What is Statistical Learning?

Prof Wells

STA 295: Stat Learning

January 23rd, 2024

Outline

In today's class, we will. . .

Outline

In today's class, we will...

- Provide a brief overview of the field of statistical learning
- Practice articulating statistical learning research questions

Section 1

Stat Learning

What is Statistical Learning?

- At its core, stat learning is collection of tools to understand data
  - Stat learning data is often *large* (both many observations and many variables)

What is Statistical Learning?

- At its core, stat learning is collection of tools to understand data
  - Stat learning data is often *large* (both many observations and many variables)
  - Modern stat learning methods often involve models built with computer assistance

What is Statistical Learning?

- At its core, stat learning is collection of tools to understand data
  - Stat learning data is often *large* (both many observations and many variables)
  - Modern stat learning methods often involve models built with computer assistance
- Stat learning methods are often divided into *supervised* and *unsupervised* methods.

What is Statistical Learning?

- At its core, stat learning is collection of tools to understand data
  - Stat learning data is often *large* (both many observations and many variables)
  - Modern stat learning methods often involve models built with computer assistance
- Stat learning methods are often divided into *supervised* and *unsupervised* methods.
  - **Supervised**: involve building models to predict the value of one or more output variables
  - **Unsupervised**: seek to learn about relationships and structure within data

What is Statistical Learning?

- At its core, stat learning is collection of tools to understand data
  - Stat learning data is often *large* (both many observations and many variables)
  - Modern stat learning methods often involve models built with computer assistance
- Stat learning methods are often divided into *supervised* and *unsupervised* methods.
  - **Supervised**: involve building models to predict the value of one or more output variables
  - **Unsupervised**: seek to learn about relationships and structure within data
- Example: Suppose we collect information on Grinnell College applicant's grades, SAT scores, high school location, family income, extracurricular activities, and recommendation letters.

What is Statistical Learning?

- At its core, stat learning is collection of tools to understand data
  - Stat learning data is often *large* (both many observations and many variables)
  - Modern stat learning methods often involve models built with computer assistance
- Stat learning methods are often divided into *supervised* and *unsupervised* methods.
  - **Supervised**: involve building models to predict the value of one or more output variables
  - **Unsupervised**: seek to learn about relationships and structure within data
- Example: Suppose we collect information on Grinnell College applicant's grades, SAT scores, high school location, family income, extracurricular activities, and recommendation letters.
  - Can we predict whether the student would accept an offer of admission to Grinnell?

What is Statistical Learning?

- At its core, stat learning is collection of tools to understand data
  - Stat learning data is often *large* (both many observations and many variables)
  - Modern stat learning methods often involve models built with computer assistance
- Stat learning methods are often divided into *supervised* and *unsupervised* methods.
  - **Supervised**: involve building models to predict the value of one or more output variables
  - **Unsupervised**: seek to learn about relationships and structure within data
- Example: Suppose we collect information on Grinnell College applicant's grades, SAT scores, high school location, family income, extracurricular activities, and recommendation letters.
  - Can we predict whether the student would accept an offer of admission to Grinnell?
  - Can we estimate the average Grinnell GPA of a student?

What is Statistical Learning?

- At its core, stat learning is collection of tools to understand data
  - Stat learning data is often *large* (both many observations and many variables)
  - Modern stat learning methods often involve models built with computer assistance
- Stat learning methods are often divided into *supervised* and *unsupervised* methods.
  - **Supervised**: involve building models to predict the value of one or more output variables
  - **Unsupervised**: seek to learn about relationships and structure within data
- Example: Suppose we collect information on Grinnell College applicant's grades, SAT scores, high school location, family income, extracurricular activities, and recommendation letters.
  - Can we predict whether the student would accept an offer of admission to Grinnell?
  - Can we estimate the average Grinnell GPA of a student?
  - Are there patterns in the type of student who apply to Grinnell? Can we sort students into a certain number of distinct groups?

What is Statistical Learning?

