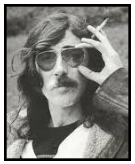
Big Data: Desafíos y Perspectivas para Datos Sociales

Walter Sosa-Escudero

Universisad de San Andrés y CONICET

6ta. Conferencia anual de la Sociedad Argentina de Economía Regional
San Carlos de Bariloche, 26 y 27 de Mayo de 2016

Preludio



'Todo es GMM. Si sabes GMM, sabes todo'

(alguien, circa 1993)

El fin del mundo

WIRED GEAR SCIENCE ENTERTAINMENT BUSINESS SECURITY DESIGN OPINION

SCIENCE : DISCOVERIES

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson 06.23.08

Illustration: Marian Bantjes

'Basta de una vez con la teoría del comportamiento humano, desde la linguística a la sociología. Olvídense de la taxonomía, la ontología y la psicología. ¿Quién sabe por qué la gente hace lo que hace? El punto es que lo hacen, y que podemos rastrearlo y medirlo con una precision antes impensable. Con suficientes datos los numeros hablan por si mismos.'

Chris Anderson, Wired, 23/6/2008

'Si bien muchos estudios muestran la utilidad de big data, estamos lejos de que reemplacen a los metodos y teorias tradicionales'

David Lazer, Science 2/2014

'Funes es big data sin estadistica'

Stephen Stigler, 2016, en The Seven Pillars of Statistical Wisdom

Clarín | Noticias | Deportes | ¡ExtraShow! | Estilo | Ser

iEco | Mercados · Indicadores · Empresas y Negocios | Economía

Clarin.com › iEco › Economía › 06/04/14

Big data: ¿Otra vez arroz?

DEBATE

2 opiná | 142 shares | 11 | 131

Walter Sosa Escudero
Profesor Asociado, Udesa

Creo conservar, en algún recóndito lugar de mi casa, mi paleta de paddle de cuando en los noventa pensaba que el juego del presente se transformaría en el deporte del futuro. También disfruto de los vinilos de mi adolescencia que escucho casi a diario. Y por alguna razón exótica guardo celosamente una caja de diskettes de mis comienzos con la computación personal, allá en los ochenta.

Esta charla

Una revision de las oportuinades que plantea el paradigma de big data / aprendizaje automatico a la economia empirica.

- Lenguaje y jerga.
- Tres casos regionales / espaciales.
- Diferencias entre el paradigma clasico en econometria (explicativo) y el de aprendizaje (predictivo, exploratorio).
- Herramientas
- Consideraciones generales
- Guia de acciones (lecturas, cursos, etc.)

Jerga

NoSQL, Hadoop, mineria, aprendizaje, visualización, fat models, funciones de pérdida, riesgo de Bayes, LASSO, CART, GARROTE, sobreajuste, muestra de entrenamiento, aprendizaje supervisado y no supervisado, validación cruzada, árboles, bosques, podas, nodos, clusters, boosting, bagging, matriz de confusion, curva ROC, regularización, shrinkage, Bayes, incertidumbre de modelo, promedio de modelos, error reproducible, prediccion fuera de muestra, bases, splines, GAM, support vector machines, subset selection, scrapping, networks, phyton, predictive analytics.

Demasiado Spanglish...

Problemas

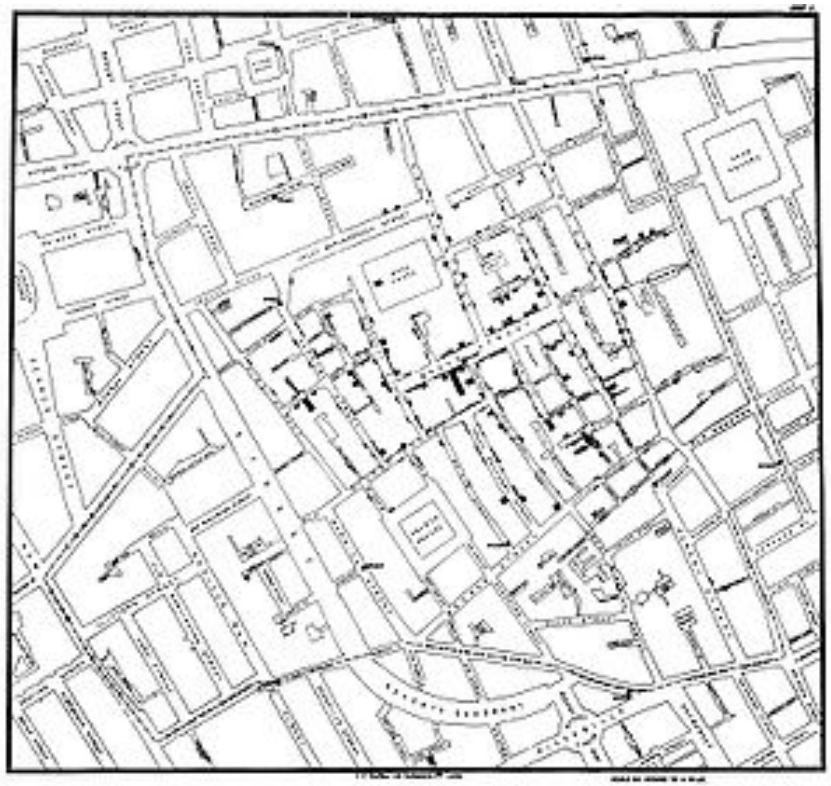
- Decidir si un email es spam.
- Identificar codigos postales escritos a mano.
- Relevar precios.
- Asignar tratamiento para pacientes con cancer.
- Identificar a los pobres.
- Encontrar el nombre de una canción a partir de un silbido o tarareo.
- Traducir un texto.
- Predecir preferencias por la redistribucion ante una eleccion.
- Medir la clase media.
- Recomendar libros o peliculas.

