Lecture 1: Introducción

Aprendizaje y Minería de Datos para los Negocios

Ignacio Sarmiento-Barbieri

Universidad de los Andes

October 16, 2021

Agenda

Motivación

La primera victoria y derrota del Big Data

- 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

Motivación

Google se ha unido a la conversación

- Google propuso un mecanismo ingenioso: **Google Flu Trends**
- Punto de partida:
  - Proporcion 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

Motivación

Google se ha unido a la conversación

- ¿Un solo 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!

Motivación

Google se ha unido a la conversación

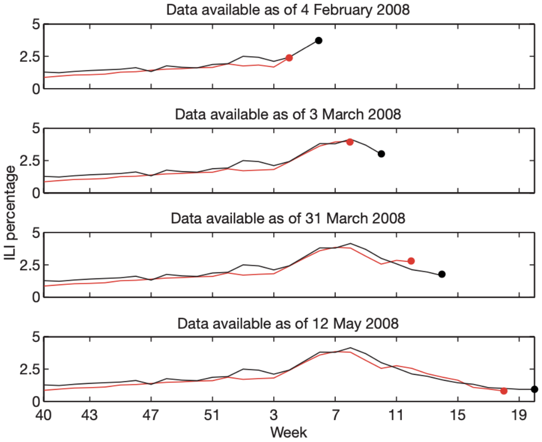


Figure 3 | ILI percentages estimated by our model (black) and provided by the CDC (red) in the mid-Atlantic region, showing data available at four points in the 2007-2008 influenza season. During week 5 we detected a sharply increasing ILI percentage in the mid-Atlantic region; similarly, on 3

Motivación

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

Motivación

- Otro ejemplo, los algoritmos de reconocimiento de cara:
  - no son reglas fijas basadas en que los humanos entendemos por rostros y a partir de ello buscar combinaciones de pixeles.
  - son algoritmos que usan datos de fotos etiquetadas con un rostro y estiman una función $f(x)$ que predice si es un rostro o no a partir de pixeles x.
- El aprendizaje de maquinas se hizo una realidad cuando los investigadores dejaron de afrontarlo de manera teórica y lo hicieron empíricamente.
- 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?

¿Qué entendemos por Big Data y ML?

¿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
    - GPS
    - Texto
    - Imágenes
- Machine Learning
  - Cambio de paradigma de estimación a predicción

Presentación: Sobre mi

- Ignacio Sarmiento Barbieri
- https://ignaciomsarmiento.github.io/
- i.sarmiento@uniandes.edu.co
- Intereses: Economia Pública y Urbana. Economia del Crime. Econometria Aplicada, Big Data y Machine Learning.
- Originario de Salta, Argentina

Presentación: Sobre mi

Motivation

- We discussed the examples of Google Flu and Facebook face detection
  - Take away, the success was driven by an empiric approach
  - Given data estimate a function $f(x)$ that predicts y from x
- This is basically what we do as economists everyday so:
  - Are these algorithms merely applying standard techniques to novel and large datasets?
  - If there are fundamentally new empirical tools, how do they fit with what we know?
  - As empirical economists, how can we use them?

Big vs Small, Classic vs Predictive

- Classical Stats (small data?)
  - Get the most of few data (Gosset)
  - Lots of structure, e.g. $X_1, X_2, ..., X_n \sim t_v$
  - Carefully curated $\rightarrow$ approximates random sampling (expensive, slow) but very good and reliable

- Big Data (the 4 V's)
  - Data **V**olume
  - Data **V**ariety
  - Data **V**elocity
  - Data **V**alue

The Classic Paradigm

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

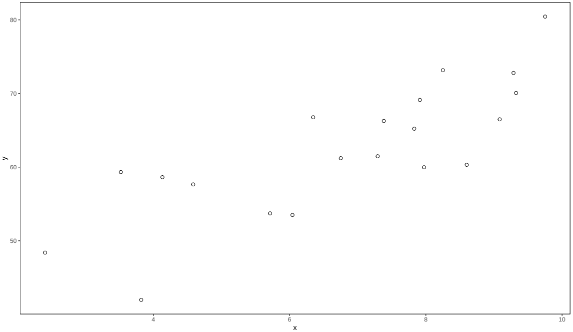
- Interest lies on inference
- "Correct" $f()$ to understand how Y is affected by X
- Model: Theory, experiment
- Hypothesis testing (std. err., tests)

The Predictive Paradigm

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

- Interest on predicting Y
- "Correct" $f()$ to be able to predict (no inference!)
- Model?

