The Predictive Paradigm

Big Data y Machine Learning para Economía Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

Agenda

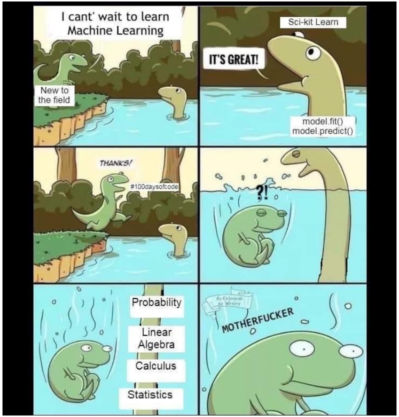
Agenda

¿Qué entendemos por Big Data y ML?

- ¿Que es Big Data?
  - Big n, es solo parte de la historia
  - Big también es big k, muchos covariates, a veces $n << k$
  - Vamos a entender Big también como datos que no surgen de fuentes tradicionales (cuentas nac., GEIH, etc)
    - Datos de la Web, Geográficos, etc.
- Machine Learning
  - Cambio de paradigma de estimación a predicción

Lenguajes

- Estadística y Econometría
- Inglés
- Código
  - Elijan el que quieran:
    - `R`, `Python`, o cualquier otro
    - no hay restricción
    - yo me basare en `R`
  - Github
  - Slack
- Aprender haciendo y mucha prueba y error!



Materiales

1. Bloque Neón
2. Statistical Learning (FREE!!! (as beer, not speech))
   https://www.gnu.org/philosophy/free-sw.en.html
   - James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (ISLR)
   - Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction
3. Libros de econometría
   - Davidson, R., & MacKinnon, J. G. (2004). Econometric theory and methods
   - Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data.
   - Hayashi, F. (2000). Econometrics

Agenda

1. About the Course
2. Machine learning is all about prediction
3. Prediction vs Causality
4. Prediction and loss functions

Machine learning is all about prediction

- Machine learning is a branch of computer science and statistics, tasked with developing algorithms to predict outcomes y from observable variables x.
- The learning part comes from the fact that we don't specify how exactly the computer should predict y from x. This is left as an empirical problem that the computer can "learn".
- In general, this means that we abstract from the underlying model, the approach is pragmatic

Machine learning is all about prediction

- Machine learning is a branch of computer science and statistics, tasked with developing algorithms to predict outcomes y from observable variables x.
- The learning part comes from the fact that we don't specify how exactly the computer should predict y from x. This is left as an empirical problem that the computer can "learn".
- In general, this means that we abstract from the underlying model, the approach is pragmatic

"Whatever works, works...."

“Whatever works, works....”

“Whatever works, works....”????

- In many applications, ML techniques can be successfully applied by data scientists with little knowledge of the problem domain.
- For example, the company Kaggle hosts prediction competitions (www.kaggle.com/competitions) in which a sponsor provides a data set, and contestants around the world can submit entries, often predicting successfully despite limited context about the problem.

“Whatever works, works....”????

- However, much less attention has been paid to the limitations of pure prediction methods.
- When ML applications are used “off the shelf” without understanding the underlying assumptions or ensuring that conditions like stability are met, then the validity and usefulness of the conclusions can be compromised.
- A deeper question concerns whether a given problem can be solved using only techniques for prediction, or whether statistical approaches to estimating the causal effect of an intervention are required.

Agenda

Policy Prediction Problems

- Empirical policy research often focuses on causal inference.
- Since policy choices seem to depend on understanding the counterfactual— what happens with and without a policy—this tight link of causality and policy seems natural.
- While this link holds in many cases, there are also many policy applications where causal inference is not central, or even necessary.

The Causal Paradigm

$$y = f(X) + u \tag{1}$$

- Interest lies on inference
- "Correct" $f()$ to understand how y is affected by X
- Model: Theory, experiment
- Hypothesis testing (std. err., tests)

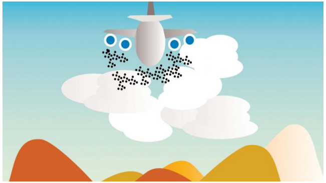
The Predictive Paradigm

$$y = f(X) + u \tag{2}$$

- Interest on predicting y
- "Correct" $f()$ to be able to predict (no inference!)
- Model? We treat $f()$ as a black box, and any approximation $\hat{f}()$ that yields a good prediction is good enough (*Whatever works, works.*).

Prediction vs. Causality

Prepare



Influence

Prediction vs. Causality

Prepare

- A loan officer wants to know the likelihood of an individual repaying a loan based on income, employment, and other characteristics.

Influence

- A mortgage lender wants to know if direct debit will increase loan repayments.

Prediction vs. Causality

Prepare

- In order to decide whether to invest in a start-up, an investor needs to know how likely the start-up is to succeed, given the entrepreneur's experience and the characteristics of the industry.

Influence

- An entrepreneur needs to know what the effect of receiving funding from a private equity investor (rather than getting a loan) is on the ultimate success of an enterprise.

Prediction vs. Causality

Prepare

- A home seller wants to know what price homes with the characteristics of his or her home typically sell for.