Babel

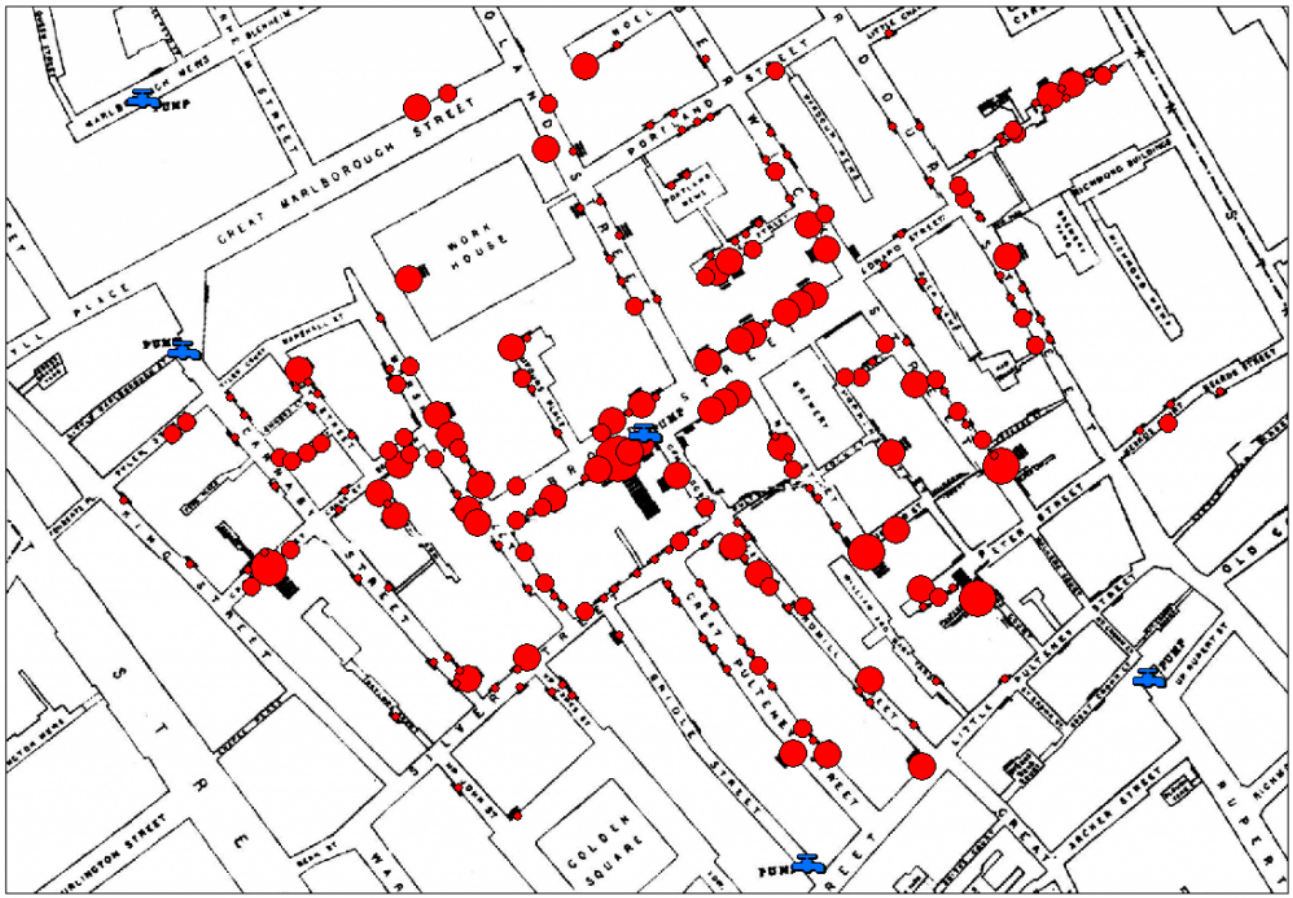
- Aprender, hurgar, encontrar patrones, clasificar, reducir dimensionalidad, visualizar, resumir, decidir.
- Manipular, mover, organizar, guardar, recuperar, explorar datos masivos.
- Data mining, statistical learning, machine learning, predictive analytics.

Más Spanglish...

Todo empieza con el mapa de Snow



Snow big data



Ascenso y caida de Google Flu Trends

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

[1]*Google Inc.* [2]*Centers for Disease Control and Prevention*

Epidemics of seasonal influenza are a major public health concern, causing tens of millions of respiratory illnesses and 250,000 to 500,000 deaths worldwide each year[1]. In addition to seasonal influenza, a new strain of influenza virus against which no prior immunity exists and that demonstrates

- 450.000.000 de modelos
- 1 dia vs 2 semanas

Ascenso y caida de GFT

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

David Lazer,[1,2]* Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[3,5,6]