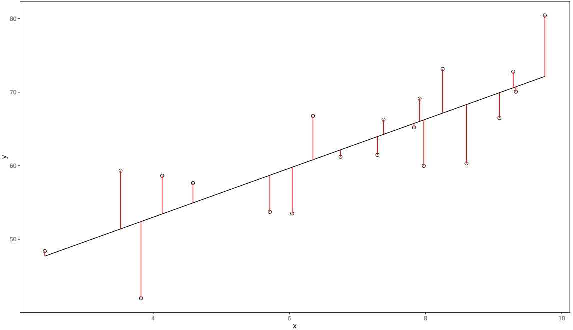
How to choose $f(.)$



Source: simulated data, see `figures` folder for scripts

How to choose $f(.)$

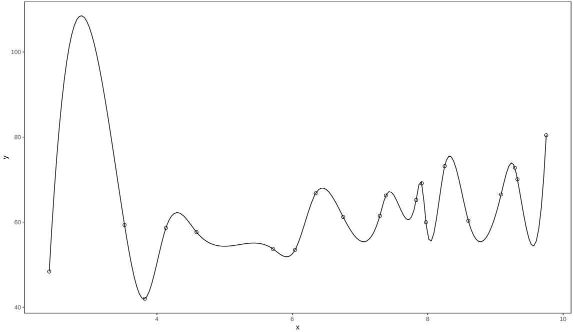
- Linear $f(X) = X\beta$



Source: simulated data, see `figures` folder for scripts

How to choose $f(.)$

- Spline $f(X) = g(X)$, where g is a spline



Source: simulated data, see `figures` folder for scripts

Statistical Decision Theory: A bit of theory

- We need a bit of theory to give us a framework for choosing f
- A decision theory approach involves an **action space** $\mathcal{A}$
- The **action space** $\mathcal{A}$ specify the possible "actions we might take"
- Some examples

Table 1: Action Spaces

Inference	Action Space
Estimation $\theta, g(\theta)$	$\mathcal{A} = \Theta$
Prediction	$\mathcal{A} = \textit{space of } X_{n+1}$
Model Selection	$\mathcal{A} = \{\textit{Model I}, \textit{Model II}, \ldots\}$
Hyp. Testing	$\mathcal{A} = \{\textit{Reject} \mid \textit{Accept } H_0\}$

Statistical Decision Theory: A bit of theory

- After the data $X = x$ is observed, where $X \sim f(X|\theta)$, $\theta \in \Theta$
- A decision is made
- The set of allowable decisions is the action space ($\mathcal{A}$)
- The loss function in an estimation problem reflects the fact that if an action a is close to θ,
  - then the decision a is reasonable and little loss is incurred.
  - if it is far then a large loss is incurred

$$L : \mathcal{A} \to [0, \infty] \tag{3}$$

Statistical Decision Theory: A bit of theory

Loss Function

- If θ is real valued, two of the most common loss functions are
  - Squared Error Loss:

$$L(a,\theta) = (a-\theta)^2 \tag{4}$$

  - Absolute Error Loss:

$$L(a,\theta) = |a-\theta| \tag{5}$$

- These two are symmetric functions. However, there's no restriction. For example in hypothesis testing a "0-1" Loss is common.
- Loss is minimum if the action is correct

Statistical Decision Theory: A bit of theory

Risk Function

In a decision theoretic analysis, the quality of an estimator is quantified by its risk function, that is, for an estimator $\delta(x)$ of θ, the risk function is

$$R(\theta,\delta) = E_\theta(L(\theta,\delta(X))) \tag{6}$$

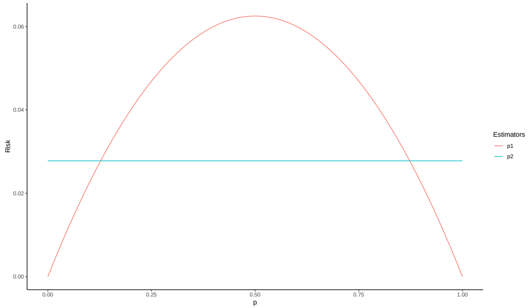
at a given θ, the risk function is the average loss that will be incurred if the estimator $\delta(X)$ is used

- Since θ is unknown we would like to use an estimator that has a small value of $R(\theta,\delta)$ for all values θ
- Loss is minimum if the action is correct
- If we need to compare two estimators (δ_1 and δ_2) then we will compare their risk functions
- If $R(\delta_1,\theta) < R(\delta_2,\theta)$ for all $\theta \in \Theta$, then δ_1 is preferred because it performs better for all θ

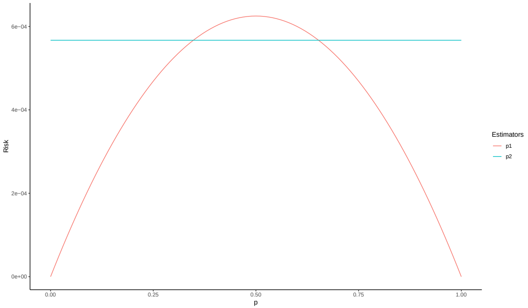
Statistical Decision Theory: A bit of theory

