STAT 432: Basics of Statistical Learning

Introduction

Shiwei Lan, Ph.D. <shiwei@illinois.edu>

http://shiwei.stat.illinois.edu/stat432.html

January 14, 2019

University of Illinois at Urbana-Champaign

Course Info

Basics of Statistical Learning

- Instructor: Shiwei Lan <shiwei@illinois.edu>
  - Office: Illini Hall 103 C
  - Office hours: MW 1 - 2PM or by appointment
- Teaching Assistants:
  - Tianyi Qu <tianyiq3@illinois.edu>
  - Office hours: T 3 - 5PM at Illini Hall 104

Course Info

- Course Schedule, Lecture Notes, etc.

  http://shiwei.stat.illinois.edu/stat432.html

- Discussion and Questions

  Compass2g or Piazza?

- Homework and Grades on Compass2g

Course Info

- Textbook:
  - **Required**: [ISL] James, G., Witten, D., Hastie, T. Tibshirani, R. "An Introduction to Statistical Learning: With Applications in R". Springer. (free online PDF)
  - **Recommended**: [ESL] Hastie, T., Tibshirani, R. Friedman, J. "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". Springer. (free online PDF)
  - **Supplemental**: [R4SL] David Dalpiaz "R for Statistical Learning". (Link)
- Programming language:
  - R is mandatary
  - All homework reports must be submitted in either Word of PDF format, no other formats accepted.
  - R Markdown is strongly recommended to generate PDF report.

Course Info

- Sending email to me or TAs:
  - Please start with "Stat 432" in your email title.

- More about homework:
  - Don't wait till the last minute.
  - Late submission penalty: 3 points lost for each day of delay. Submission late for more than 3 days will NOT be accepted.
  - Simplify R output: No excessively long output. Total page limit.
  - You are encouraged to discuss with anyone, but DO NOT COPY each other. Otherwise, there will be severe penalties!

Statistical Learning Problems

Examples

Statistical learning is the process of extracting statistical regularities from datasets. They are motivated from real world problems. A few examples from HTF:

- Email Spam (classification)
- Handwritten Digits (classification)
- DNA microarray (clustering)

