MS&E 226: Fundamentals of Data Science

Lecture 1: Introduction

Ramesh Johari
rjohari@stanford.edu

What is this class about?

Example data: Houses

Prices of a selected subset of houses in Saratoga County, New York in 2006.

1,728 observations on 16 variables (e.g., `price`, `lotSize`, `livingArea`, etc.)
Available via `mosaicData` package (background at `mosaic-web.org`).

```
> install.packages("mosaicData")
> library(mosaicData)
```

Example data: Houses

The `SaratogaHouses` dataset contains 16 columns, including:

`price`	price (in dollars)
`livingArea`	living area (in square feet)
`age`	age of house (in years)
`bedrooms`	number of bedrooms
`bathrooms`	number of bathrooms
`heating`	type of heating system
`newConstruction`	whether the house is new construction

Example data: Houses

We pick out a few columns to focus on:

```
> library(tidyverse)
> sh = SaratogaHouses %>%
  select(price,
         livingArea,
         age,
         bedrooms,
         bathrooms,
         heating,
         new = newConstruction)
```

Example data: Houses

```
> sh
       price livingArea age bedrooms bathrooms          heating new
1     132500        906  42        2       1.0         electric  No
2     181115       1953   0        3       2.5 hot water/steam  No
3     109000       1944 133        4       1.0 hot water/steam  No
4     155000       1944  13        3       1.5          hot air  No
5      86060        840   0        2       1.0          hot air Yes
6     120000       1152  31        4       1.0          hot air  No
...
```

A sample

The Saratoga County houses data set is an example of a *sample*:

Data we observe on a specific collection of units (in this case, a subset of houses in Saratoga County).

(Note that Saratoga County contains tens of thousands of housing units...)

The population

We use the sample to reason about a population: a larger "universe" of units, from which our sample was observed.

E.g., we might use the Saratoga County houses data set to reason about:

- All houses in Saratoga County
- Houses in upstate New York (kind of plausible...)
- Houses in the entire state of New York (less plausible...)
- Houses on the East Coast (even less plausible...)

Generalization

Our goal is to *generalize* conclusions from the sample to the population.

Broadly, data science is about developing a collection of tools that allow us to confidently generalize.

Relationships

In this course, we will focus on using the sample to understand *relationships*: how some variables are related to an *outcome*.

For example, in the housing data, we might be interested in using the sample to understand the relationship in the population between the other variables and the *price*.

Modeling relationships

Formally:

- Y_i, $i = 1, \ldots, n$: i'th observed (real-valued) *outcome*.
  $\mathbf{Y} = (Y_1, \ldots, Y_n)$
- X_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, p$: i'th observation of the j'th (real-valued) *covariate*.
  $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$.
  $\mathbf{X}$ is the matrix whose *rows* are $\mathbf{X}_i$.

Matrix $\mathbf{X}$ and vector $\mathbf{Y}$ notation

House data with this notation ($n = 1728$, $p = 6$):

```
> sh
      price livingArea age bedrooms bathrooms heating new
1        Y1        X11 X12      X13       X14     X15 X16
2        Y2        X21 X22      X23       X24     X25 X26
3        Y3        X31 X32      X33       X34     X35 X36
4        Y4        X41 X42      X43       X44     X45 X46
5        Y5        X51 X52      X53       X54     X55 X56
6        Y6        X61 X62      X63       X64     X65 X66
...
```

Names

Names for the Y_i's:
outcomes, response variables, target variables, dependent variables

Names for the X_{ij}'s:
covariates, features, regressors, predictors, explanatory variables, independent variables

Continuous variables

Variables such as `price` and `livingArea` are *continuous* variables: they are naturally real-valued.

To start we will consider outcome variables that are continuous (like `price`).
Note: even continuous variables can be constrained:

- Both `price` and `livingArea` must be positive.
- `bedrooms` must be a positive integer.

Categorical variables

Other variables take on only finitely many values, e.g.:

- `new` is `Yes` or `No` if the house is or is not new construction.
- `heating` is one of the following:
    - `electric`
    - `hot water/steam`
    - `hot air`

These are *categorical variables* (or *factors*).

The population model: A probabilistic view

Generalization views the population as a *probability distribution*.

The population model: A probabilistic view

Generalization views the population as a *probability distribution*.

- There is a probability distribution of $\vec{X} = (X_1, \ldots, X_p)$ in the population.

The population model: A probabilistic view

Generalization views the population as a *probability distribution*.

- There is a probability distribution of $\vec{X} = (X_1, \dots, X_p)$ in the population.
- And Y has a conditional probability distribution *given* $\vec{X}$.

The population model: A probabilistic view

Generalization views the population as a *probability distribution*.

- There is a probability distribution of $\vec{X} = (X_1, \ldots, X_p)$ in the population.
- And Y has a conditional probability distribution *given* $\vec{X}$.

