

Introducción

Jorge Gallego

Facultad de Economía
Universidad del Rosario

Junio 20 de 2017

Introducción

“Los datos son el nuevo petróleo; como el petróleo, deben refinarse antes de usarse”, Andreas Weigend (Stanford University)

- No dudamos de la importancia de los datos para analizar políticas públicas y tomar decisiones en el sector privado
- Cruciales en las etapas de diseño, formulación, implementación, seguimiento y evaluación
- Pero, ¿cómo usarlos según sea el caso? ¿Qué herramientas tenemos disponibles para sacar el mayor provecho de ellos?
- Veamos dos ejemplos de “juguete” de Kleinberg (2015)

Introducción

¿Podemos hacer llover si invertimos en la “danza de la lluvia”?



Introducción

¿Es beneficioso salir a la calle hoy con una sombrilla?



Introducción

- Dos escenarios de política pública relacionados con el clima
- En los que los datos podrían informar sobre qué decisiones tomar
- En el primer ejemplo es crucial poder explicar. Importa la **causalidad**
- En el segundo es clave poder pronosticar. Importa la **predicción**

Introducción

- En el primer ejemplo es crucial poder explicar. Importa la *causalidad*
- “¿Las danzas de lluvia causan la lluvia?”
- En el segundo es clave poder pronosticar. Importa la *predicción*
- “¿La probabilidad de que llueva es tan grande como para justificar sacar la sombrilla?”

Introducción

A veces se confunden las cosas



BOGOTÁ 18 ENE 2012 - 8:45 AM

'Chamán' fue contratado para que no lloviera en la posesión de Santos

Según Jorge Elías González Vásquez le pagaron tres millones de pesos.

Por: Elespectador.com

COMPARTIDO

60

INSERTAR



CHAMÁN DICE QUE EVITÓ LLUVIA EN LA POSESIÓN DE SANTOS

12.35.05

NOTICIAS PARACOL

Predicción y Causalidad

- Formalicemos algunas de estas ideas
- Sea Y una variable de resultado (e.g. lluvia) que depende de las variables X_0 y X
- El gobernante debe decidir sobre X_0 (e.g. sombrilla o danza) para maximizar una función de bienestar conocida $\pi(X_0, Y)$
- Luego,

$$\frac{d\pi(X_0, Y)}{dX_0} = \frac{\partial\pi}{\partial X_0}(Y) + \frac{\partial\pi}{\partial Y} \frac{\partial Y}{\partial X_0}$$

Predicción y Causalidad

Dos argumentos (incógnitas) interesantes en esta expresión:

1. $\frac{\partial Y}{\partial X_0}$ es el componente de inferencia causal ¿Qué efecto tiene la política sobre el *outcome*?
2. Mientras que $\frac{\partial \pi}{\partial X_0}(Y)$ es el de predicción. ¿Cuál es el efecto de la política sobre el bienestar cuando el *outcome* es Y ?

Cuál sea relevante depende del problema de política pública en cuestión

Predicción y Causalidad

Por lo general hay una restricción de exclusión de por medio

- En el caso de la danza $\frac{\partial \pi}{\partial X_0}(Y) = 0$ y nos interesa la parte causal: si la danza afecta la lluvia $\frac{\partial Y}{\partial X_0}$
- En el caso de la sombrilla $\frac{\partial Y}{\partial X_0} = 0$ y nos interesa la predicción: si llueve o no la sombrilla es necesaria, $\frac{\partial \pi}{\partial X_0}(Y)$

El análisis contemporáneo de políticas se centra en $\frac{\partial Y}{\partial X_0}$. Pero muchos problemas relevantes son de tipo $\frac{\partial \pi}{\partial X_0}(Y)$

Machine Learning y Políticas Públicas

- La batería de herramientas para analizar políticas públicas crece
- Gracias a los desarrollos teóricos, empíricos y computacionales
- Pero también a la disponibilidad de más y mejores datos
- Haremos un recorrido por las principales técnicas del modelaje predictivo
- Y veremos algunas aplicaciones

