Introducción

Big Data and Machine Learning for Applied Economics

Ignacio Sarmiento-Barbieri

Universidad de los Andes

August 5, 2025

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., etc)
    - Datos de la Web, Geográficos, etc.
- Machine Learning
  - Cambio de paradigma de estimación a predicción

Agenda

¿Qué es la máquina de aprender?

Aprendizaje de máquinas es todo sobre predicción

- El aprendizaje de máquinas es una rama de la estadística y la informática , encargada de desarrollar algoritmos para predecir los resultados y a partir de las variables observables X.
- La parte de aprendizaje proviene del hecho de que no especificamos cómo exactamente la computadora debe predecir y a partir de X.
- Esto queda como un problema empírico que la computadora puede ”aprender”.
- En general, esto significa que nos abstraemos del modelo subyacente, el enfoque es muy pragmático

¿Qué es la máquina de aprender?

Aprendizaje de máquinas es todo sobre predicción

- El aprendizaje de máquinas es una rama de la estadística y la informática , encargada de desarrollar algoritmos para predecir los resultados y a partir de las variables observables X.
- La parte de aprendizaje proviene del hecho de que no especificamos cómo exactamente la computadora debe predecir y a partir de X.
- Esto queda como un problema empírico que la computadora puede ”aprender”.
- En general, esto significa que nos abstraemos del modelo subyacente, el enfoque es muy pragmático

“Whatever works, works....”

“Whatever works, works....”

“Whatever works, works....”????

- En muchas aplicaciones, los científicos de datos pueden aplicar con éxito las técnicas ML con poco conocimiento del dominio del problema.
- Por ejemplo, el sitio web Kaggle organiza concursos de predicción (`www.kaggle.com/competitions`) en los que un patrocinador proporciona un conjunto de datos y los concursantes de todo el mundo pueden enviar entradas, a menudo prediciendo con éxito a pesar del contexto limitado sobre el problema.

“Whatever works, works....”????

- Sin embargo, cuando las aplicaciones ML se utilizan ”off the shelf” sin comprender los supuestos subyacentes o garantizar que se cumplan condiciones básicas las conclusiones pueden verse comprometidas. (Athey, 2017)
- Una pregunta más profunda (y dificil?) se refiere a si un problema dado se puede resolver usando solo técnicas de predicción, o si se requieren enfoques estadísticos para estimar el efecto causal de una intervención.

“Whatever works, works....”????

La primera victoria y derrota de ML

- Contexto ¿similar? al de 2020: Epidemia de la gripe A en 2009
- En EEUU la forma de monitorear es a través de reportes de la CDC
- La CDC agrega a nivel de región y a nivel nacional
- Todo esto llevaba aproximadamente 10 días $\rightarrow$ demasiado tiempo para una epidemia

Policy Prediction Problems

Google se ha unido a la conversación

- Google propuso un mecanismo ingenioso: **Google Flu Trends**
- Punto de partida:
  - Proporción de visitas semanales por Gripe A en hospitales
  - 9 regiones $\times$ 5 años (2003-2007) = $2,340$ datos
  - Estos son los datos que tomaban 10 dias en elaborarse
- Google cruzó estos datos con las búsquedas sobre la gripe A
- Con estos datos, construyeron un modelo para predecir intensidad de gripe A

Policy Prediction Problems

Google se ha unido a la conversación

- ¿Un sólo modelo?

Policy Prediction Problems

Google se ha unido a la conversación

- ¿Un sólo modelo?
- Los investigadores de Google estimaron **450 millones** de modelos
- Eligieron el que mejor predice sobe la intensidad de búsqueda
- Les permite tener información diaria, semanal o mensual para cualquier punto de EEUU y el mundo
- A Google le toma 1 día lo que a la CDC 10!

Policy Prediction Problems

Google se ha unido a la conversación

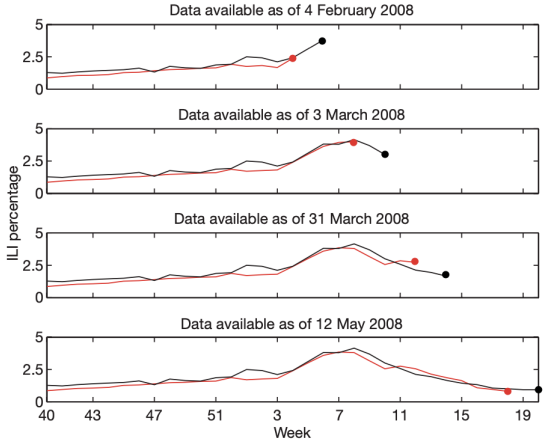


Figure 3 | ILI percentages estimated by our model (black) and provided by the CDC (red) in the mid-Atlantic region, showing data available at four points in the 2007-2008 influenza season. During week 5 we detected a sharply increasing ILI percentage in the mid-Atlantic region; similarly, on 3