Together, these give a *joint* distribution over $\vec{X}$ and Y: the *population model*.

From population to sample ... and back again

The idea behind this probabilistic viewpoint:

From population to sample … and back again

The idea behind this probabilistic viewpoint:

- The population model represents the distribution of features and outcomes in the *entire* population of interest (e.g., all houses in Saratoga County).
  We don't know this distribution!

From population to sample ... and back again

The idea behind this probabilistic viewpoint:

- The population model represents the distribution of features and outcomes in the *entire* population of interest (e.g., all houses in Saratoga County). **We don't know this distribution!**
- Throughout this course, we will assume the *sample* consists of *independent* random draws from the population. (Be careful though ... this isn't always a valid assumption!)

From population to sample ... and back again

The idea behind this probabilistic viewpoint:

- The population model represents the distribution of features and outcomes in the *entire* population of interest (e.g., all houses in Saratoga County).
  We don't know this distribution!
- Throughout this course, we will assume the *sample* consists of *independent* random draws from the population. (Be careful though ... this isn't always a valid assumption!)
- By reasoning about the sample, we can *infer* properties of population model, and in particular, the relationship between features and the outcome.

From population to sample ... and back again

The idea behind this probabilistic viewpoint:

- The population model represents the distribution of features and outcomes in the *entire* population of interest (e.g., all houses in Saratoga County).
  We don't know this distribution!
- Throughout this course, we will assume the *sample* consists of *independent* random draws from the population. (Be careful though ... this isn't always a valid assumption!)
- By reasoning about the sample, we can *infer* properties of population model, and in particular, the relationship between features and the outcome.

All of generalization involves reasoning about the population model using the sample.

The nemesis of generalization: Uncertainty

What makes generalization difficult?

If we are only studying a sample, we are *uncertain* about whether what we observe and conclude is due to true properties of the population distribution, or just due to "random chance" (since our data was randomly drawn from the population).

The hardest job of a data scientist is to "separate truth from chance": to reason rigorously about this uncertainty.

A key goal of MS&E 226 is to teach you how the methodology we use in data science grapples with this uncertainty.

Examples of generalization

What are some examples of generalization?

Examples of generalization

What are some examples of generalization?

- *Prediction*: Predicting the price of a new house drawn from the population

Examples of generalization

What are some examples of generalization?

- *Prediction*: Predicting the price of a new house drawn from the population
- *Inference*: In the population, what is the relationship between the number of bedrooms in a house, and the price of the house?

Examples of generalization

What are some examples of generalization?

- *Prediction*: Predicting the price of a new house drawn from the population
- *Inference*: In the population, what is the relationship between the number of bedrooms in a house, and the price of the house?
- *Causality*: Would adding a bedroom to a house increase its value?

Examples of generalization

What are some examples of generalization?

- *Prediction*: Predicting the price of a new house drawn from the population
- *Inference*: In the population, what is the relationship between the number of bedrooms in a house, and the price of the house?
- *Causality*: Would adding a bedroom to a house increase its value?

These are all questions that might be investigated using *the same data*.

Examples of generalization

What are some examples of generalization?

- *Prediction*: Predicting the price of a new house drawn from the population
- *Inference*: In the population, what is the relationship between the number of bedrooms in a house, and the price of the house?
- *Causality*: Would adding a bedroom to a house increase its value?

These are all questions that might be investigated using *the same data*.

MS&E 226 is about understanding the foundations and differences between these forms of generalization.

Why are these even different?

Before we dive in, let's pause to think a little bit about whether these are even different from each other.

Can we make good predictions without good inference, i.e., without understanding which covariates matter? Or without understanding causality, i.e., how covariates affect the outcome?

Example: Breast cancer risk and wealth

Consider the following story:

abc NEWS | U.S. | World | Politics | Entertainment | Health | Tech | ...

Breast Cancer Risk Associated With Wealth

By JOY VICTORY • Dec. 1, 2005

Share with Facebook | Share with Twitter

0 SHARES

Women who live in regions of the United States known as breast cancer "hot spots" may have an increased risk because of personal wealth and not pollution or electrical wires, researchers say.

Deborah Winn, a scientist with the National Institutes of Health, states in the December issue of the journal Nature Reviews Cancer that the most likely reason that women in certain communities -- such as Long Island or San Francisco -- have increased breast cancer risk is that those areas are populated by wealthy women. Winn's article analyzes a series of studies conducted by the Long Island Breast Cancer Study Project in New York.

These women tend to have children later, have fewer children, and are more likely to receive costly replacement hormone therapy -- all of which are linked to increased breast cancer risk.

Example: Breast cancer risk and wealth

What can we say?

- Wealth is predictive of breast cancer.

Example: Breast cancer risk and wealth

What can we say?