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Cen-

- Sobrestimacion de GFT
- Busqueda endogena
- Granularidad a nivel region

Track the poor

ECONOMICS

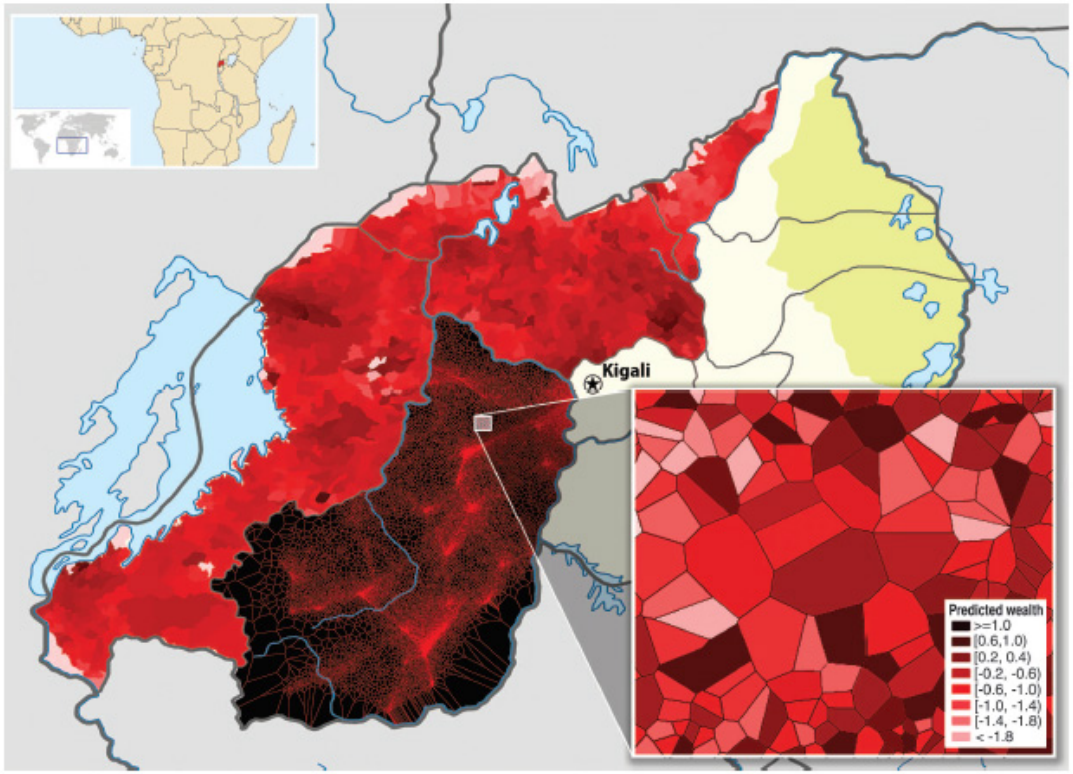
Predicting poverty and wealth from mobile phone metadata

Joshua Blumenstock,[1*] Gabriel Cadamuro,[2] Robert On[3]

Accurate and timely estimates of population characteristics are a critical input to social and economic research and policy. In industrialized economies, novel sources of data are enabling new approaches to demographic profiling, but in developing countries, fewer sources of big data exist. We show that an individual's past history of mobile phone use can be used to infer his or her socioeconomic status. Furthermore, we demonstrate that the predicted attributes of millions of individuals can, in turn, accurately reconstruct the distribution of wealth of an entire nation or to infer the asset distribution of microregions composed of just a few households. In resource-constrained environments where censuses and household surveys are rare, this approach creates an option for gathering localized and timely information at a fraction of the cost of traditional methods.

Deliable, quantitative data on the economic characteristics of a country's population are essential for sound economic policy and unemployment (*9*), electoral outcomes (*10*), and economic development (*8*). Although most comparable sources of big data are scarce in the

Track the poor



Impuesto a las ventas

American Economic Review 2014, 104(1): 1–26
http://dx.doi.org/10.1257/aer.104.1.1

Sales Taxes and Internet Commerce†

By Liran Einav, Dan Knoepfle, Jonathan Levin, and Neel Sundaresan*

We estimate the sensitivity of Internet retail purchasing to sales taxes using eBay data. Our first approach exploits the fact that a seller's location—and therefore the applicable tax rate—is revealed only after a buyer has expressed interest in an item. We document how adverse tax "surprises" reduce the likelihood of purchase and shift subsequent purchases toward out-of-state sellers. We then use more aggregated data to estimate that every one percentage point increase in a state's sales tax increases online purchases by state residents by almost 2 percent, while decreasing their online purchases from state retailers by 3–4 percent. (*JEL* H71, L81, L86)

Impuesto a las ventas

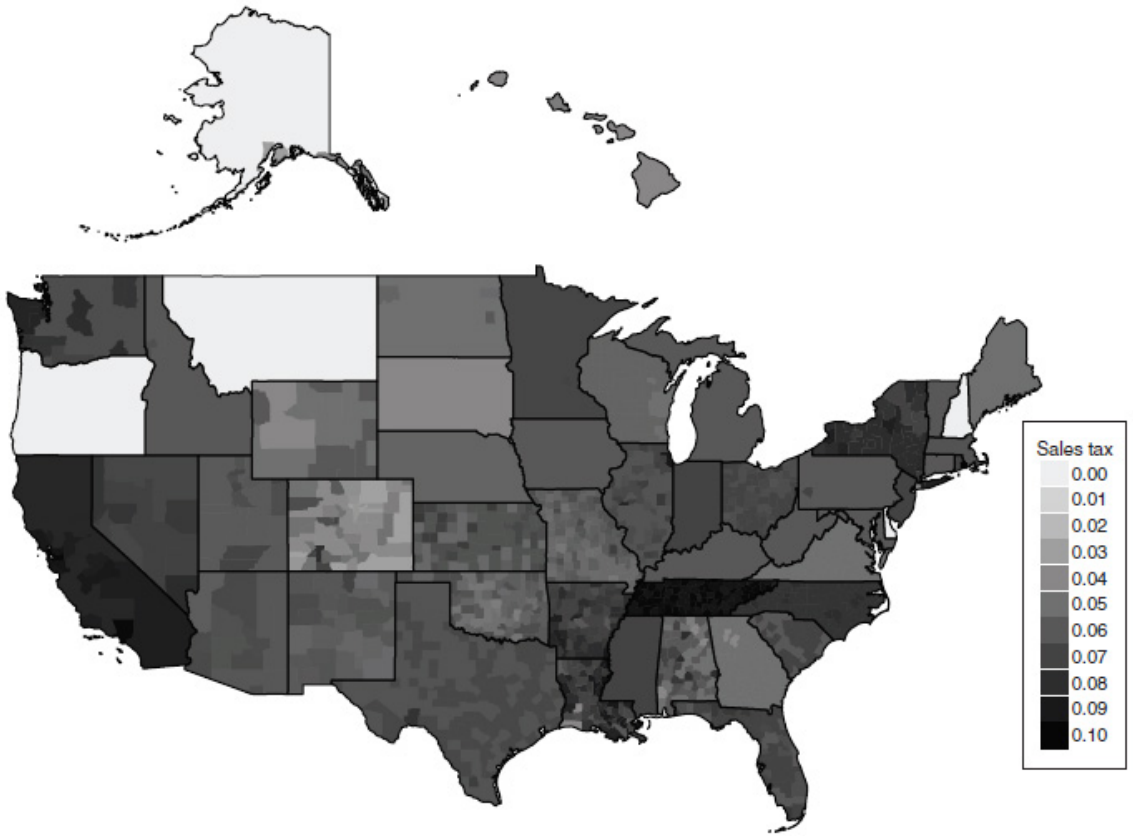


FIGURE 1. CROSS SECTION VARIATION IN SALES TAX RATES

El paradigma explicativo

Gran parte de la actividad academica en economia:

$$Y = f(X) + u$$

- Objetivo: conocer $f(\ )$ o sus caracteristicas (derivadas, por ejemplo).
- Exito: insesgadez, varianza minima, consistencia, etc.
- ¿f? Viene de una estructura (teoria) o representa un experimento (¿teoria?).
- Validez interna ('ajusta bien a los datos').
- Foco en f

Dinamica

- Estimar correctamente f (puntual, intervalos).
- Inferencia estadistica. Errores estandar y test de hipotesis.
- Paradigma de estimacion insesgada / consistente y varianza minima.
- Mayoritariamente frecuentista.

El paradigma predictivo

$$Y = f(X) + u$$

- Objetivo: predicir Y en base a X, sin observar u y sin saber $f(.)$ ('aprender' f).
- Exito: precedir bien *fuera de la muestra.*
- ¿f? Desconocido, hay que 'aprenderla'
- Datos de entrenamiento: usados para construir / estimar un modelo predictivo.
- Datos de test / evaluación: usados para validar y evaluar.
- Predecir bien en los datos de *test*.
- Foco en Y.

Dinamica

- Decidir qué puede ser f.
- Aprender (estimar) f. Volver al paso inicial. Iterar.
- Decidir la dimensión de X.

Cuidado (para econometristas): el objetivo es predicion fuera de la muestra!.

Exito

Perdida esperada: $E\big[L(Y, \hat{f}(X))\big]$.

Perdida cuadratica:

$$E\big[(Y - \hat{f}(X))^2\big]$$

Perdida cuadratica en los datos de test:

$$E\big[(Y - \hat{f}(X))^2 \mid \mathcal{T}\big]$$

Trade-off sesgo-varianza

$$\begin{aligned} \text{Err}(x_0) &= E\left[(Y - \hat{f}(x_0))^2 \mid X = x_0\right] && (1) \\ &= \text{Sesgo}^2 \hat{f}(x_0) + \text{Var}\ \hat{f}(x_0) + \sigma_\epsilon^2 && (2) \end{aligned}$$

- σ^2 es el *erorr irreducible*
- Trade-off: cuanto mas complejo es $\hat{f}$ menor el sesgo y menor la varianza.
- El paradigma de estimacion insesgada no resuelve el problema, solo lo aisla.
- Sesgo pequeño puede llevar a una dramatica reduccion en la varianza. Abre la puerta para estimacion sesgada.

Validacion cruzada

Como medir $Err(X \mid \mathcal{T})$?. Una estrategia muy importante es *k-fold cross-validation*.

- Partir los datos en K partes iguales.
- Ajustar el modelo dejando afuera una de las particiones.
- Computar el error de prediccion para los datos dejados afuera.
- Repetir para $k = 1, \ldots, K$.

La estimacion de validacion cruzada del error de prediccion es:

$$CV(\hat{f}) = \frac{1}{N}\left(y_i - \hat{f}_{-k}(x_i)\right)^2,$$

$\hat{f}_{-k}(x_i)$ es la estimacion con la $k-$esima parte excluida.

- Como elegir K?: Folclore: 5, 10. Y N ('leave-one-out)?. Trade off sesgo-varianza nuevamente: mas datos para el modelo, menos parra la prediccion.
- CV para seleccion de modelos: $f(x, \alpha)$ en donde α parametriza la complejidad (numero de variables, nudos en splines, ancho de banda, etc.). Entonces: computar CV para valores alterantivos de α y minimizar.

$$CV(\hat{f}, \alpha) = \frac{1}{N}\Big(y_i - \hat{f}_{-k}(x_i, \alpha)\Big)^2,$$

Modelos para Y continuo

Toda la ferreteria de la econometria

- Regresion multiple
- Polinomios, dummys, interacciones y modelos saturados.
- Regresion no parametrica. Splines. Bases.
- La maldicion de la dimensionalidad.

Eleccion de modelos

Big data: la dimension de X puede ser ridiculamente alta. El numero de columans pude exceder el numero de filas (fat models).

Recordar: el objetivo es medir la performance predictiva fuera de la muestra. Problema: la performance dentro de la muestra sobreestima a la performance fuera de la muestra (overfit).

Ejemplo: R^2 is no decreciente *dentro de la muestra*.

Shrinkage

Estimar y elegir variables en un solo paso.