- At its core, stat learning is collection of tools to understand data
  - Stat learning data is often *large* (both many observations and many variables)
  - Modern stat learning methods often involve models built with computer assistance
- Stat learning methods are often divided into *supervised* and *unsupervised* methods.
  - **Supervised**: involve building models to predict the value of one or more output variables
  - **Unsupervised**: seek to learn about relationships and structure within data
- Example: Suppose we collect information on Grinnell College applicant's grades, SAT scores, high school location, family income, extracurricular activities, and recommendation letters.
  - Can we predict whether the student would accept an offer of admission to Grinnell?
  - Can we estimate the average Grinnell GPA of a student?
  - Are there patterns in the type of student who apply to Grinnell? Can we sort students into a certain number of distinct groups?
  - Is it sufficient to describe applicants using fewer attributes?

Supervised Learning

- We will spend majority of semester investigating supervised methods (Why?)

Supervised Learning

- We will spend majority of semester investigating supervised methods (Why?)
- Supervised learning models often have several objectives:
  - **Explain**: understand which input variables affect the output variable, and how

Supervised Learning

- We will spend majority of semester investigating supervised methods (Why?)
- Supervised learning models often have several objectives:
  - **Explain**: understand which input variables affect the output variable, and how
  - **Predict**: accurately estimate the value of the output variable for unobserved cases

Supervised Learning

- We will spend majority of semester investigating supervised methods (Why?)
- Supervised learning models often have several objectives:
  - **Explain**: understand which input variables affect the output variable, and how
  - **Predict**: accurately estimate the value of the output variable for unobserved cases
  - **Assess**: quantify the quality of predictions and explanations

Supervised Learning

- We will spend majority of semester investigating supervised methods (Why?)
- Supervised learning models often have several objectives:
  - **Explain**: understand which input variables affect the output variable, and how
  - **Predict**: accurately estimate the value of the output variable for unobserved cases
  - **Assess**: quantify the quality of predictions and explanations
- Supervising learning is divided into two tasks, depending on output variable's type:

Supervised Learning

- We will spend majority of semester investigating supervised methods (Why?)
- Supervised learning models often have several objectives:
  - **Explain**: understand which input variables affect the output variable, and how
  - **Predict**: accurately estimate the value of the output variable for unobserved cases
  - **Assess**: quantify the quality of predictions and explanations
- Supervising learning is divided into two tasks, depending on output variable's type:
  - **Regression**: Predicting (estimating) numeric value of quantitative variables
  - **Classification**: Predicting (classifying) qualitative level of categorical variables

Supervised Learning

- We will spend majority of semester investigating supervised methods (Why?)
- Supervised learning models often have several objectives:
  - **Explain**: understand which input variables affect the output variable, and how
  - **Predict**: accurately estimate the value of the output variable for unobserved cases
  - **Assess**: quantify the quality of predictions and explanations
- Supervising learning is divided into two tasks, depending on output variable's type:
  - **Regression**: Predicting (estimating) numeric value of quantitative variables
  - **Classification**: Predicting (classifying) qualitative level of categorical variables
- Methods of model building and assessment differ between the two tasks.
  - Research goals and measures of success might also differ between the tasks

Exploration

Each of the data contexts below prompts a broad research goal. As a group, pick **ONE** context and brainstorm as many precise research questions as possible that seem worth investigating:

1. **Media** The New York Times is trying to understand the popularity of its different articles with the hope of using this understanding to improve the hiring of writers, as well as their web layout.
2. **Public Health** The Iowa Department of Health is trying to better understand the different health trajectories of people who have contracted certain illnesses to improve funding for health services.
3. **Politics** The campaign management team for a poiltical candidate is trying to better understand voter turnout in different regions of the country in preparation for an upcoming election campaign.
4. **Technology** A text messenger app developer is trying to develop a tool for auto-completing a partially typed word, as well as for predicting the next word in a sentence.

Reflection

Consider a few of your brainstormed questions.

- Which suggest supervised learning methods? Which suggest unsupervised learning methods?
- Which questions can be framed as regression tasks? Which can be framed as classification tasks?
- Do any questions not not clearly fall under any of these categories?