Example: Binomial Risk Function

- Let $X_1, X_2, ...X_n \sim_{iid} Bernoulli(p)$
- Consider 2 estimators for p: $\hat{p}^1 = \frac{1}{n}\sum X_i$ and $\hat{p}^2 = \frac{\sum X_i + \sqrt{n/4}}{n+\sqrt{n}}$
- Their risks are: $R(\hat{p}^1, p) = \frac{p(1-p)}{n}$ and $R(\hat{p}^2, p) = \frac{n}{4(n+\sqrt{n})^2}$



(a) n=4



(b) n=400

Decision Theory for prediction

How to choose f?

- In a prediction problem we want to predict Y from $f(X)$ in such a way that the loss is minimum
- Assume also that $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}$ with joint distribution $Pr(X,Y)$

$$R(Y,f(X)) = E[(Y - f(X))^2] \tag{7}$$

$$= \int (y - f(x))^2 Pr(dx,dy) \tag{8}$$

conditioning on X we have that

$$R(Y,f(X)|X) = E_X E_{Y|X}[(Y - f(X))^2|X] \tag{9}$$

this risk is also known as the **mean squared (prediction) error** $MSE(f)$

Decision Theory for prediction

It suffices to minimize the $MSE(f)$ point wise so

$$f(x) = argmin_m E_{Y|X}[(Y-m)^2|X=x) \tag{10}$$

Y a random variable and m a constant (predictor)

$$min_m E(Y-m)^2 = \int (y-m)^2 f(y)dy \tag{11}$$

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

Decision Theory for prediction

Proof

FOC

$$\int -2(y-m)f(y)dy = 0 \tag{12}$$

Dividing by -2 and reorganizing

$$m\int f(y)dy = \int yf(y)dy \tag{13}$$

Decision Theory for prediction

$$m \int (y)dy = \int yf(y)dy \tag{14}$$

$$m = E(Y|X = x) \tag{15}$$

The best prediction of Y at any point $X = x$ is the conditional expectation function (CEF), when best is measured using a square error loss

- What shape does the CEF take?
- Linear
  - (y, X) are jointly normal
  - When models are saturated.

Linear Regression

- Note the following from the *Regression-CEF Theorem*
  The function $X'\beta$ provides the minimum risk linear approximation to $E(Y|X)$, that is

$$\beta = \underset{b}{argmin}\, E\left\{(E(Y|X) - X'b)^2\right\} \tag{16}$$

- Proof

$$(Y - X'b)^2 = ((Y - E(Y|X)) + (E(Y|X) - X'b))^2 \tag{17}$$

$$= (Y - E(Y|X))^2 + (E(Y|X) - X'b)^2 + 2(Y - E(Y|X))(E(Y|X) - X'b) \tag{18}$$

- The CEF approximation problem then has the same solution as the population least square problems

Linear Regression

- Regression provides the best linear predictor for the dependent variable in the same way that the CEF is the best unrestricted predictor of the dependent variable.
- The fact that Regression approximates the CEF is useful because it helps describe the essential features of statistical relationships, without necessarily trying to pin them down exactly.
- Linear regression is the "work horse" of econometrics and (supervised) machine learning.
- Very powerful in many contexts.
- Big 'payday' to study this model in detail.

Linear Regression Model

$f(X) = X\beta$, estimating $f(.)$ boils down to estimating β

$$y = X\beta + u \tag{19}$$

where

- y is a vector $n \times 1$ with typical element y_i
- X is a matrix $n \times k$
  - Note that we can represent it as a column vector $\underset{n\times k}{X} = [\underset{n\times 1}{X_1} \; \underset{n\times 1}{X_2} \dots \underset{n\times 1}{X_k}]$
- β is a vector $k \times 1$ with typical element β_j

Thus

$$\begin{aligned} y_i &= X_i'\beta + u_i \\ &= \sum_{j=1}^{k} \beta_j X_{ji} + u_i \end{aligned} \tag{20}$$

Recap

- We start shifting paradigms
- Tools are not that different (so far)
- Decision Theory: Risk with square error loss $\rightarrow$ MSE
- OLS is a "work horse" approximates the $E[Y|X]$ quite well
- Next Class:
  - Next Class: OLS, Geometry, Properties

Further Readings

- Angrist, J. D., & Pischke, J. S. (2008). Mostly harmless econometrics. Princeton university press.
- Casella, G., & Berger, R. L. (2002). Statistical inference (Vol. 2, pp. 337-472). Pacific Grove, CA: Duxbury.
- Tom Shaffer The 42 V's of Big Data and Data Science. https://www.kdnuggets.com/2017/04/42-vs-big-data-data-science.html