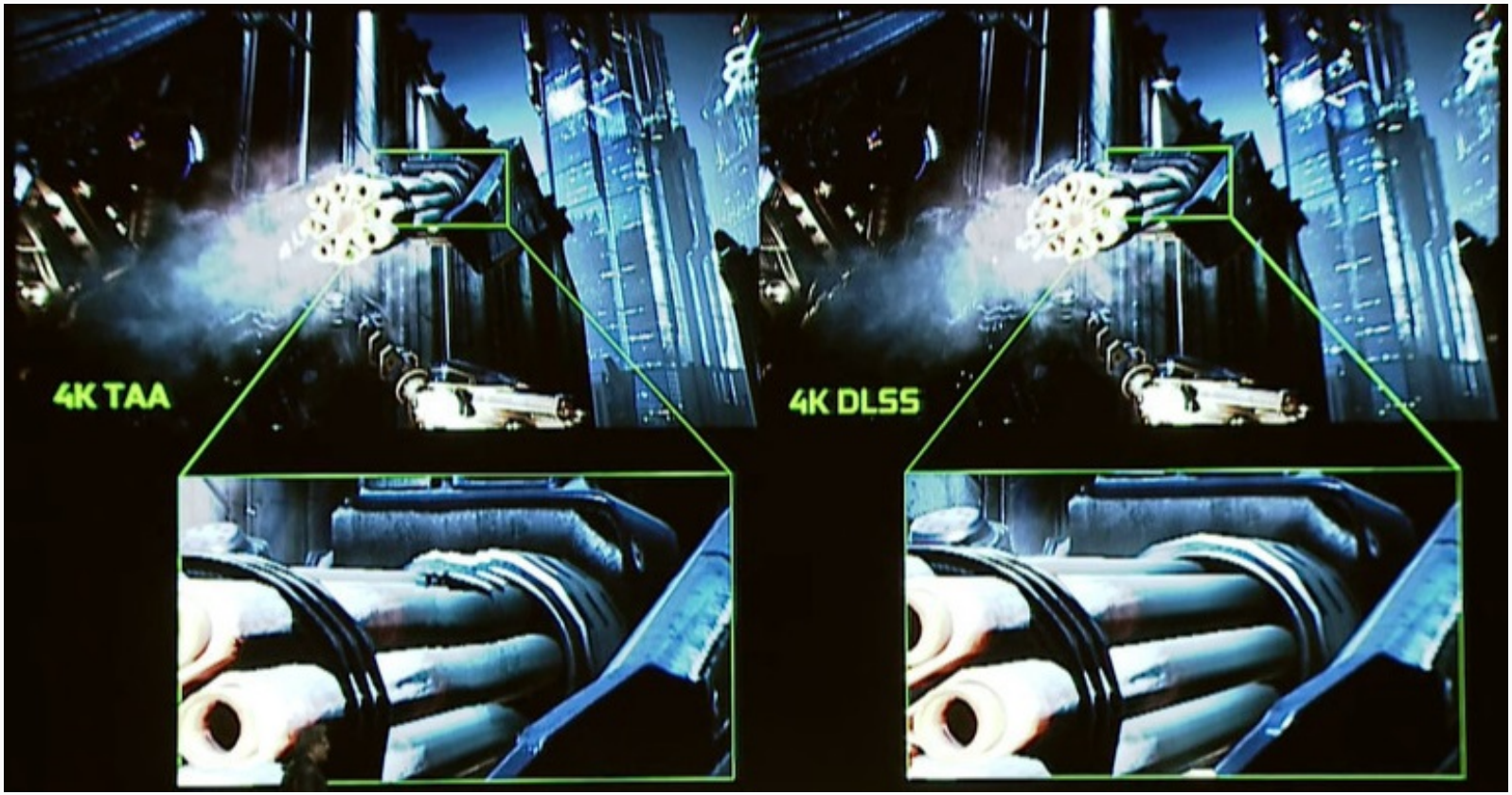
Examples

A compressive sensing method is used to decompose a surveillance video image into background and moving objects. Matrix decomposition and ℓ_1 penalization are used in this method.

Examples from Jiang, Hong, Wei Deng, and Zuowei Shen. “Surveillance video processing using compressive sensing.” arXiv preprint arXiv:1302.1942 (2013).

Examples

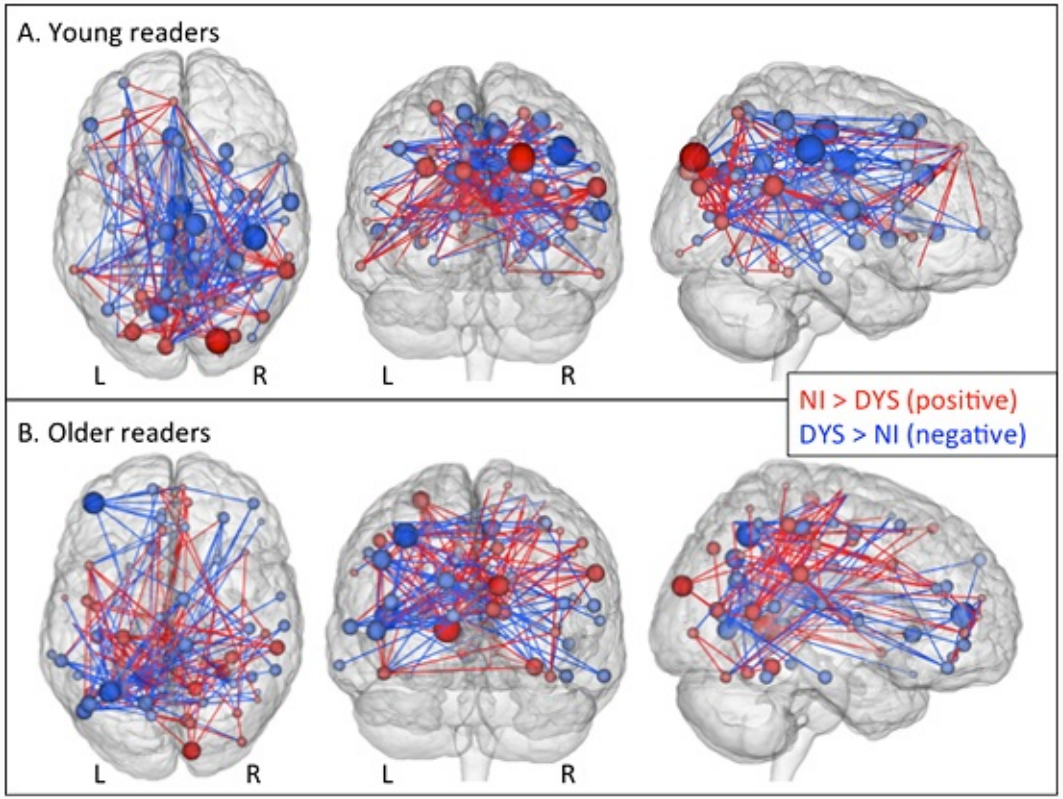
Single image super-resolution is used for sharping an image, or recovering a high-resolution image from a low-resolution one. Deep learning can be used to recognize the objects.