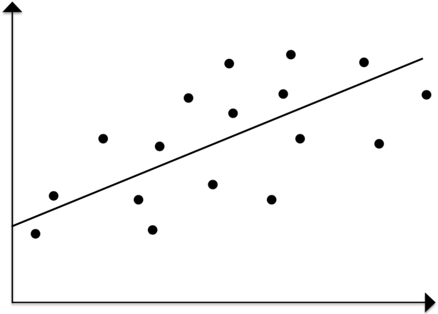
Influence

- A home seller wants to know by how much installing new windows will raise the value of his or her home.

Prediction vs. Causality: Target

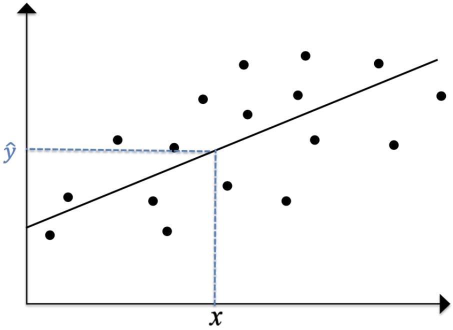
$$y = f(x) + \epsilon \tag{3}$$

$$y = \alpha + \beta x + \epsilon \tag{4}$$



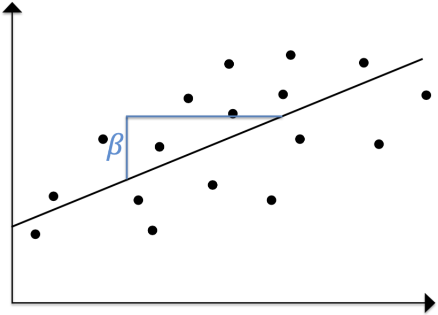
Prediction vs. Causality: Target

$$y = \underbrace{\alpha + \beta x}_{\hat{y}} + \epsilon \tag{5}$$



Prediction vs. Causality: Target

$$y = \alpha + \beta x + \epsilon \tag{6}$$



Prediction vs. Causality: The garden of the parallel paths?

- We've seen that prediction and causality
  - Answer different questions
  - Serve different purposes
  - Seek different targets
- Different strokes for different folks, or complementary tools in an applied economist's toolkit?

Agenda

Getting serious about prediction

$$y = f(X) + u \tag{7}$$

- Interest on predicting Y
- Model? We treat $f()$ as a black box, and any approximation $\hat{f}()$ that yields a good prediction is good enough (*"Whatever works, works..."*).
- How do we measure "what works"?

Getting serious about prediction

$$y = f(X) + u \tag{7}$$

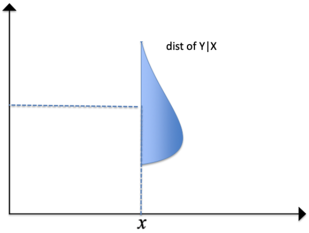
- Interest on predicting Y
- Model? We treat $f()$ as a black box, and any approximation $\hat{f}()$ that yields a good prediction is good enough (*"Whatever works, works..."*).
- How do we measure "what works"?
- Formal statistics can help figure out this: what is a good prediction.

Minimizing our losses

- Want our prediction to be "close" i.e. minimize the expected loss function
- Formally, a supervised learning algorithm takes as an input a loss function $L(\hat{y}, y)$ and searches for a function $\hat{f}$ within a function class $\mathcal{F}$ that has a low expected prediction loss

$$E_{(y,X)}[L(\hat{f}(X), y)] \tag{8}$$

on a new data point from the same distribution.



Minimizing our losses

- A very common loss function in a regression setting is the squared loss $L(d) = d^2$
- Under this loss function the expected prediction loss is the mean squared error (MSE)
- Can we find the function f^* within a function class $\mathcal{F}$ that has a low expected prediction loss?

Minimizing our losses

- By conditioning on X, it suffices to minimize the $MSE(f)$ point wise so

$$f(x) = argmin_{f^*} E_{Y|X}[(Y - f^*)^2 | X = x) \tag{9}$$

- f^* a random variable and we can treat f^* as a constant (predictor)

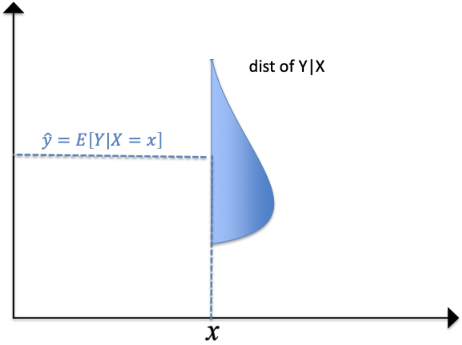
$$min_m E(Y - f^*)^2 = \int (y - f^*)^2 f(y) dy \tag{10}$$

- **Result**: The best prediction of Y at any point $X = x$ is the conditional mean, when best is measured using a square error loss

Minimizing our losses

- **Result**: The best prediction of y at any point $X = x$ is the conditional mean, when best is measured using a square error loss

$$f^* = E[y|X = x] \tag{11}$$



Minimizing our losses

- Prediction problem solved if we knew $f^* = E[y|X = x]$

Minimizing our losses

- Prediction problem solved if we knew $f^* = E[y|X = x]$
- But we have to settle for an estimate: $\hat{f}(x)$
- The EMSE of this

$$E(y - \hat{y})^2 = E(f(X) + u - \hat{f}(X))^2 \tag{12}$$

Review

- The predictive paradigm
  - Machine Learning is all about prediction
  - ML targets something different than causal inference, they can complement each other
- Next Class: Bias/Variance Trade Off and Linear Regression