Introducción

Ciencia de Datos para la toma de decisiones en Economía

Ignacio Sarmiento-Barbieri

September 2, 2022

Agenda

Agenda

¿Qué es la máquina de aprender?

- 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?

- 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.

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.

Policy Prediction Problems

- Consider two toy examples:
  1. One policymaker facing a drought must decide whether to invest in a rain dance to increase the chance of rain.
  2. Another seeing clouds must decide whether to take an umbrella to work to avoid getting wet on the way home.
- Both decisions could benefit from an empirical study of rain.

Policy Prediction Problems

Illustrative Application

- Osteoarthritis (joint pain and stiffness) is a common and painful chronic condition among the elderly.
- Replacement of the affected joints, most commonly hips and knees, provide relief each year to around 500, 000 Medicare beneficiaries in the United States.
- The benefits accrue over time, so surgery only makes sense if someone lives long enough to enjoy them;
  - joint replacement for someone who dies soon afterward is futile—a waste of money and an unnecessary painful imposition on the last few months of life

Policy Prediction Problems

Illustrative Application

- The payoff to surgery depends on (eventual) mortality, creating a pure prediction problem.
- Put differently, the policy challenge is: can we predict which surgeries will be futile using only data available at the time of the surgery?
- This would allow us save both dollars and disutility for patients.

Policy Prediction Problems

Illustrative Application

TABLE 1—RISKIEST JOINT REPLACEMENTS

Predicted mortality percentile	Observed mortality rate	Futile procedures averted	Futile spending ($ mill.)
1	0.435 (0.028)	1,984	30
2	0.422 (0.028)	3,844	58
5	0.358 (0.027)	8,061	121
10	0.242 (0.024)	10,512	158
20	0.152 (0.020)	12,317	185
30	0.136 (0.019)	16,151	242

Source: Kleinberg et al (2015)

Policy Prediction Problems

La primera victoria y derrota de ML

- Contexto ¿similar? al de hoy: 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 ciudad, condado, estado, 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 (comparemos con la Colombia de 2009)
- 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

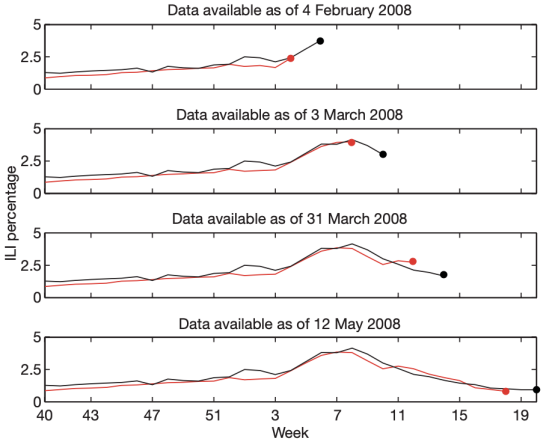


Figure 3 | ILI percentages estimated by our model (black) and provided by the CDC (red) in the mid-Atlantic region, showing data available at four points in the 2007-2008 influenza season. During week 5 we detected a sharply increasing ILI percentage in the mid-Atlantic region; similarly, on 3

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

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.

Libros

- Statistical Learning (FREE!!! beer)
  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
- Taddy, M. (2019). Business data science: Combining machine learning and economics to optimize, automate, and accelerate business decisions. McGraw Hill Professional.

. . . y otros libros y papers que voy a mencionar en clase.

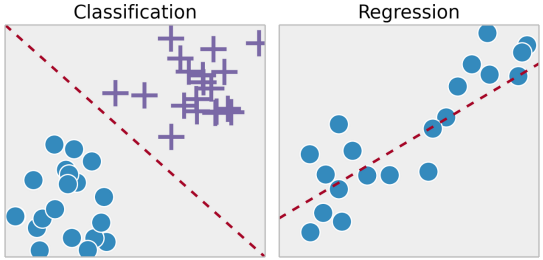
Agenda

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



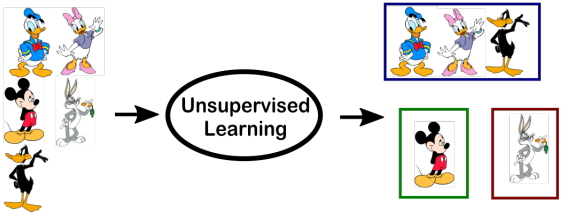
Source: shorturl.at/opqKT

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

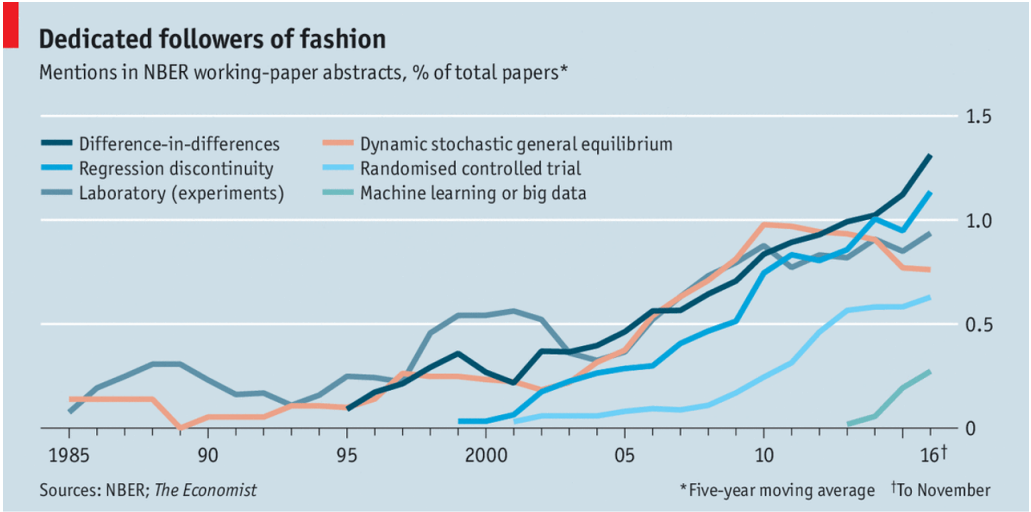
Agenda

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

Agenda

Further Readings

- Athey, S. (2017). Beyond prediction: Using big data for policy problems. Science, 355(6324), 483-485.
- Einav, Liran, and Jonathan D. Levin. The data revolution and economic analysis. No. w19035. National Bureau of Economic Research, 2013.
- Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). "Prediction policy problems" link
- Mullainathan y Spiess (2017), "Machine Learning: An Applied Econometric Approach" link
- Sosa Escudero, W. (2019). Big Data. Siglo Veintiuno Editores
- Taddy, M. (2019). Business data science: Combining machine learning and economics to optimize, automate, and accelerate business decisions. McGraw Hill Professional.
- Varian (2014), "Big Data: New Tricks for Econometrics" link