Definiciones

¿Qué es big data? No hay consenso

- ¿Bases de datos con muchos datos?
- ¿Datos estáticos (digitalizados, e.g. censos) o dinámicos (creados en tiempo real, e.g. redes sociales)?
- *"Real time, socially-created and socially-driven data that could be harvested without having to purposely collect it or budget for its collection"* Raftree, 2015
- Datos que tienen su propia vida

Definiciones

Dos cosas para definir: Big data y Análisis con Big Data

1. Big Data: las tres V's

- ▶ Volúmen
- ▶ Variedad
- ▶ Velocidad

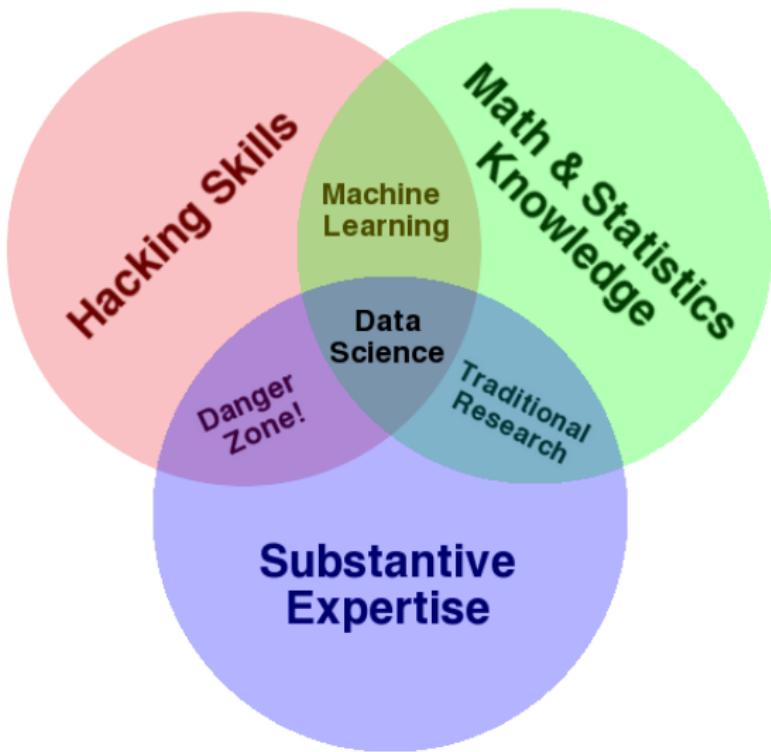
2. Análisis con Big Data

- ▶ Conjunto de herramientas y metodologías
- ▶ Transforman grandes cantidades de datos brutos
- ▶ En datos sobre los datos

Machine Learning

- Es el campo interesado en desarrollar algoritmos para transformar los datos en acción inteligente
- Estudio de sistemas (algoritmos) que mejoran su desempeño con la experiencia
- Como el spam de gmail, los carros sin conductor, Alexa de Amazon, las recomendaciones en Netflix

