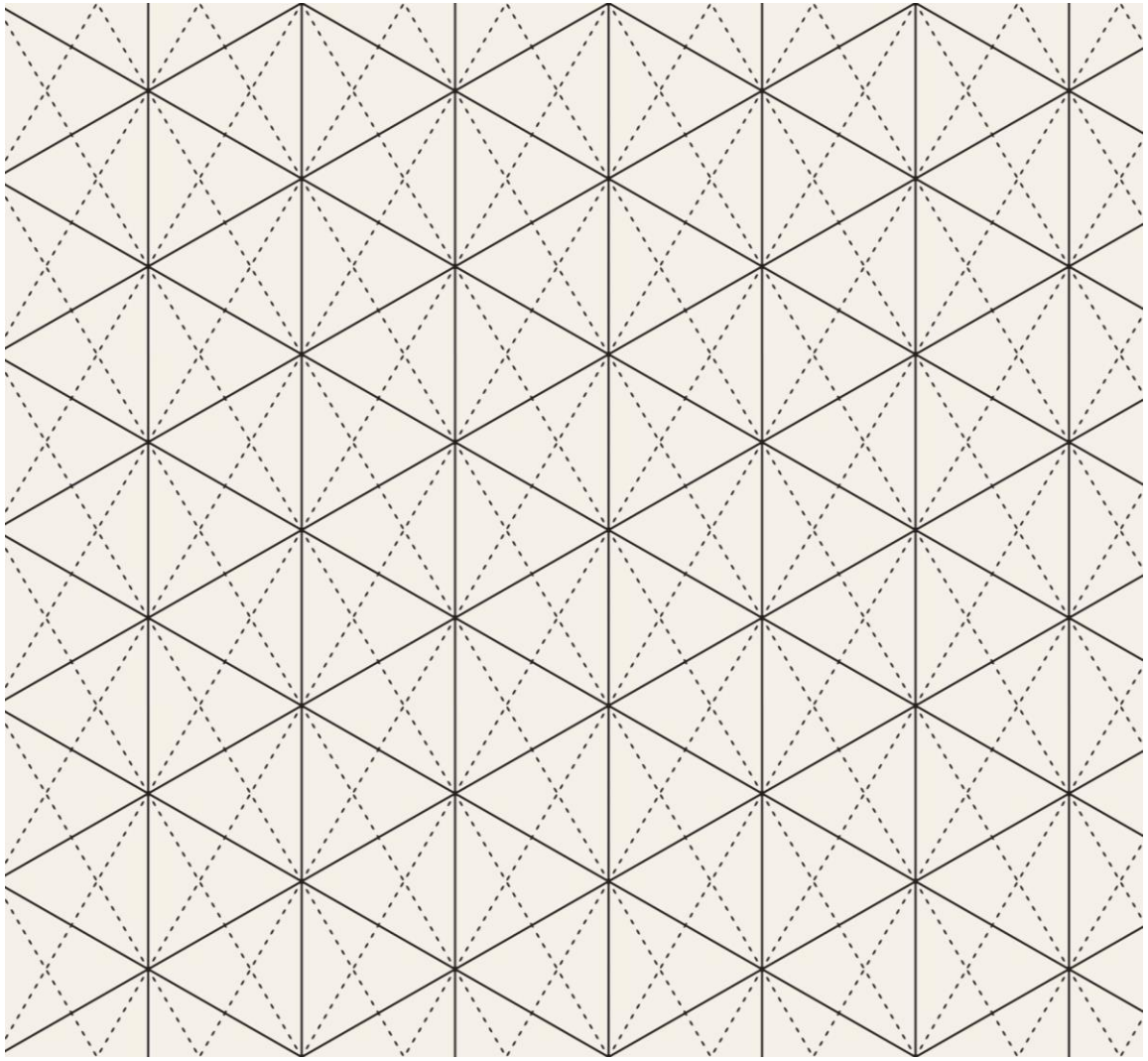


Introduction to Statistical Learning

ECON 258

PREPARE BY ARDINA HASANBASRI

Outline

1) Important Definitions and Terminology

- statistical learning versus machine learning

- parametric versus non-parametric

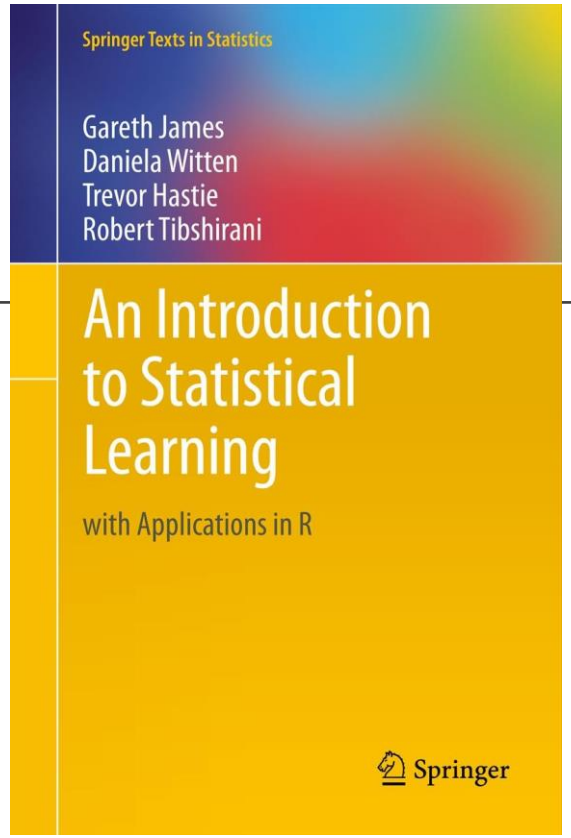
- training versus test data

- supervised/unsupervised

- continuous vs categorical problems

2) Prediction versus Causality

3) Coding Exercise with Machine Learning Flavor



- Our textbook for the ML part of the course.
- Pdf freely available online and is on canvas

1.1 Definition of Machine Learning

From the famous machine learning course by Andrew Ng on Coursera

- Arthur Samuel (1959): “Field of study that gives computers the ability to learn without explicitly being programmed.”

- Tom Mitchell (1998): “A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T as a measure by P improves with experience.”

1.2 Statistical Learning vs Machine Learning

- Statistical learning has to do with the relationship between different variables.

- Machine learning encompasses so much more:
  for example, determining whether an email is a spam or not

- Techniques are pretty much the same!

1.3 Statistical Learning using ML Definition

- We want to **TRAIN** our "machine" using data (**E: experience**)
- The machine should come-up with a rule evaluated by (**P: performance**)
- P will be used to predict an outcome variable (**T: task**)

To some extent, you have been doing this already in econometrics classes by estimating a function f(.).

Some Definitions:

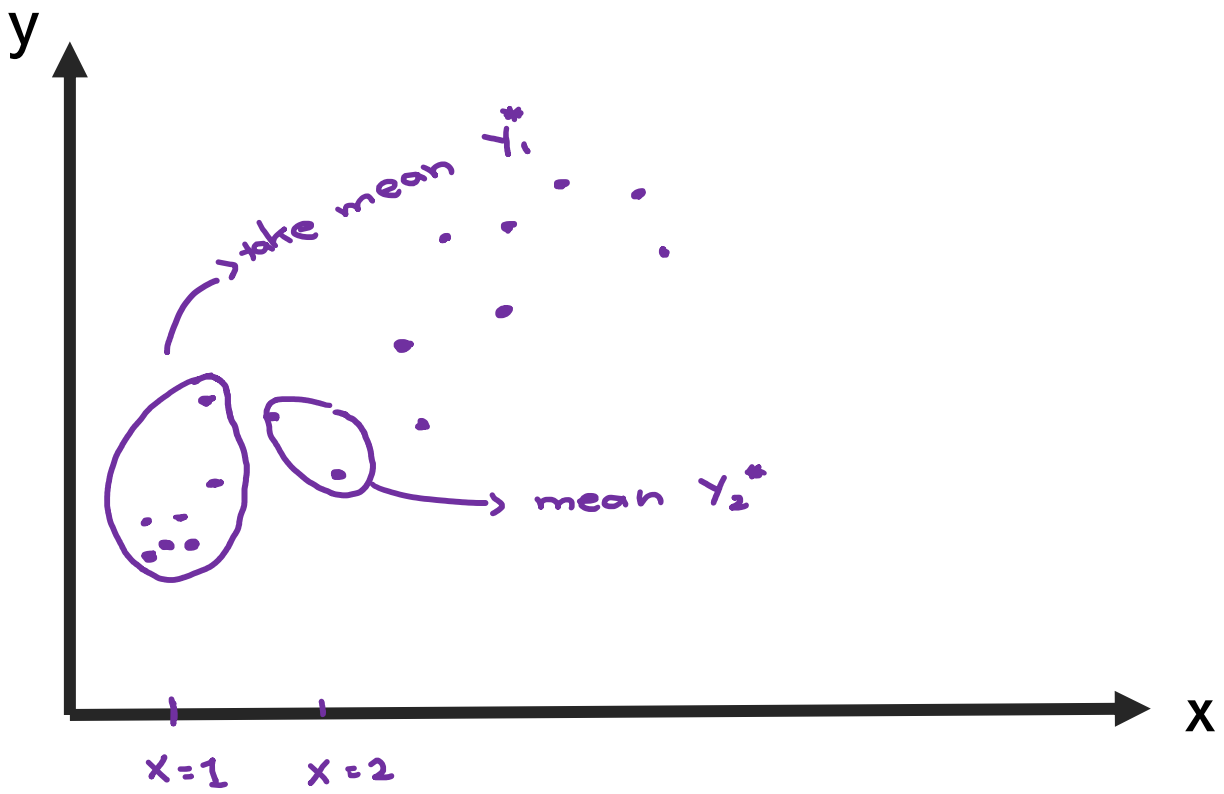
1.4 Parametric vs Non-Parametric

- Parametric requires you to make an assumption on the functional form of f(x).
- Non-parametric does not rely on assumptions.

1) $$wages_i = \alpha + \beta schooling_i + \gamma \mathbf{X_i} + \epsilon_i$$

2) $$wages_i = f(schooling_i, \mathbf{X_i})$$

1.5 Example of Non-Parametric: K Nearest Neighbor



Basic intuition:

- For each x, I estimate f(x) by finding the nearest k observation of y and take the mean of it. This mean is my estimate of x.

$f(x=1) = Y_1^*$

$f(x=2) = Y_2^*$

$f(x=3) = Y_3^*$

1.6 How to choose one over the other?

1) Data availability
   - non-parametric requires a lot of data

2) Interpretability versus Accuracy
   - parametric can be interpreted much easily
   - non-parametric more accurate because they are more flexible

More Definitions:

1.7 Training/Learning Data and Test Data

Remember that the machine needs to learn from data, then needs to find a way to evaluate the performance of the machine.

Common practice:

- Estimate a model using training data
- Then check whether model predicts well using test data
- For OLS, for example, we can look at the mean square error of the model on the test data.

1.8 Supervised vs Unsupervised Learning

When we want to understand the relationship between inputs and outputs, what we do is **supervised learning.** There is an associated y for each observation that we care about.

Unsupervised looks at relationship between observations. There is no particular y that we care about. Instead, we want the machine to categorize or group the different observation to find those who are similar.

1.9 Unsupervised Learning Example

- What happens when you do not have a label for your “primary variable”?

  Usually, we have an output variable:
  - **wages** depend on a, b, c
  - **individual decision** depends on d, e, f

- What type of questions concerns no labels?
- Can you group observations based on these patterns thus creating the labels?

Clustering Mechanism

Clustering algorithms can be used to group observations based on similarity.

1) Hierarchical Clustering

2) K-Means Clustering

Type of Questions:

- Finding similarities in gene sequencing (we'll do a coding example with this!)

- Finding networks (individuals with the same mutual friends)

Example: Grouping Landowners in Sub-Saharan Africa

Motivation:

There are some evidence that owning property rights relates to measures of poverty and economic growth.

Problem:

The idea of land ownership is quite nuance: right to bequeath, right to sell, right to improve, etc. There is also ability to decide on plot harvest and economic ownership.

Questions:

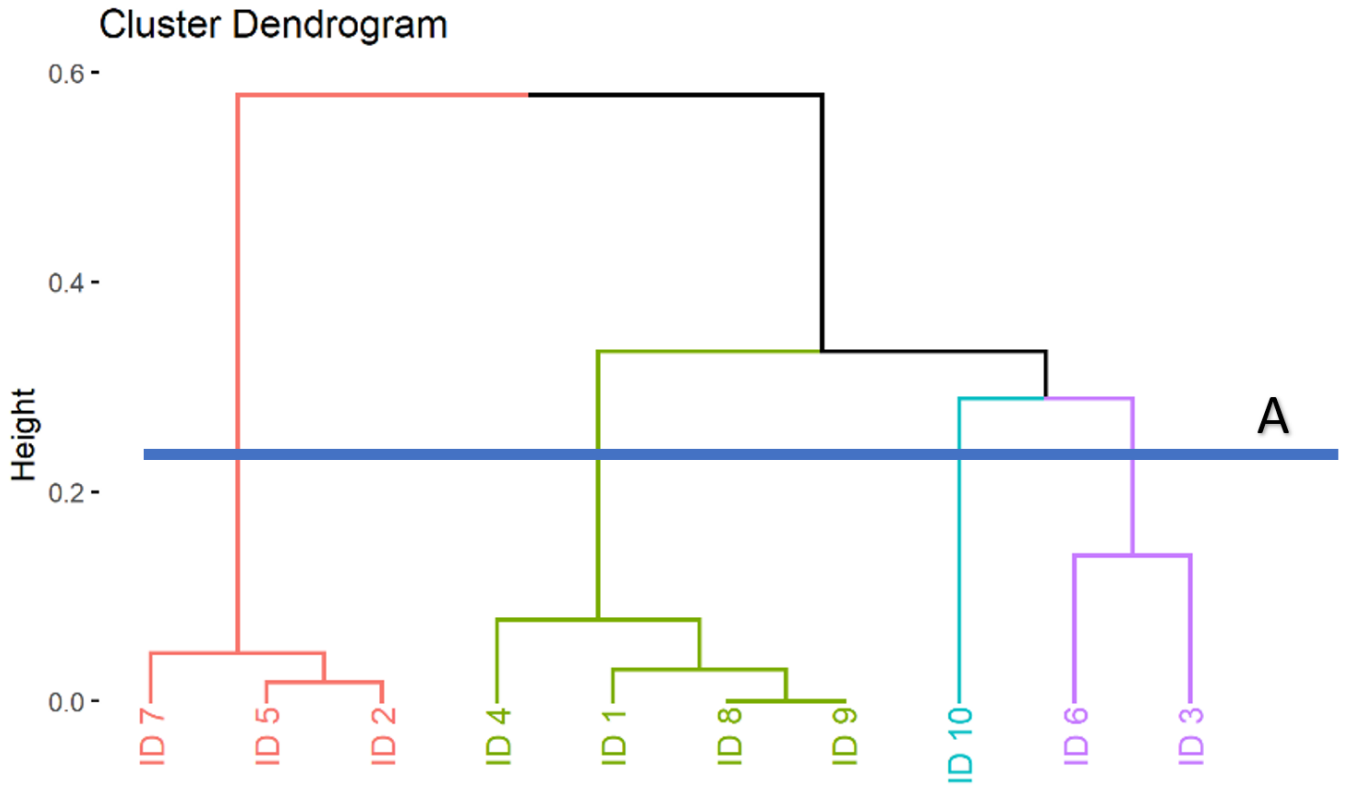
How do these right bundles correlate with each other?

Table 2: Sample Landowner for HCPC Illustration

	bequeath	**collateral**	**rent**	**improvement**	**plot decision**	**economic decision**	**document ownership**
ID 1	E	E	E	E	J	J	N
ID 2	J	J	J	J	J	J	E
ID 3	E	N	N	E	J	J	N
ID 4	E	E	E	N	J	J	E
ID 5	J	J	J	J	N	J	E
ID 6	N	N	N	E	N	J	E
ID 7	J	J	J	J	J	J	N
ID 8	E	E	E	E	J	J	E
ID 9	E	E	E	E	J	J	E
ID 10	E	N	E	N	N	N	E

Notes: The table randomly samples 10 individuals from Ethiopia and reports the ownership status on 7 land variables. The types of ownership status are exclusive (E), joint (J), and not hold (N).

Figure 7: Hierarchical Tree (Dendrogram) Illustration of 10 Landowners



Notes: The figure shows the resulting hierarchical tree (dendrogram) from the sample Ethiopia data of 10 individuals. Line A partitions the data into four clusters.

1.9 Continuous vs Categorical Data in Predictions

Disclaimer:

Be careful when dealing with continuous or categorical outcome (y) since the methods and interpretation differ slightly depending on which one you are using.

- Keep this in mind moving forward.

2.1 Causality versus Prediction

Causal inference methods require a lot of assumptions.

- Strict functional form assumption and on the distribution of the error term.
- Causal methods are not very good on prediction.

Machine learning methods are good on prediction.

- Sometimes unclear what is driving/causing y to change.
- No assumption is needed.

2.2 Difficulty of Causal Inference Methods

Difference-in-Differences

- Same time trend assumption
- Composition of the population over time does not change (to make sure fixed effect is eliminated).

Instrumental Variable

- Instrument cannot be weak.
- Instrument cannot affect outcome Y through ANY other channel except for variable we are instrumenting.

Propensity Score Matching

- If there is an omitted variable that determines treatment and not taken into account in matching, will lead to bias.

Regression Discontinuity Design

- Need to find a cutoff to find randomization.
- Can only calculate LATE

These are just some of the issues with causal inference models.

2.3 Benefits of Machine Learning?

1. Limited or free of assumptions
2. Some ML techniques such as random forest have intuitive ways of listing which variables are more important for prediction
3. Using a learning and test sample brings peach of mind

- One could argue, we should combine ML techniques with other econometric techniques.

3 Coding Exercise

Download R code “Intro to Statistical Learning Practice Code”, solutions also available already.

You will do a machine learning flavor exercise with fake data.

- Flavor in the sense that you will use training and test data and check how well the model does through MSE.

A toy model of y on x.

$$y = x^3 + \epsilon$$

with $\epsilon \sim N(0, 1)$.

Reference

James et. al. (2017) An Introduction to Statistical Learning with Applications in R, Springer Texts in Statistics