Policy Prediction Problems

El rey ha muerto, larga vida al rey

- ¿Qué tienen en común Google Flu y Elvis?
  - Abanderados de la revolución
  - Definió y redefinió las reglas sistemáticas para hallar la solución a un problema
  - Éxito rotundo → Publicación en Nature!
    https://www.nature.com/articles/nature07634
  - Pero como a Elvis el éxito fue efímero
  - La predicciones comenzaron a sobre-estimar considerablemente la incidencia de la gripe A
  - Google Flu esta ahora archivado (disponible al público)
  - Continúa recolectando datos pero solo algunas instituciones científicas tienen accesso

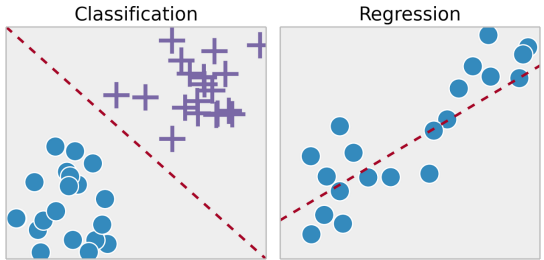
Agenda

ML branches

- ML tasks can (¿?) be divided into two main branches:
  1. Supervised Learning

ML branches

- Supervised Learning
  - for each predictor x_i a 'response' is observed y_i.
  - everything we have done in econometrics is supervised



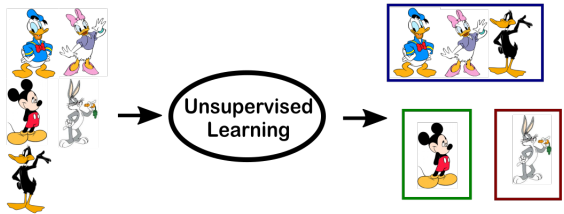
Source: shorturl.at/opqKT

ML branches

- ML tasks can (¿?) be divided into two main branches:
  1. Supervised Learning
  2. Unsupervised Learning

ML branches

- Unsupervised Learning
  - observed x_i but no response.
  - example: cluster analysis



Source: shorturl.at/opqKT

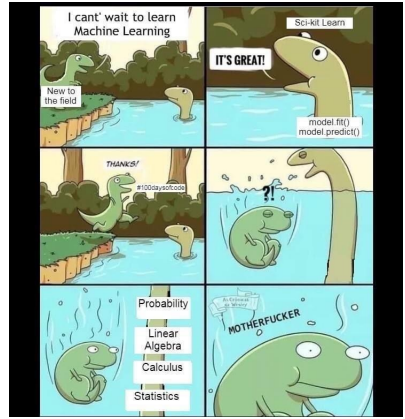
Agenda

Sobre el curso

- El aprendizaje automático en economía es muy nuevo y dinámico.
  - Las similitudes con la econometría plantea interrogantes:
    - ¿Estos algoritmos están simplemente aplicando técnicas estándar a nuevos y grandes conjuntos de datos?
    - Si hay herramientas empíricas fundamentalmente nuevas, ¿cómo encajan con lo que conocemos?
    - Como economistas empíricos, ¿cómo podemos utilizarlas?
- Estas clases darán un *"snapshot"* de este campo en evolución.
- Estudiaremos ML a través de ejemplos, centrándonos en algunas aplicaciones y algoritmos de ML.

Lenguajes

- Estadística y Econometría
- Inglés
- Código
  - Elijan el que quieran:
    - `R`, `Python`, o cualquier otro
    - no hay restricción
    - nosotros usaremos `R`
  - Github
- Materiales en BN
- Aprender haciendo y mucha prueba y error!



Evaluaciones

Table 1: Puntajes

	Puntaje Individual	Puntaje Total	Fecha entrega
Quices		28%	
Quiz 0*	0%		Agosto 12, 2025
Quiz 1	7%		Agosto 26, 2025
Quiz 2	7%		Octubre 7, 2025
Quiz 3	7%		Octubre 28, 2025
Quiz 4	7%		Noviembre 18, 2025
Talleres		60%	
Taller 1	20%		Septiembre 7, 2025
Taller 2	20%		Octubre 19, 2025
Taller 3	20%		Noviembre 23, 2025
Presentaciones		12%	
Taller 1	4%		Septiembre 9, 2025
Taller 2	4%		Octubre 21, 2025
Taller 3	4%		Noviembre 25, 2025