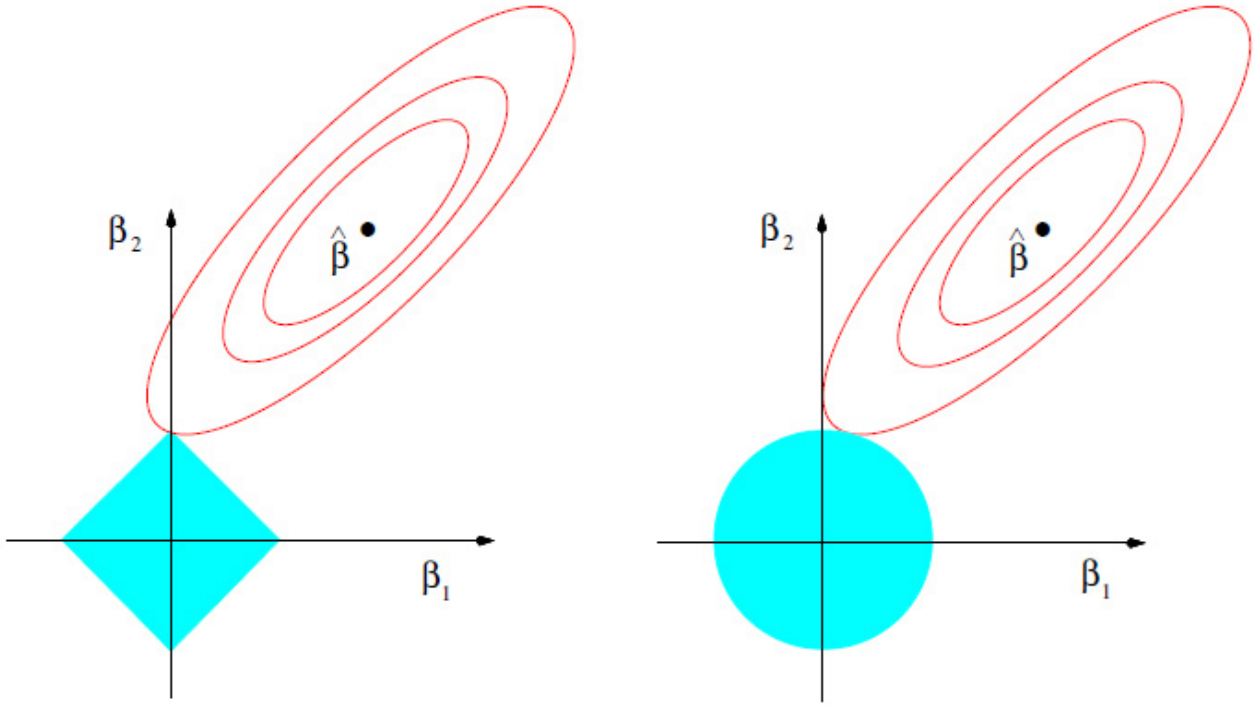
- Ridge

$$\min_{\beta} \sum_{i=1}^{n} e_i^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

- Lasso

$$\min_{\beta} \sum_{i=1}^{n} e_i^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

λ penaliza elegir parametros no nulos (¿incluir variables?). Son similares pero lo hacen de una manera completamente distinta.



Fuente: Hastie, Tibshirani and Friedman (2008)

- Ridge se invento para el problema de multicolinealidad. Sesga los coeficientes a cero, para reducir varianza
- Ambos 'encogen' hacia cero. Lasso elige exactamente cero: 'smooth subset selecction' (Least Absolute Shrinkage and Selection Operator).
- Ambos tienen una 'lectura' bayesiana.
- Como elegir λ?: cross validation.
- Lasso formaliza lo que en la practica mucha gente hace 'artesanalmente'.

Clasificacion

Y es binaria (spam o no, pobre o no, pago un credito o no, etc.)

- Econometria: estimar $Pr(Y = 1 \mid X)$. Interes en los efectos marginales.
- Data mining: predecir Y.

Modelos:

- MLP, logits, probits (not popular).
- Analisis discriminante.
- $K-$ vecinos cercanos. Voto por mayoria local.
- Arboles (mas en un rato

Clasificador de Bayes: predecir el resultado mas probable (1 if $\hat{p} > 0{,}5$). Minimiza el riesgo de Bayes. La eleccion de un umbral no es trivial. El riesgo de bayes podria ser irrelevante. Ej: credit scoring.

Exito

Matriz de confusion (ejemplo):

		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

Source: Hastie, Tibshirani and Friedman (2008)

- Empezar con $c = 0$ aumentarlo progresivamente. Un error crece y el otro baja: curva ROC.
- *Criterio:* tasa de error: $1/N \sum I(y_i \neq \hat{y}_i)$
- Eleccion de modelos: todo lo anterior funciona (CV, en particular)

Arboles

Victimas del Titanic: logit

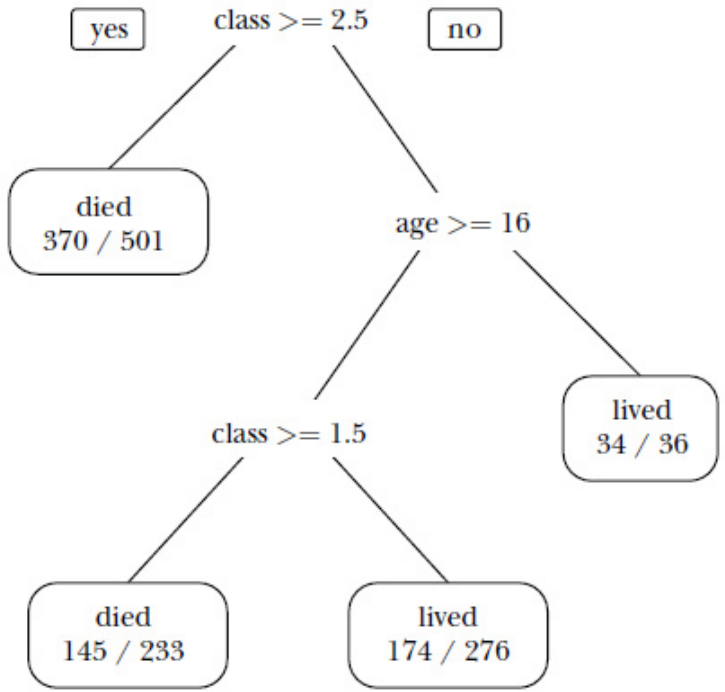
Logistic Regression Table

Predictor	Coef	SE Coef	Z	P
Constant	-0.191839	0.175568	-1.09	0.275
Class				
First	0.971320	0.0952002	10.20	0.000
Gender				
Male	-1.03799	0.200630	-5.17	0.000
Age	0.0044963	0.0033885	1.33	0.185
ChildorAdult				
Child	0.387517	0.174976	2.21	0.027
Gender*Age				
Male	-0.0123825	0.0040596	-3.05	0.002