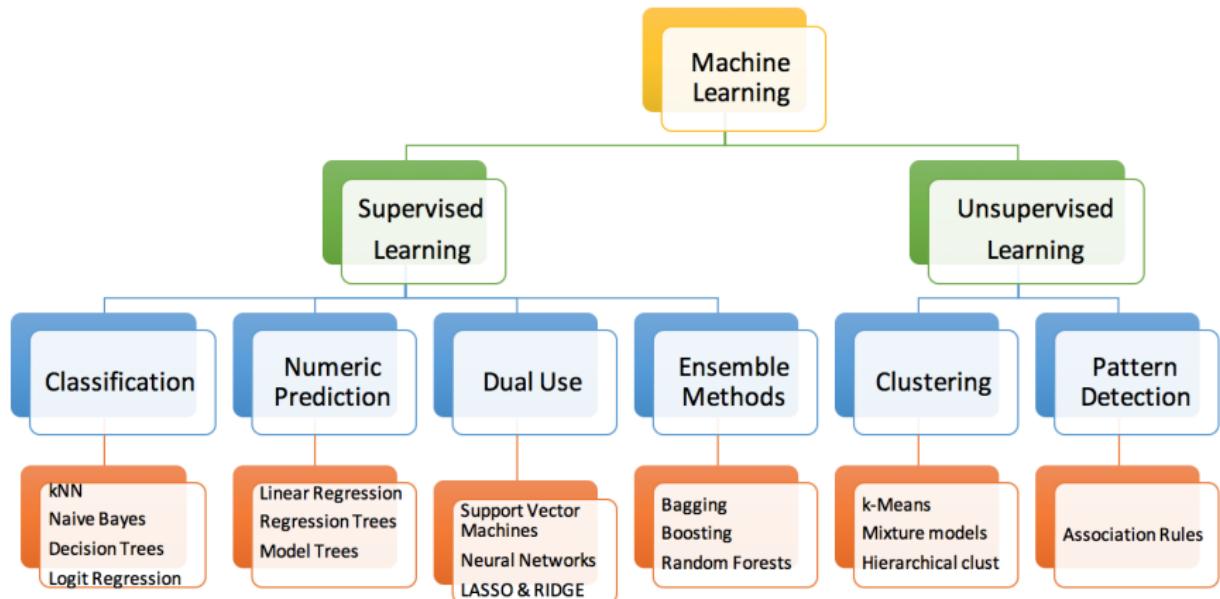
Diagrama de Conway



Machine Learning

- Para hacer *machine learning* combinamos tres cosas:
 1. Datos disponibles
 2. Poder computacional
 3. Métodos estadísticos
- El rápido desarrollo de estos tres campos ha promovido la expansión del *machine learning*

Taxonomía (No-Exhaustiva) de Algoritmos



Algoritmos de *Machine Learning*

- **Modelos predictivos:** predicción de un valor usando otros valores de los datos
- Un modelo de aprendizaje busca descubrir la relación entre una característica (outcome) de interés y las demás
- No siempre tienen que ser predicciones del futuro
- **Aprendizaje supervisado:** optimizar función para encontrar combinación de características que resulten en el outcome

Algoritmos de *Machine Learning*

- **Clasificación:** aprendizaje supervisado para predecir a qué categoría pertenece un caso
 - ▶ Correo spam
 - ▶ Célula cancerosa
 - ▶ Equipo deportivo que ganará o perderá
 - ▶ Aplicante que hará default crediticio
- El outcome de interés a predecir es una variable categórica llamada clase.
- Las categorías posibles son los niveles

Algoritmos de *Machine Learning*

- El aprendizaje supervisado también puede usarse para hacer predicciones numéricas
- Por ejemplo ingreso, valores de laboratorio, puntajes o conteos.
- **Modelo descriptivo:** busca resumir los datos de formas útiles y novedosas
- En contraste con un modelo predictivo, ninguna característica es más importante que las demás
- El proceso de entrenar un modelo descriptivo se conoce como **aprendizaje no-supervisado**

Algoritmos de *Machine Learning*

- Los modelos de aprendizaje no supervisado suelen usarse en minería de datos
- **Descubrimiento de patrones:** se usa para identificar asociaciones útiles dentro de los datos
- Ejemplo: análisis de canastas de mercado. Identifican qué artículos tienden a ser comprados al tiempo
- Patrones en comportamiento fraudulento, defectos genéticos o hot spots de actividad criminal

Algoritmos de *Machine Learning*

- **Clustering:** División de una base de datos en grupos homogéneos.
- **Análisis de segmentación:** clustering que identifica grupos de individuos con comportamiento o info demográfica similar
- Útil para campañas publicitarias. Incluso en política
- La máquina construye clusters. El humano interpreta
- **Meta-aprendizaje:** algoritmos para a aprender a aprender mejor

