BIOS 362 Advances Statistical Inference
(Statistical Learning)

Matthew S. Shotwell, Ph.D.

Data Science Institute
Vanderbilt University
Nashville, TN, USA

January 6, 2020
My Bio

- Matthew (Matt) S. Shotwell, Ph.D.
- Assoc. Prof. in Biostatistics
- 9 years at VU/VUMC
- 70% Biomed. Research / 30% Teaching
- R user 10+ years
- Teach “Statistical Learning” (BIOS 8362); 5 years
- Hastie et al. *Elements of Statistical Learning*
Why study Machine Learning and Statistics?

Data science is **HOT**. From [glassdoor.com](http://glassdoor.com):

**Data Scientist Salaries in Nashville, TN Area**

14 Salaries  Updated Feb 20, 2018

**Average Base Pay**

$96,751 /yr

20% below national average

**Additional Cash Compensation**

<table>
<thead>
<tr>
<th>Average</th>
<th>$10,100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range</td>
<td>$3,441 - $23,078</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Low</th>
<th>$73K</th>
<th>$97K</th>
<th>$123K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Not enough reports to show salary distribution

**Salaries for Related Job Titles**

- **Data Analyst**: $58K
- **Data Scientist Intern**: $72K
- **Quantitative Analyst**: $74K
- **Senior Data Scientist**: $114K
Why study Machine Learning and Statistics?

source: https://towardsdatascience.com/introduction-to-statistics-e9d72d818745
Why study Machine Learning and Statistics?

source: https://everett.wsu.edu/majorsdegrees/data_analytics/
Why study Machine Learning and Statistics?

Impact!

Data scientists:
- Can contribute to almost any worthwhile effort
- Can have large-scale impact
- Are the first to “know”
- Provide crucial interpretation
Why study Machine Learning in R?

Employers are increasingly looking for data scientists and statisticians with experience using R. From indeed.com:
Why study Machine Learning in R?

Employers are increasingly looking for data scientists and statisticians with experience using R. From indeed.com:

Course Overview

- Syllabus:
- [http://biostat.mc.vanderbilt.edu/wiki/Main/CourseBios362](http://biostat.mc.vanderbilt.edu/wiki/Main/CourseBios362)
- Prediction vs. Inference
- Principles of Machine Learning
- Regression Modeling Methods
- Supervised Learning
- Unsupervised Learning
- Statistical Inference
Prediction vs. Inference

Generally two types of tasks:

- Prediction: Is this a picture of a cat or a dog?
- Inference: Does greenhouse gas affect global average temperature?
Prediction vs. Inference

Generally two types of tasks:

- Prediction: What’s the probability that Dr. Shotwell will survive the semester?
- Inference: How does student effort in class affect the happiness of instructors?
Prediction vs. Inference

Generally two types of tasks:

▶ Prediction: Predicting an outcome.
▶ Inference: Making inferences about an unknown structure, mechanism, or relationship: e.g., effect of an exposure on an outcome.

▶ Need to ask: What is my primary task?
▶ Need to ask: Can I do both at the same time?
▶ Need to ask: What to consider when doing prediction/inference? Principles of prediction/ML.
Machine learning

source: https://www.wordstream.com/blog/ws/2017/07/28/machine-learning-applications
Supervised learning

- Primarily used for prediction (versus inference)
- Have input (‘features’) AND output (‘target’)
- Create a model (‘learner’) using observed inputs and outputs
- Goal is to predict outputs from new inputs
- “Supervised” because both inputs and outputs to guide model

[Image: Supervised Learning Workflow]

FIGURE 1.1. Scatterplot matrix of the prostate cancer data. The first row shows the response against each of the predictors in turn. Two of the predictors, svi and gleason, are categorical.
Unsupervised learning

- Have only input, no output
- Discover organization or clustering of input
- Prediction or inference task?

source: https://www.leveredge.com/blogpost/machine-learning-course-iot
- Gene expression array
- Rows - tumor samples
- Cols - genes
- Green - overexpressed
- Red - underexpressed
- Similar samples?
- Similar genes?
Unsupervised learning

- Have only input, no output
- Dimension reduction of input

source:

https://hackernoon.com/a-1aymans-introduction-to-principal-components-2fca55c19fa0
(Statistical) Inference

- Populations, samples, sampling biases
- How methods and tools for inference relate to those for ML
- Fundamentals of frequentist statistics

source: https://mahritaharahap.wordpress.com/teaching-areas/inferential-statistics/