Victimas del Titanic: CART

Figure 1

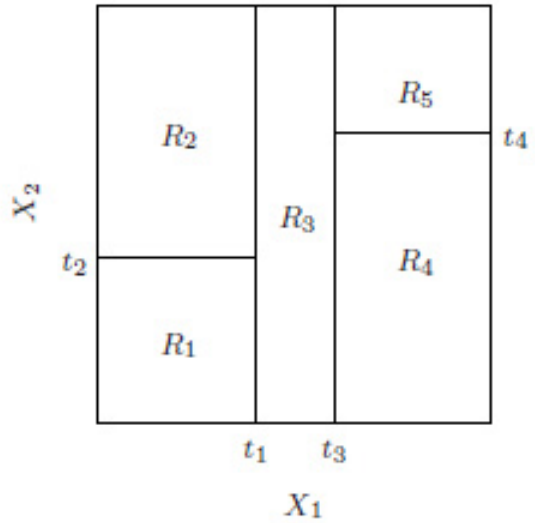
A Classification Tree for Survivors of the *Titanic*

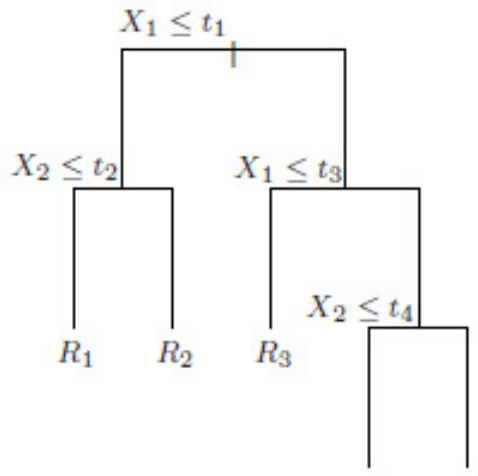


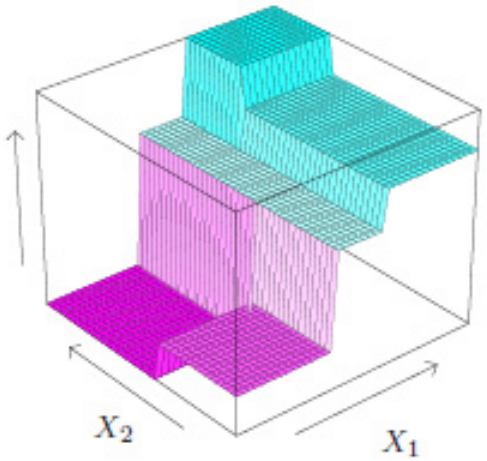
Note: See text for interpretation.

Algoritmo CART

- Elegir la j-esima variable explicativa. Partir los datos de acuerdo a $X_j > s$. Dentro de cada submuestra predecir (media, mayoria).
- Problema: como elegir j y s de modo de minimizar el error de prediccion? Fuerza bruta: todas las variables, todas las particiones! (big data, big computer!).
- Repertir dentro de la particion anterior.
- Seguir hasta que el error de prediccion no caiga.