Algunos Ejemplos

1. Servicio público y corrupción
2. Justicia
3. Seguridad
4. Educación
5. Salud

Servicio Público y Corrupción



Servicio Público y Corrupción

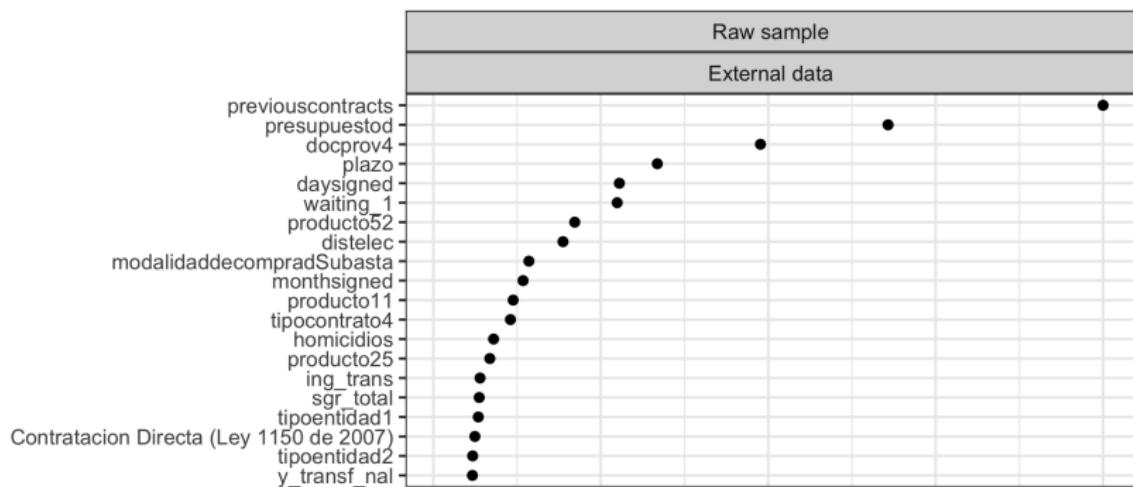
Pregunta de Política Pública

¿Qué tipo de contratos públicos tienen el mayor riesgo de terminar en pleitos judiciales y con fallas en la gestión?

- En Gallego, Rivero y Martínez (2017) usamos la información de contratos públicos en SECOP
- Contratos con prórrogas y adiciones de dinero
- Cruce con otras bases para indagar por controversias judiciales
- Algoritmos para predecir la probabilidad de que un contrato termine “mal”

Servicio Público y Corrupción

Figure: Variable Importance

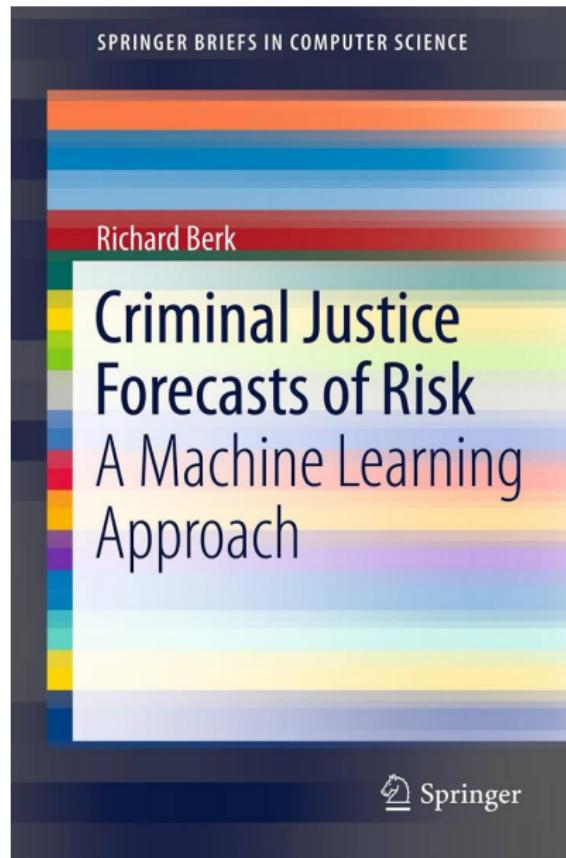




Pregunta de Política Pública

Cuando se captura a un sospechoso de un delito, ¿debe ser enviado a prisión de manera preventiva?

