Statistical Learning

Welcome!

T. ARNOLD

Statistical Learning

Today we are going to get all of the administrative details dealt with. Here is a quick outline:

1. syllabus
2. course content
3. introductions
4. install course materials

There should be plenty of time for questions throughout the class.

Course Website

As I mentioned in my email, these notes and all others for this semester will be posted on the course website:

https://statsmaths.github.io/dsst389-s22/

1. Syllabus

Course Expectations

There are four things I expect from you this semester:

– Regularly attend and participate in class.
– Record daily attendance (or reasons for absence) using the class form.
– Complete and present four class projects throughout semester.
– Complete end-of-semester self assessment of your work.

Grading

I try to keep grading simple. You will get a grade out of 95 points (based on posted rubric) for each of the projects and another one for the self-assessment. The final grade is based on the average of these five scores.

Letter grades are assigned as follows: A (93–95), A- (90–92), B+ (87–89), B (83–86), B- (80–82), C+ (77–79), C (73–76), C- (70–72), and F (0–69).

Groups

The projects can be done in groups. I prefer these groups to stay fixed throughout the semester, but we can discuss changes if issues arise.

I strongly recommend working in a group of two people. Working alone or as a group of three is also okay.

I prefer that you all try to organize yourselves into these groups. Think about who you want to be in a group with; we will put them together in the near future.

Attendance

This is a course where it is very important to be present in class. I usually have a very strict attendance policy for 389, but that is hard this semester.

Instead, please fill out the class form on the website for each course meeting. If absent, please explain why. I will follow up with anyone with a warning if there are any issues.

If you need to miss the day you are supposed to present a project, I expect you to (1) email me before class and (2) send me your slides. Note that there is always the option of presenting remotely.

Schedule and Workload

The course schedule is posted on the website. I will make every attempt to follow this schedule. I tried hard to avoid projects being due during typical busy periods.

The workload for this class is not particularly heavy but note that it is inconsistent. Make sure you plan on carving out some time before the projects are due.

Getting Help

There is usually a lot of time in class and right after class to ask questions and get help with the course material.

I can answer quick questions by email. This is particularly helpful if you have a coding question.

Of course, I am also happy to set up a time to meet outside of class. I don't have fixed office hours, but generally am free to meet on Mondays and Wednesdays before 1:30pm and after 5:30pm. Just send me an email (ideally the day beforehand) with some times that work for you.

2. Course Content

Machine Learning?

Machine Learning vs Statistical Learning: different histories but the same thing

Machine Learning (ML) is a branch of artificial intelligence that uses data to create models. Methods mostly fall into three groups:

- **Supervised Learning**: detect patterns in order to make predictions about new data
- **Unsupervised Learning**: detect patterns in order to organize and structure data
- **Reinforcement Learning**: detect patterns in order to make complex decisions

These are not disjoint areas, though, many tasks require a mixture of methods.

Examples

Here are some examples of ML tasks. They are organized by the canonical type of machine learning that each is usually associated with, though the optimal approach may involve a mixture of methods.

Supervised Learning

- Predict the sale price of a house based on its size and location.
- Predict whether an email message should be put into a user's spam box.
- Find and identify all the faces found in a video feed.

Unsupervised Learning

- Cluster a collection of news paper articles by themes.
- Find and recommend similar products to users on a digital commerce website.
- Flag suspicious product reviews that should be manually investigated for fraud.

Reinforcement Learning

- Build an algorithm to play a game, such as checkers or chess, against a human.
- Determine a way to optimally schedule elevators in a large office building.
- Program a self-driving car.

Teaching ML

There are three different approaches to teaching machine learning:

Mathematical Approach Focus on theoretical properties of various methods, using the language of probability and numerical analysis.

CS/Engineering Approach Focus on implementation and performance of ML techniques and algorithms.

Data Science Approach Focus on the application of ML techniques in order to understand complex datasets.

This course will take the **Data Science Approach**.

What will you learn?

- understand the terminology of predictive and unsupervised ML methods
- how to apply a set of methods using the open-source R programming language
- how to use and understand a core set of general-purpose, interpretable ML methods
- how to use and understand several specific methods for working with textual data
- how to summarise and present the results of an exploratory analysis of data that integrates ML methods

What won't you (directly) learn?

- a laundry-list of dozens of ML methods
- theoretical justification/analysis of ML methods
- implementation details of ML methods
- a full introduction to R
- deep learning models

Project Oriented Class

In order to embody the data science approach, this course is centered around four projects.

The primary goal will be learning how to use machine learning algorithms to tell a story about data. We will get a lot of practice doing this.

Our primary focus will be on text analysis. In addition to being generally interesting and one of my research areas, it is application domain for a first course in machine learning.

Example Projects

Here are some digital projects that highlight what the data science approach to machine learning can do:

- **pixplot**: https://s3-us-west-2.amazonaws.com/lab-apps/pix-plot/index.html
- **addi:** https://statsmaths.github.io/addi_project/06_interactive_viz/build/?id=2014712029
- **signsat40:** http://signsat40.signsjournal.org/topic-model/

Project Format

The projects take the form of a short presentation. Rather than a textual write-up, I want you to focus on producing clean, professional slides for the presentation.

Here are the planned topics:

- **IMDb movie reviews**: predicting how many stars a movie review gives
- **Amazon product reviews**: predict the author of a review
- **Yelp reviews**: predict author of the review and cluster the corpus authors
- **Wikipedia**: detect themes in a subset of Wikipedia articles

Programming

This semester we will use the open-source programming language R. No prior experience with R is required. I have written a number of wrapper functions that minimise the amount of code you need to actually write.

However, if you are interested, there will be plenty of opportunities to dive into the wrapper functions and create your own takes on the code during the
semester.

Programming: Example

```
dsst_erate(model)
```

```
dsst_erate <- function(model, segmented = FALSE)
{
  if (segmented)
  {
    res <- model$docs %>%
      group_by(train_id, real_label) %>%
      summarize(
        class_rate = mean(pred_label != real_label)) %>%
      pivot_wider(
        values_from = class_rate, names_from = train_id
      )
  } else {
    res <- model$docs %>%
      group_by(train_id) %>%
      summarize(
        class_rate = mean(pred_label != real_label)) %>%
      pivot_wider(
        values_from = class_rate, names_from = train_id
      )
  }

  return(res)
}
```

3. Introductions

About Me

- From New England: born in Maine, school in MA, ME, CT
- Moved to Richmond in 2016
- Research on large text and image datasets in linguistics and cultural studies

Photogrammar

R Journal

ACLS

Distant Viewing Lab

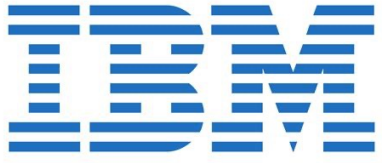
THE ANDREW W. MELLON FOUNDATION

NATIONAL ENDOWMENT FOR THE HUMANITIES

COLLEGIUM DE LYON
Université de Lyon

About Me

- Lots of industry experience in DS:
  - IBM (Healthcare)
  - Travelers (Insurance)
  - DARPA (social media)
  - AT&T (location analytics)
  - Telperian (pharmaceuticals)



About Me

- I have two Shih-Tzus: Roux and Sargent
- Roux is often in my office; please come say hello

4. Course Setup

[see other slides]