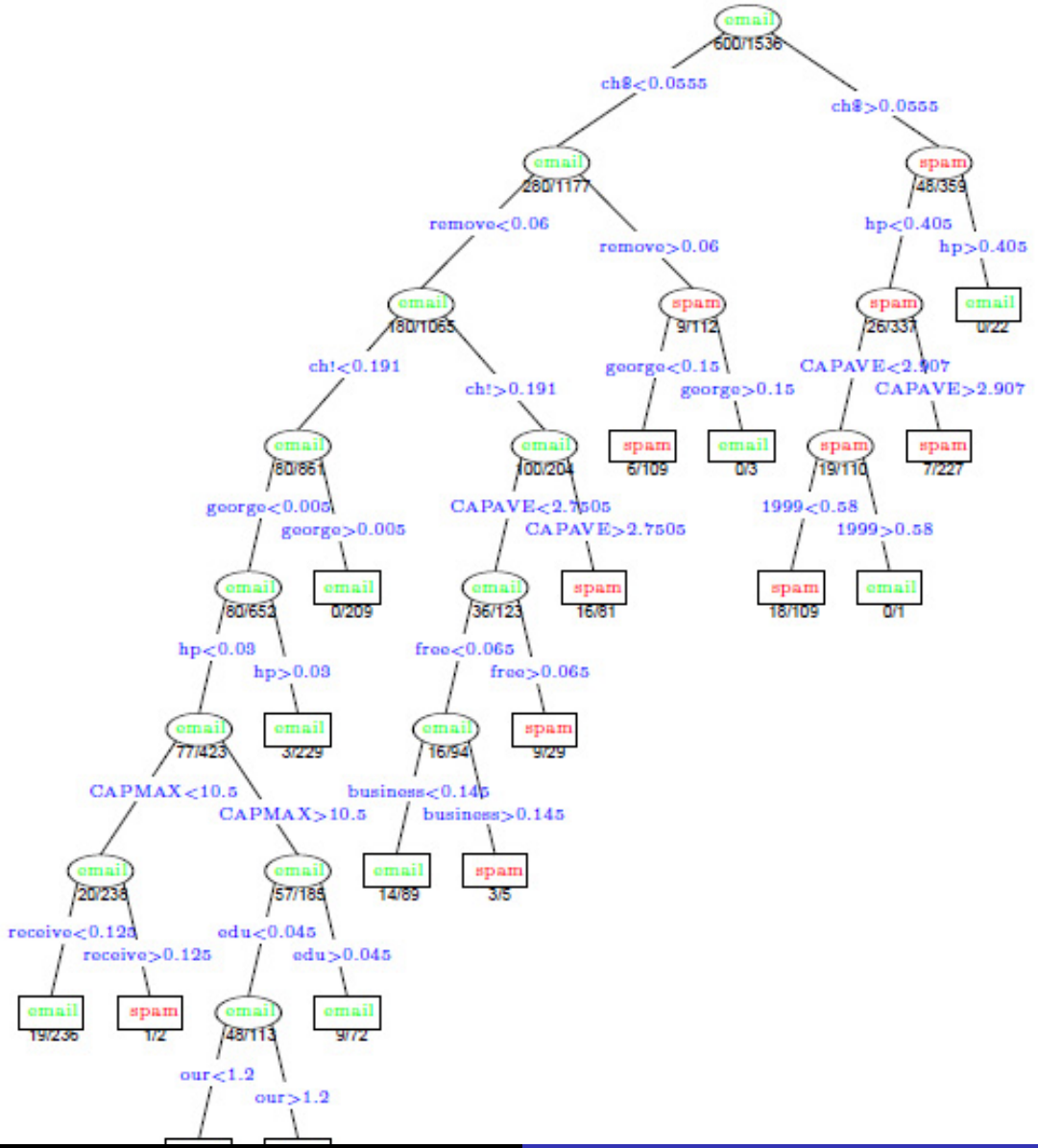

R_1
R_2
R_3
R_4
R_5
t_1
t_2
t_3
t_4
X_1
X_2


$X_1 \leq t_1$
$X_2 \leq t_2$
$X_1 \leq t_3$
$X_2 \leq t_4$
R_1
R_2
R_3


X_2
X_1

Ideas:

- Muy popular. Facil de comunicar.
- Las ramas crecen en cada particion. Poda: como parar el arbol.
- Variables importantes: aparecen 'arriba' del arbol.
- Anda mal si el modelo es demasiado lineal (logit o MLP mejor).
- Bagging: muestra bootstrap y CART en cada muestra boostrap. Voto por mayoria.
- Random forests: multiples arboles eligiendo predictores al azar en cada ramificacion.
- Mejora la prediccion pero empeora la comunicabilidad.


email
600/1536
ch$<0.0555
ch$>0.0555
email
280/1177
spam
48/359
remove<0.06
remove>0.06
hp<0.405
hp>0.405
email
180/1065
spam
9/112
spam
26/337
email
0/22
ch!<0.191
ch!>0.191
george<0.15
george>0.15
CAPAVE<2.907
CAPAVE>2.907
email
80/861
email
100/204
spam
6/109
email
0/3
spam
19/110
spam
7/227
george<0.005
george>0.005
CAPAVE<2.7505
CAPAVE>2.7505
1999<0.58
1999>0.58
email
80/652
email
0/209
email
36/123
spam
16/81
spam
18/109
email
0/1
hp<0.03
hp>0.03
free<0.065
free>0.065
email
77/423
email
3/229
email
16/94
spam
9/29
CAPMAX<10.5
CAPMAX>10.5
business<0.145
business>0.145
email
20/238
email
57/185
email
14/89
spam
3/5
receive<0.125
receive>0.125
edu<0.045
edu>0.045
email
19/236
spam
1/2
email
48/113
email
9/72
our<1.2
our>1.2

Precios sombra

- Basis expansion
- Neural networks.
- Support vector machines.
- Unsupervised vs. supervised
- Clusters and groups
- Principal components and dimension reduction
- Fat models
- Data handling / programming / scrapping

Lo bueno

- Rol de la induccion y la descripción en las ciencias sociales (y en la ciencia).
- ¿Big data o new data?
- Experimentos
- Construccion de contrafactuales.
- Complejidad, alta frecuencia.
- Re-evaluar la 'tirania de la insesgadez'.
- ¡Bayes!

Lo malo

- ¿Small data? ¿Realmente hay un problema de pocos datos en las ciencias sociales?
- Mas datos no es necesariamente mejor.
- Choice based sampling. Papelones historicos. El origen de Gallup.
- Small data es un problema poblacional: no observabilidad de contrafactuales.
- Polítcas: ¿prohibir paraguas? ¿regalar televisores?
- Daniel Heymann y el PBI diario. Datos de la frecuencia correcta. Raices unitarias.

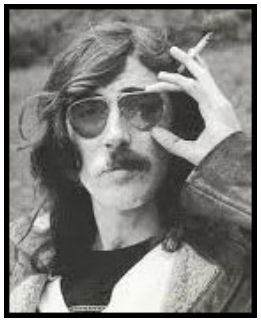
Lo feo

¿El fin de la teoria? Intentos 'ateoricos' en economia.

- VAR y econometria dinamica.
- La 'revolucion de credibilidad': experimentos, instrumentos. Mostly Harmless.
- ¿Big data? Mejoras sustanciales con la teoria (linguistica basada en viejas traducciones, meteorologia). Acciones de politica. Analisis causal.

Teoria y conocimiento causal como necesidad atavica.

Perspectivas



'... yo ya soy parte del mar'

- Posturas extremas sobre el tema.
- Discusion filosofica produnda sobre el rol de la teoria y la induccion.
- ¿Cambios en la enseñanza?

Caja de herramientas

Review

- Breiman (2001, Statistical Modeling: The Two Cultures, StatSci)
- Varian (2014, Big Data: New Tricks for Econometrics, JEP).
- Einav y Levin (2014, Economics in the age of big data).
- Sosa Escuedero (2105, Big data, mining and learning: new challenges for applied econometrics, mimeo)

Metodos

- Hastie, Tibshirani,Friedman (2009)
- James, Witten, Hastie and Tibshirani (2014).
- Murphy (2012, Machine Learning)

Papers

- Edicion especial de JEP sobre Big Data (JEP, 2014)
- Papers: Keely and Tan (2008, JPubE)
- Bajari et al. (2015, American Economic Review)
- Einav et al. (2015, AER).