- Decisión crucial para los jueces
- Depende de la predicción de la probabilidad de que el sospechoso cometa un crimen
- ¿Quién predice mejor? ¿El juez o un algoritmo?
- Kleinberg et al. (2015) muestran que con machine learning mejoran las predicciones de los jueces y se reduce el crimen



Seguridad



Daniel E. Ortega

@dortegaeval

Siguiendo

#PuntosCalientesBogotá "Ni en Europa ni en EE.
UU. ha habido una intervención en puntos
calientes de esa magnitud" goo.gl/9R9Y3j



RETWEET

1

ME GUSTA

4

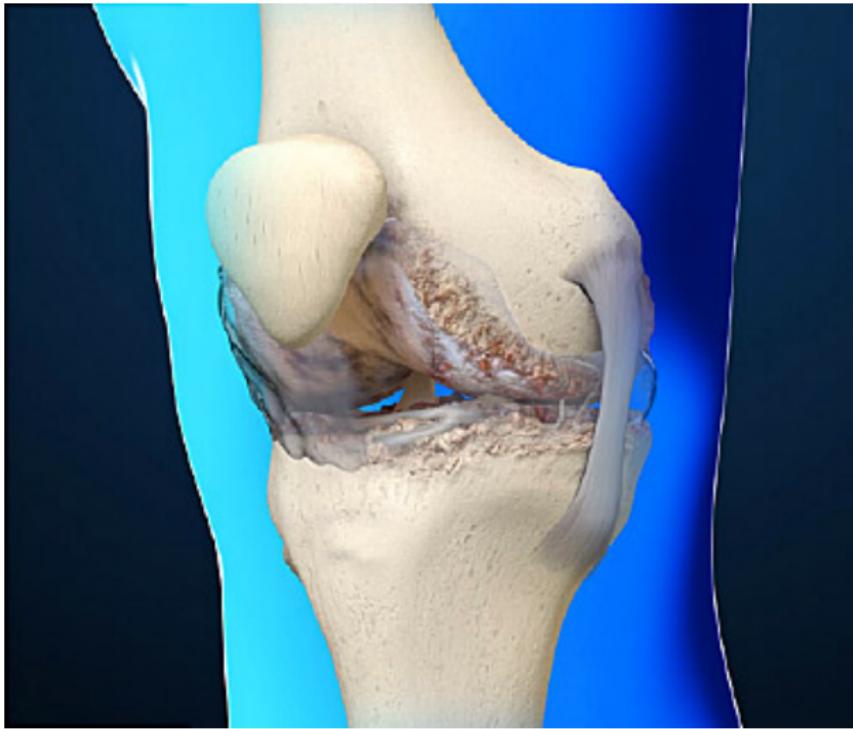


7:05 - 24 ene. 2017

Pregunta de Política Pública

¿Qué tipo de estrategias policivas son más eficientes para la prevención de la criminalidad en las grandes ciudades?

- Hotspots en grandes ciudades. Big data y machine learning para predecir crimen
- Cámaras de seguridad, patrullaje, carteles disuasivos, etc.
- Se predice dónde ocurrirán los crímenes y se aleatorizan estrategias



Pregunta de Política Pública

¿Cómo asignar un tratamiento médico costoso a grupos de la población con alto riesgo de muerte?

- Tratamiento contra la osteoartritis
- Cirugía costosa y complicada cuyos beneficios tardan en llegar
- ¿Tiene sentido proveer el servicio a pacientes con baja expectativa de vida?
- Kleinberg et al. (2015) desarrollan un algoritmos para predecir sobre quiénes tiene más sentido proveer el tratamiento

Mis Estudiantes...

No olviden visitar los posters



Predicción de deserción universitaria y consumo de sustancias psicoactivas

Conclusiones

- La importancia de los datos para analizar políticas públicas es creciente
- Aumentan las fuentes de información. Aumentan los métodos para analizarlos
- Pero es importante saber cuál es el objetivo. Explicar o predecir.
- Porque según sea el caso los métodos cambian