- Wealth is predictive of breast cancer.
- Breast cancer is predictive of wealth.

Example: Breast cancer risk and wealth

What can we say?

- Wealth is predictive of breast cancer.
- Breast cancer is predictive of wealth.
- The reason certain women have breast cancer is that they are wealthier.

Example: Breast cancer risk and wealth

What can we say?

- Wealth is predictive of breast cancer.
- Breast cancer is predictive of wealth.
- The reason certain women have breast cancer is that they are wealthier.
- The reason certain women are wealthier is that they have breast cancer.

Example: Breast cancer risk and wealth

What can we say?

- Wealth is predictive of breast cancer.
- Breast cancer is predictive of wealth.
- The reason certain women have breast cancer is that they are wealthier.
- The reason certain women are wealthier is that they have breast cancer.
- If wealth increases, then incidence of breast cancer increases.

Example: Breast cancer risk and wealth

What can we say?

- Wealth is predictive of breast cancer.
- Breast cancer is predictive of wealth.
- The reason certain women have breast cancer is that they are wealthier.
- The reason certain women are wealthier is that they have breast cancer.
- If wealth increases, then incidence of breast cancer increases.
- If we made everyone poorer, there would be fewer cases of breast cancer.

Example: Breast cancer risk and wealth

What can we say?

- Wealth is predictive of breast cancer.
- Breast cancer is predictive of wealth.
- The reason certain women have breast cancer is that they are wealthier.
- The reason certain women are wealthier is that they have breast cancer.
- If wealth increases, then incidence of breast cancer increases.
- If we made everyone poorer, there would be fewer cases of breast cancer.

Moral:

Prediction relies on correlation, not causation.

Example: Education and income

Economist David Card, in his paper "The Causal Effect of Education on Earnings":

> *In the absence of experimental evidence, it is very difficult to know whether the higher earnings observed for better educated workers are caused by their higher education, or whether individuals with greater earning capacity have chosen to acquire more schooling.*

Another example: Internet marketing

Suppose a customer sees multiple channels of advertising from you: a social media ad, a display ad, a promoted tweet, e-mail ad, etc..

At the time of placing ads, you have demographic information about the customer.

- *Prediction* asks: Will this customer purchase or not? How much is this customer going to spend?
- *Inference* asks: Which campaign is most responsible for the customer's spend?

Often you can make great predictions, even if you cannot infer the value of the different campaigns.[1]

[1] The latter problem is the *attribution* problem.

Learning goals

- Defining your goal (the objective)
- Frameworks to compare methods
- Understanding assumptions
- Defining and quantifying uncertainty
- An “index” beyond MS&E 226

Topic-specific learning goals

- *Prediction*: Understand the basics of machine learning and generalization to new samples.
- *Inference*: Reason about the data-generating population or system.
- *Causality*: Reason about when we can make cause-and-effect claims from data.

Along the way we will learn about a variety of methods in support of each of these goals.

What this course is not!

MS&E 226 is not a vocational course; it is a conceptual course.

What about AI?

It's a transformative time to be teaching. Important note:

You are strongly encouraged to use AI tools to help you learn for any part of this class, except (of course!) the midterm exam, the final exam, and your in-person oral presentation.

Organization

1. **Prediction** (3 weeks): Train-test-validate; cross validation; binary classification; using optimization to build predictive models (maximum likelihood; linear and logistic regression; regularization, lasso, and ridge; other methods); model complexity and the bias-variance decomposition.

Organization

1. **Prediction** (3 weeks): Train-test-validate; cross validation; binary classification; using optimization to build predictive models (maximum likelihood; linear and logistic regression; regularization, lasso, and ridge; other methods); model complexity and the bias-variance decomposition.
2. **Inference** (3 weeks): Sampling distributions; p-values, confidence intervals, and hypothesis testing; application to linear and logistic regression; bootstrap; multiple hypothesis testing; post-selection inference.

Organization (continued)

3. **Causality** (3 weeks): The Rubin causal model, potential outcomes, and counterfactuals; randomized experiments; causal inference from observational data.

Organization (continued)

3. **Causality** (3 weeks): The Rubin causal model, potential outcomes, and counterfactuals; randomized experiments; causal inference from observational data.
4. **Bayesian statistics and decision-making** (1 week): Basics of Bayesian statistics; priors and posteriors; Bayesian vs. frequentist statistics; a Bayesian approach to decision-making.

Who is this class for?

- Targeted as a *first course* in data science (machine learning, statistical inference, and causality).
- Students with either deep backgrounds in one of machine learning *or* statistics tend to benefit from seeing both treated on a common footing, though there may be some redundancy in technical concepts with things you've seen before. You should decide whether the redundancy is worth the conceptual unification.
- Students with deep backgrounds in machine learning *and* statistics should probably not take this class.