Example by NVIDA using deep learning super-sampling in computer games.

Examples

Graphical models can be used to understand the network connectivity of multiple brain regions. This information can help diagnose diseases such as parkinson and dyslexia (difficulty reading).



Example from Finn, Emily S., et al. “Disruption of functional networks in dyslexia: a whole-brain, data-driven analysis of connectivity.” Biological psychiatry 76.5 (2014): 397-404.

Examples

Auto piloting is made possible by recognizing objects such as lines, cars, pedestrians, etc. in real time images. These are essentially classification problems, carried out through deep learning.

Example from Tesla: https://www.tesla.com/autopilot; And the video of accident that killed a woman in Tempe, Arizona.

Examples

OpenAI Five is an AI system that plays Dota 2 against real human professionals. "... OpenAI Five sees the world as a list of 20,000 numbers which encode the visible game state, and chooses an action ...". They lose to human professional team 0:2 in TI 2018.

OpenAI Five: https://openai.com/five/;

Course Overview

Learn pattern from data

- Suppose we have a set of training data $\mathcal{D}_n = \{x_i, y_i\}_{i=1}^n$.
- Each x_i is a p-dimensional covariate vector that may represent
  - Gray scale of pixels in an image
  - Frequency of a particular word in an email
  - Concentration of a gene expression in blood
  - Brain electrical impulses
  - ...
- Each y_i is an outcome variable (or vector) that may represent
  - Cancer status (binary classification)
  - Type of object shown in the image (multicategory classification)
  - Surface temperature of a planet (regression)
  - Weight and volume of a tumor (multivariate regression)
  - ...
- Goal: learn patterns

Supervised vs. Unsupervised Learning

Oftentimes, there are two types of problems. And the goals are different:

- Supervised learning: data contains both “input” variables (covaraites) and “output” variable(s). And we want to use the inputs to predict the values of the outputs.
- Unsupervised learning: data contains only the “input” variables, and we want to know what is the underlying mechanism that generates the data.

Supervised Learning

- Response variable or outcome Y (a random variable)
  - Regression: Y is quantitative/continuous:

$$y_i \in \mathbf{R}$$

  - Classification: Y is categorical/discrete:

$$y_i \in C = \{0, 1\} \quad \text{or} \quad C = \{-1, +1\}$$

- Based on the training data $\mathcal{D}_n = \{x_i, y_i\}_{i=1}^n$, we aim to learn/estimate a function f, which describes the relationship between (variable) X and Y:

$$Y \longleftarrow f(X)$$

- Goals:
  1. achieve small prediction error
  2. (and/or) assess the effect of each features on Y

Unsupervised Learning

- No Y, just a set of features $X \in \mathbf{R}^p$
- Training data $\mathcal{D}_n = \{x_i\}_{i=1}^n$ where each $\mathbf{x}_i$ is a random draw of X
- Goal: find patterns in the data (such as clusters), understand the data generating process of X, etc.
- Sometimes, it is difficult to measure the performance of an unsupervised leaning method.
- The goal is fuzzy, but nevertheless it is an important problem.