Nota: * *Opcional*

Complementarias

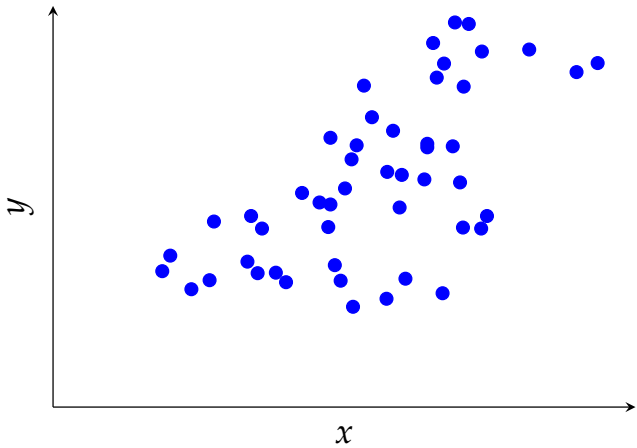
Agenda

Agenda

The basic logic of prediction

Agenda

Predicción vs Causalidad



Agenda

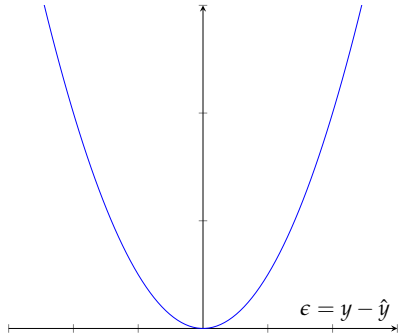
Prediction error

$$\epsilon = y - \hat{y} \tag{1}$$

Prediction error

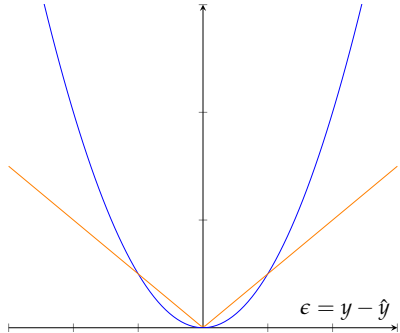
$$\epsilon = y - \hat{y} \tag{1}$$

$$L(y, \hat{y}) = (y - \hat{y})^2 \tag{2}$$



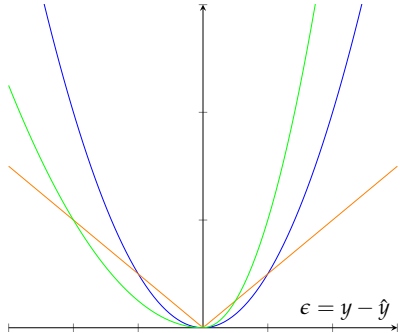
Minimizing our losses

$$L(y, \hat{y}) \tag{3}$$



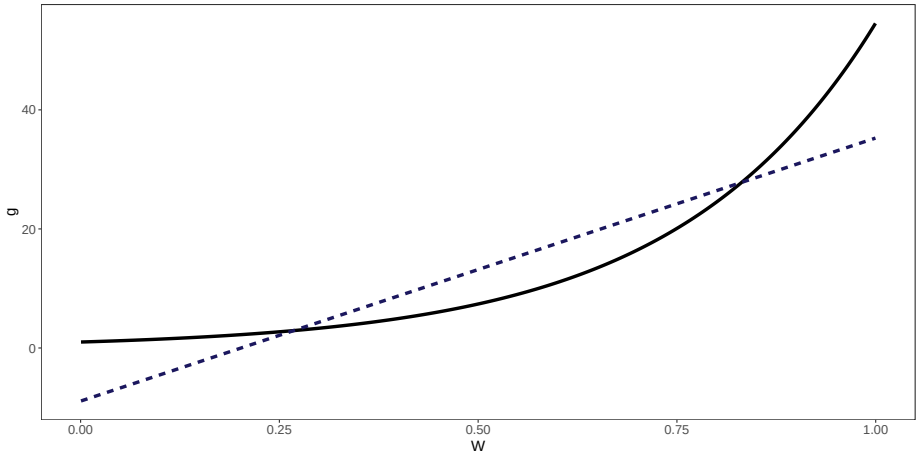
Minimizing our losses

$$L(y, \hat{y}) \tag{4}$$

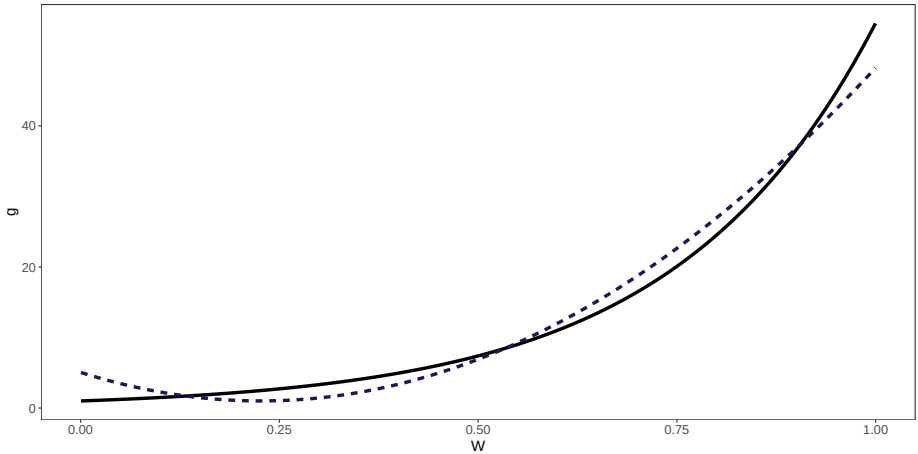


Can we find the best predictor of Y?