Varios

- Mayer-Schonberger y Cukier (Big Data, 2013).
- Curso de Susan Athey (NBER).
- Charla de Tim Harford sobre 'The Big Data Trap'.
- Nota en Clarin (6/4/2014).
- Computer intensive. Olvidense de Stata
- Curso (learning): Hastie and Tibshirani (Stanford)
- Libros gratis!

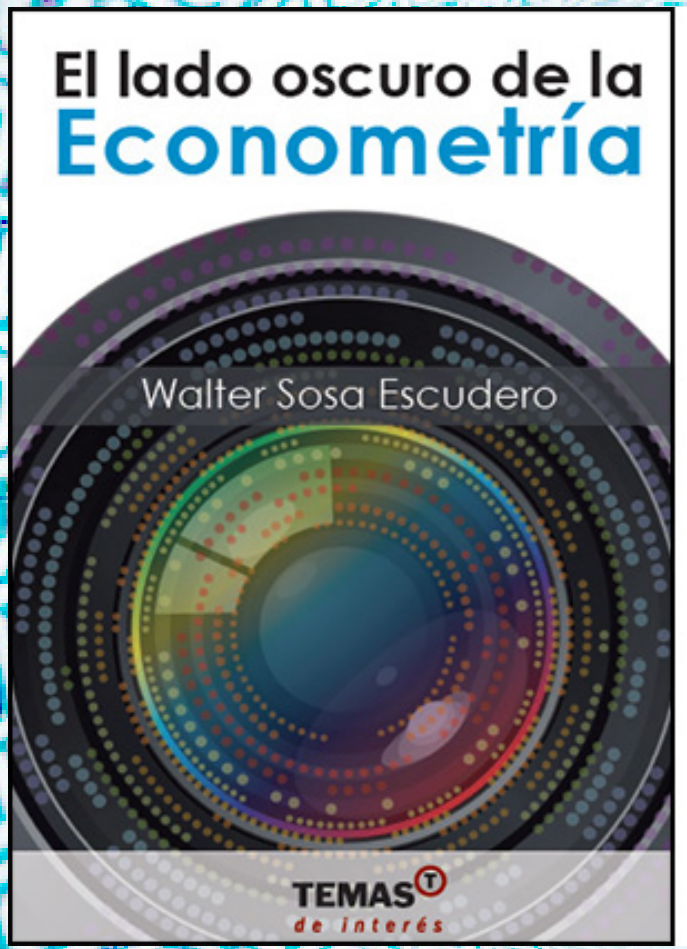
JLB

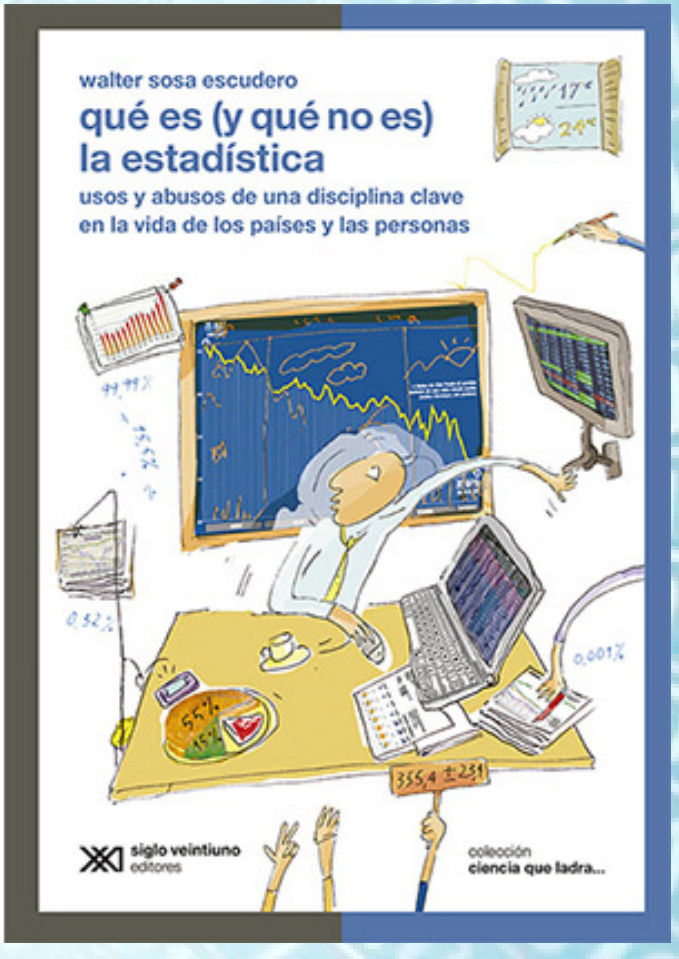
'... su antepasado no creia en un tiempo uniforme, absoluto. Creia en infinitas series de tiempos, en una red creciente y vertiginosa de tiempos divergentes, convergentes y paralelos. ... No existimos en la mayoría de esos tiempos; en algunos existe usted y no yo; en otros, yo, no usted; en otros, los dos. En este, que un favorable azar me depara, usted ha llegado a mi casa; en otro, usted, al atravezar el jardín, me ha encontrado muerto; en otro, yo digo estas mismas palabras, pero soy un error, un fantasma.'

El jardin de senderos que se bifurcan

'Ireneo tenía diecinueve años; había nacido en 1868; me parecio monumental como el bronce, más antiguo que Egipto, anterior a las profecias y a las piramides. Pense que cada una de mis palabras (que cada uno de mis gestos) perduraria en su implacable memoria; me entorpecio el temor de multiplicar ademanes inutiles.'

Funes, el memorioso


El lado oscuro de la
Econometría
Walter Sosa Escudero
TEMAS
de interés


walter sosa escudero
qué es (y qué no es)
la estadística
usos y abusos de una disciplina clave
en la vida de los países y las personas
siglo veintiuno
editores
colección
ciencia que ladra...