Types of Learning Problems

- Supervised learning with labeled data: regression or classification
- Unsupervised learning with unlabeled data: clustering, network, graphical model
- There is also a semi-supervised learning that use a large set of unlabeled data to learn the distribution of independent variables and utilize that information for analyzing a smaller set of labeled data. However, we will not cover that in this course.

Prerequisites

Prerequisites

- Probability: probability and random variables, distributions
- Statistics: estimators, likelihood, linear regressions
- Mathematics: linear algebra and calculus
- Programming and software:
  - programming in R
  - use R Markdown for homework (covered in week 1 - 2)
  - optimization basics

Course Overview

Course Overview

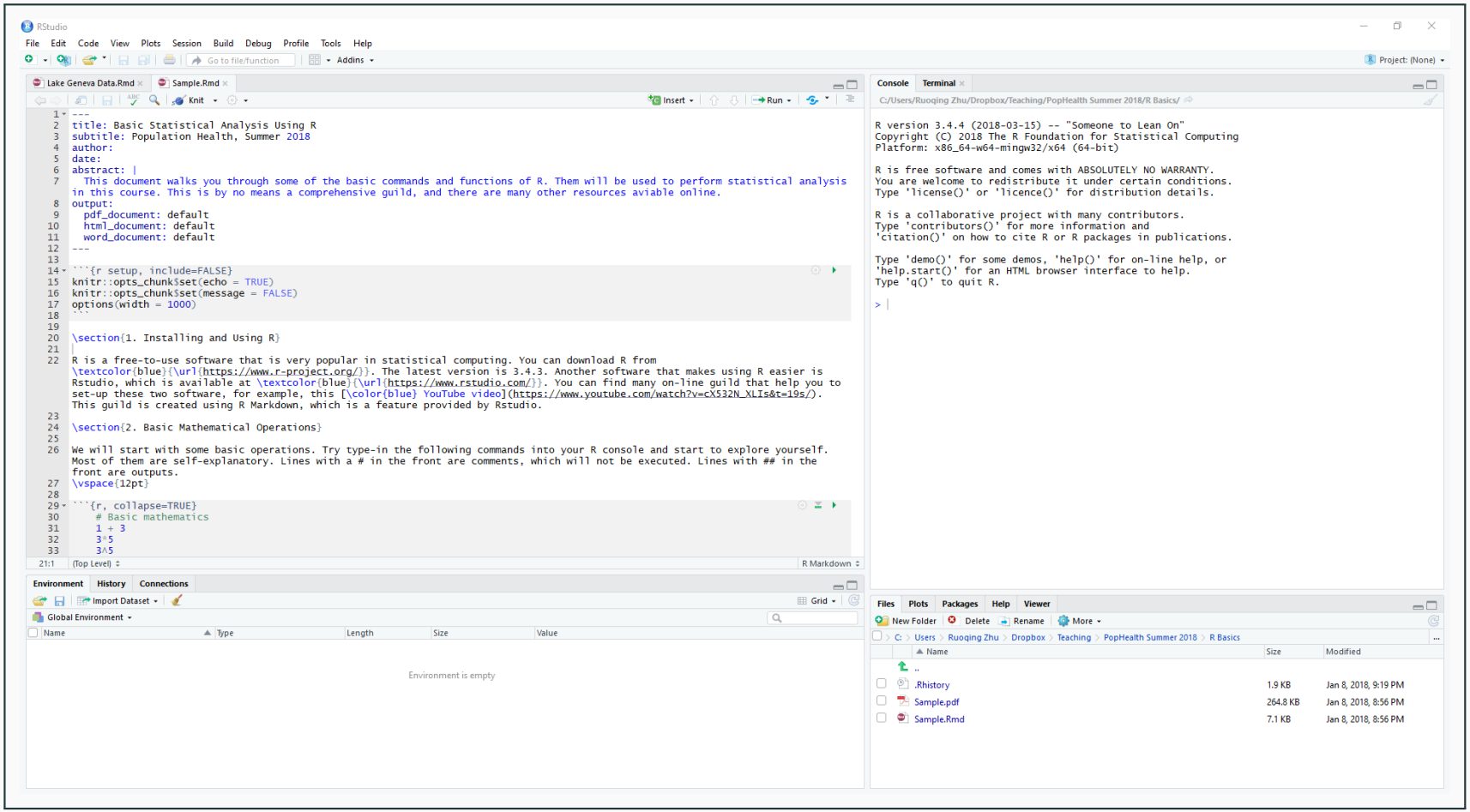
- Unsupervised Learning:
  - PCA, K-mean and hierarchical clustering
- Supervised Learning:
  - Linear models and penalization
  - Discriminant analysis, Naive Bayes
  - K nearest neighbor, tree, random forests
  - Support vector machine, neural networks
- Other concepts
  - Bias-variance trade-off
  - Variable selection
  - Cross-validation

