time?

Hacked Webpages

Out of all the webpages that exist, how could you try and identify in near-real time that a webpage has been hacked?

Classifying Objects

Given different kinds of animals, how might we classify each animal into its right class (cat, dog, mouse, etc)?



The Machine Learning (ML) culture

- It's fast.
- Papers are well written (and I expect your thoughts to be well written)!
- Code is reproducible (or it should be). And thus, I expect yours to be!
- In this course, we will look at standard techniques that are important to data mining and machine learning and are a good start for building your knowledge basis.

Machine Learning in Practice

https://www.youtube.com/watch?v=Nj2YSLPn6OY

The Book

- The book we will use is the required text (Intro to Statistical Learning with R)
- Please read all of it and follow it as we go.
- Please do the lab exercises in ISL on your own.
- For extra information that you might find interesting that is beyond this class, read the Rise of the Machine by Larry Wasserman.

The Lectures

- Lectures will be via slides, code, examples.
- They will be based on the book.
- Expect about 8–10 homeworks throughout the semester.
- All information will be posted on Sakai regarding deadlines.

The Homeworks

1. All code must be written to be reproducible in Markdown.
2. All derivations can be done in any format of your choosing (word, latex, markdown) but must be converted to a pdf document. It must be legible.
3. All files must be zipped together and submitted to Sakai as one file. Otherwise, you cannot submit. (Try this early to avoid not submitting anything).
4. Ask your TA questions early if you have a problem.
5. Your lowest homework will be dropped.

The Exams

1. All exams will be in class, closed book.
2. They will be cumulative as they go, but they will focus on the most recent material we have covered.
3. Make up exams will not be given.
4. You must take the second exam of the class to pass the class (this will be cumulative)!
5. You must also attend Datathon to pass the class.

Datathon

- Duke's first-ever datathon.
- Attack a dataset and come up with an analysis and visualization while racing against the clock.
- Top teams will win prizes, and all students will have the opportunity to interact with industry and academic sponsors.
- There will be additional bootcamps available to prepare for this event as part of the MLBytes seminar series.

Datathon

- Saturday, October 27
- `http://dukeml.org/datathon/`
- Students will work in pairs of three

Expectations:

1. Attend and complete Datathon
2. You'll need to prepare a 3 minute presentation for the class and a three page report (with your code) in the appendix. This should all be submitted to Sakai similar to a homework
3. Recommend having a plan of how you'll spend your 12 hours of Datathon

- What should your talk/presentation look like?

`https://events.technologyreview.com/video/watch/innovators-under-35-rebecca-steorts/`

Datathon

Suggestions for Datathon:

1. Slack may be a good way to communicate with your group
2. Github is a good place to store your code
3. Using a Google document to share your ideas.

- Test these out before the day of so that you don't have issues.

Course Material

1. Introduction (Today)
2. Statistical Learning
3. Information Retrieval
4. Linear Regression
5. Classification
6. Re-sampling Methods
7. Linear Model Selection and Regularization
8. Moving Beyond Linearity
9. Tree Based Methods
10. Support Vector Machines
11. Unsupervised Learning

Typos

- If you find a typo on a slide, please write it down neatly with the slide number and give it to me after class.
- I do my best to fix typos within 48 hours after the lecture with highlights in red so you can spot them.
- Thanks for helping me spot typos in advance!

Coding using R, RStudio, and Markdown

- In this class, I will assume that you are very familiar with R.
- If you need to refresh certain R skills, please do this on your own.

Reproducible Code

In this course, all code you turn in will be expected to be reproducible.

What does this mean?

Reproducible Code

- Suppose I write a lecture on linear regression with many plots explaining my analysis.
- You might wonder how exactly I created my analysis.
- If I simply write my code in a way such that you can reproduce my entire lecture, then you can verify all the results that I show you.
- Similarly, if you write your code in this way for your homeworks, then the TA and the I can also verify the results very quickly.

R versus RStudio

- RStudio mediates your interaction with R.
- RStudio is a driver of the emergergence of R Markdown, knitr, R + Github.

What is Markdown?

- Markdown is a lightweight markup language for creating PDF (or other documents).
- Markup languages produce documents from human readable text (and annotations).
- Some of you might be familiar with LaTeX (less friendly). This is another mark up language for creating PDF documents.

Why I like Markdown

1. It's very easy to learn.
2. The focus is on content rather than coding and debugging of errors.
3. Once you have the basics, you can get fancy!

(In fact, my slides right now are in markdown, so you can even make slides)!

R Markdown

This just means that you include R code in your **markdown** document.

```
x <- 2 + 2
x
```

```
## [1] 4
```

Installing RStudio and Markdown

- How do I install Rstudio?
- How do I include markdown?

Use the command

```
install.packages("rmarkdown")
```

More behind R Markdown

R Markdown files are the source code for rich, reproducible documents. You can transform an R Markdown file in two ways.

1. knit: You can knit the file.

- The rmarkdown package will call the knitr package.
- knitr will run each chunk of R code in the document and append the results of the code to the document next to the code chunk.
- This workflow saves time and facilitates reproducible reports.

In the R Markdown paradigm, each report contains the code it needs to make its own graphs, tables, numbers, etc. The author can automatically update the report by re-knitting.

More behind R Markdown

2. convert: You can convert the file.

- The rmarkdown package will use the pandoc program to transform the file into a new format. - For example, you can convert your .Rmd file into an HTML, PDF, or Microsoft Word file.
- You can even turn the file into an HTML5 or PDF slideshow.
- rmarkdown will preserve the text, code results, and formatting contained in your original .Rmd file.

Conversion lets you do your original work in markdown, which is very easy to use. You can include R code to knit, and you can share your document in a variety of formats.

More behind R Markdown

In practice, authors almost always knit and convert their documents at the same time. In this article, I will use the term render to refer to the two step process of knitting and converting an R Markdown file.

You can manually render an R Markdown file with

`rmarkdown::render()`.

Takeaways

1. Make sure that you understand the class policies.
2. Make sure you can access the Google group.
3. Make sure that you can access Sakai, the first homework assignment, and start your assigned readings for the course.
4. If you feel unprepared for the class, please speak with me now, rather than later.
5. Install RStudio and markdown. Use the lab with this lecture to make sure that you can do basic exercises in R and that they are reproducible. If you have trouble with the lab or first few homework assignments and exam, then please see me quickly to resolve issues.
6. Make sure you can access the lectures on Github.

Machine Learning at Duke

There is now a space for undergraduates to do ML at Duke!

1. Board of students + faculty advisor
2. We're taking applications for new student board members

Events: Datathon, fundraising for events, dinners, bootcamps, seminar series (MLBytes), ML day in the spring.

`http://dukeml.org/`

Data+

Data+ is a 10-week summer research experience that welcomes Duke undergraduates interested in exploring new data-driven approaches to interdisciplinary challenges. Many students work on machine learning related research problems.

Data+

"The most surprising thing I gained from Data+ was patience. Starting the project, I imagined an explosion of results around every corner. Beautiful graphs with clear trends and machine learning models with near-perfect performance. But then I realized it isn't like picking plump fruit off a tree. It's more like hoeing potatoes from the ground. You need the persistence to keep digging and the humility to get dirty. Only then you start getting those golden nuggets."

- 2018 Data+ Student