From Best Linear Predictor to Best Predictor



From Best Linear Predictor to Best Predictor



From Best Linear Predictor to Best Predictor

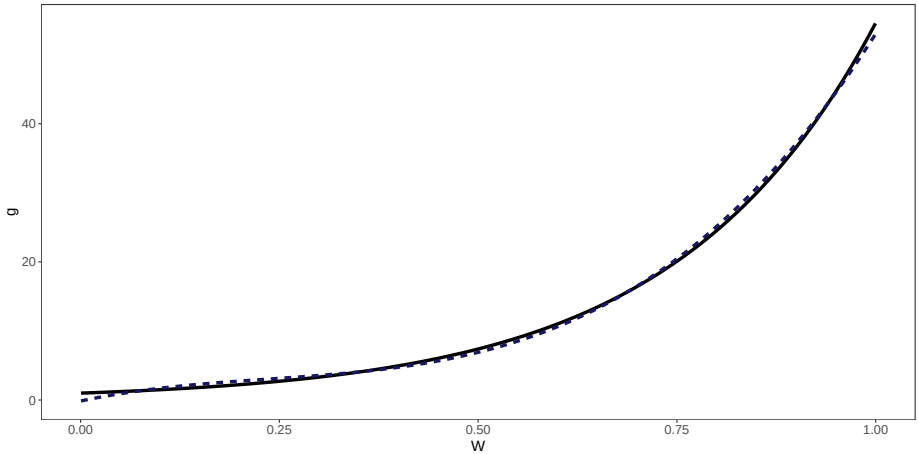


photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/

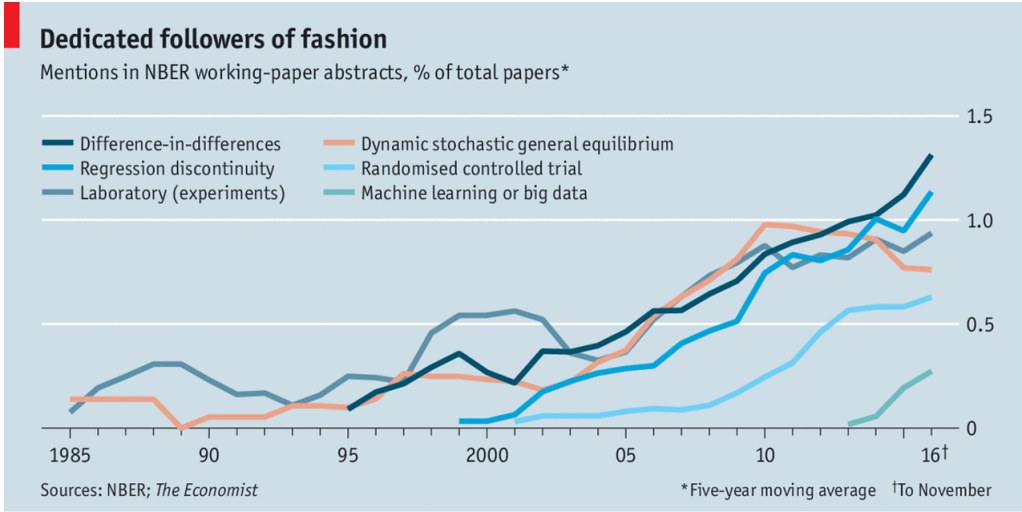
Agenda

Machine Learnists

The master-economist must possess a rare combination of gifts...He must be mathematician, historian, statesman, philosopher and data scientist – in some degree. He must understand symbols and speak in words. He must contemplate the particular, in terms of the general, and touch abstract and concrete in the same flight of thought. He must study the present in the light of the past for the purposes of the future. No part of man's nature or his institutions must be entirely outside his regard. He must be purposeful and disinterested in a simultaneous mood, as aloof and incorruptible as an artist, yet sometimes as near to earth as a politician."

adaptado de *Keynes (1924), Economic Journal*

Machine Learnists



Source: https://www.economist.com/finance-and-economics/2016/11/24/economists-are-prone-to-fads-and-the-latest-is-machine-learning