R, RStudio and R Markdown

R, RStudio and R Markdown

- R is a free software environment for statistical computing and graphics.
  — https://cran.r-project.org/
- RStudio is an integrated development environment (IDE) for R.
  — https://www.rstudio.com/
- R packages are “add-ons” for R that offers additional datasets and functionalities. They can be managed within RStudio.
- R Markdown is a feature provided in RStudio that can integrate text with R code and outputs, and generate nice looking reports.

RStudio


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Lake Geneva Data.Rmd
Sample.Rmd
Knit
Insert
Run
---
title: Basic Statistical Analysis Using R
subtitle: Population Health, Summer 2018
author:
date:
abstract: |
This document walks you through some of the basic commands and functions of R. Them will be used to perform statistical analysis in this course. This is by no means a comprehensive guild, and there are many other resources aviable online.
output:
pdf_document: default
html_document: default
word_document: default

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(message = FALSE)
options(width = 1000)
```
\section{1. Installing and Using R}
R is a free-to-use software that is very popular in statistical computing. You can download R from \textcolor{blue}{\url{https://www.r-project.org/}}. The latest version is 3.4.3. Another software that makes using R easier is Rstudio, which is available at \textcolor{blue}{\url{https://www.rstudio.com/}}. You can find many on-line guild that help you to set-up these two software, for example, this [\color{blue} YouTube video](https://www.youtube.com/watch?v=cX532N_XLIs&t=19s/). This guild is created using R Markdown, which is a feature provided by Rstudio.
\section{2. Basic Mathematical Operations}
We will start with some basic operations. Try type-in the following commands into your R console and start to explore yourself. Most of them are self-explanatory. Lines with a # in the front are comments, which will not be executed. Lines with ## in the front are outputs.
\vspace{12pt}
```{r, collapse=TRUE}
# Basic mathematics
1 + 3
3*5
3^5
21:1
(Top Level)
R Markdown
Console
Terminal
C:/Users/Ruoqing Zhu/Dropbox/Teaching/PopHealth Summer 2018/R Basics/
R version 3.4.4 (2018-03-15) -- "Someone to Lean On"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
>
Environment
History
Connections
Import Dataset
Grid
Global Environment
Name
Type
Length
Size
Value
Environment is empty
Files
Plots
Packages
Help
Viewer
New Folder
Delete
Rename
More
C: Users Ruoqing Zhu Dropbox Teaching PopHealth Summer 2018 R Basics
Name
Size
Modified
..
.Rhistory 1.9 KB Jan 8, 2018, 9:19 PM
Sample.pdf 264.8 KB Jan 8, 2018, 8:56 PM
Sample.Rmd 7.1 KB Jan 8, 2018, 8:56 PM
```

Week 1

- R comes with some standard mathematical and statistical functionalities, such as basic calculations, and regressions.
- We will walk through some basics, especially, for dealing with data.
- Then we will introduce R Markdown and show some examples for generating a report.
- Bring your laptop next lecture!