Aprendizaje Dirigido para Estimación en el marco de la Enfermedad de Alzheimer Esporádica

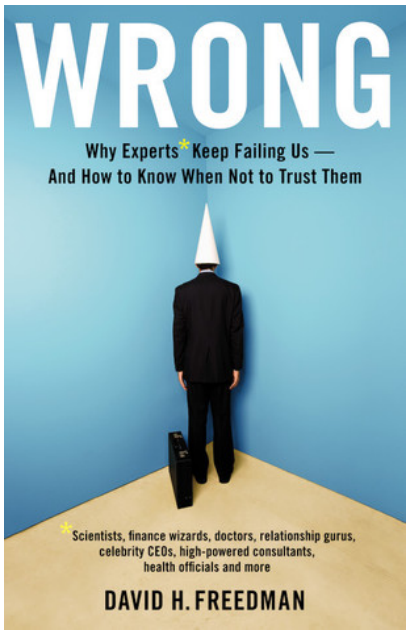
Dra. Laura Ación

Instituto de Cálculo
FCEN, UBA – CONICET
laura.acion@ic.fcen.uba.ar

Biometría II
7 de Noviembre de 2016

Problema

- Es común que haya resultados de investigación que no son reproducibles y/o replicables.
- Cuando los resultados no reproducibles/replicables provienen de estudios de investigación grandes y bien diseñados, la falta de reproducibilidad y/o replicación puede derivar en falta de credibilidad en la estadística utilizada para obtener esos resultados.
- Estas contradicciones derivan en publicaciones tales como:



Los Sospechosos de Siempre

- **Sesgo en los efectos estimados** suele ser el resultado de imponer modelos mal especificados a los datos.
- Para el análisis de datos es usual proponer varios modelos candidatos y elegir un modelo final usando algún algoritmo de selección. Habitualmente se reporta un único modelo final.
- Este tipo de práctica usualmente produce **errores estándar inválidos** y **sobreajuste** que pueden causar descubrimientos falsos y baja tasa de reproducibilidad de resultados.

Una Solución: Aprendizaje Dirigido

- Aprendizaje dirigido (*Ad*, targeted learning) es una herramienta desarrollada por Mark van der Laan y equipo en la última década.
- *Ad* es una herramienta novedosa que combina el uso de:
  - Estadística clásica
    - Modelos lineales generalizados (Biometría II)
    - Estimadores semi-paramétricos
    - Intervalos de confianza, tests de hipótesis (Biometría I)
  - Aprendizaje automático (machine learning), más común en ciencias de la computación (data mining).

Etapas de *Ad*

1. Datos: n observaciones iid de la variable $O = (Y, A, W) \sim P_0$.
2. Modelo estadístico: conjunto de posibles distribuciones de O.
3. Parámetro objetivo ($\varphi(P_0)$ – targeted parameter): cualquier función de P_0, por ejemplo un odds ratio (o cualquier medida de efecto de interés que aplique al problema y al tipo de datos).
4. Súper aprendizaje (SA): Estima la parte relevante de P_0, usando aprendizaje automático de conjunto (ensemble machine learning).
5. Máxima verosimilitud dirigida: Actualiza el estimador anterior, logra un balance óptimo de sesgo-varianza para el parámetro objetivo.
6. Inferencia e interpretación: Se calcula el error estándar para la estimación del parámetro objetivo y se interpreta los resultados en forma habitual.

¿Por qué un Parámetro Objetivo…

…y no todo un modelo con muchos parámetros?

- Habitualmente se estima todo un modelo y no un único parámetro.
- Cuando se estima todo un modelo, se estiman porciones de P_0 que no son necesarias.
- En general, estimar únicamente la porción relevante de la distribución de probabilidad provee estimadores con la menor variabilidad y sesgo posibles.
- Único supuesto de *Ad*: Los datos observados pueden ser representados como n observaciones de $O = (Y, A, W)$ iid.

Abordaje General para Estimar $\varphi(P_0)$

- Buscamos un estimador del parámetro objetivo ($\varphi(P_0)$) que aprenda de los datos en lugar de usar un modelo paramétrico (por ej, algún modelo lineal generalizado elegido por el analista) que asume que los datos siguen cierta estructura.
- El objetivo es encontrar el mejor estimador del parámetro de interés adaptándose a los datos sin sobreajustarlos.
- El "mejor" estimador resultará del método que minimice alguna función de pérdida (por ejemplo, el error cuadrático medio – ECM).

Súper Aprendizaje (Etapa 4)

- Súper aprendizaje (SA, super learning) es un método de conjunto para aprendizaje automático (ensemble machine learning) que permite usar múltiples métodos y modelos para establecer el mejor estimador del parámetro objetivo.
- El uso de varios métodos simultáneamente le permite a SA tener mejor desempeño en la estimación que cualquiera de los métodos individuales.
- Para evitar el sobreajuste, SA usa validación cruzada (VC).

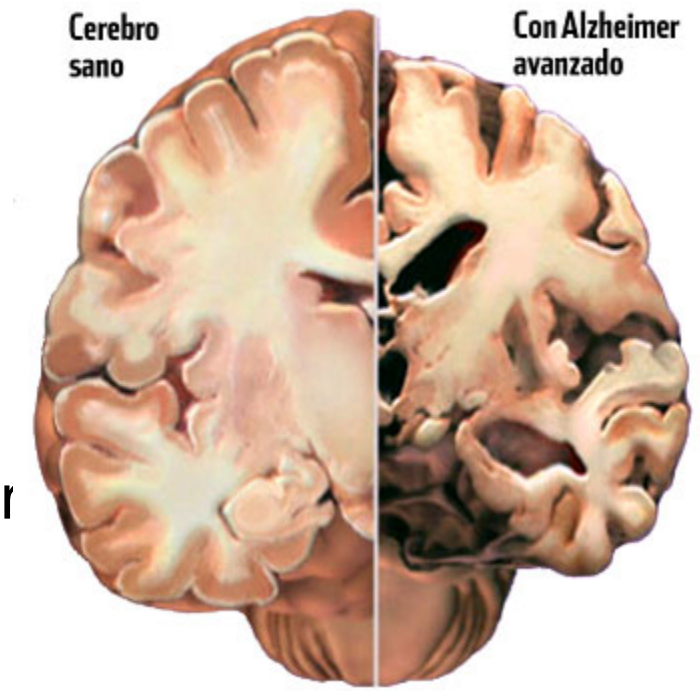
Súper Aprendizaje: Validación Cruzada

Validación cruzada:

1. Se ajustan todos los modelos usando, por ejemplo, 90% de la muestra para el ajuste y un 10% de la muestra para evaluar el ajuste.
2. Se repite esto 9 veces más usando en cada iteración un conjunto de datos para evaluación de los modelos disjunto del conjunto de datos de la iteración anterior.
3. Para cada modelo se calcula el error cuadrático medio (ECM de VC) usando los valores predichos resultantes de cada iteración.
4. Se elige el modelo/algoritmo con el menor ECM.

Ejemplo: Enfermedad de Alzheimer Esporádica

- La enfermedad de Alzheimer esporádica (EAE) es la causa más común de demencia.
- EAE es un desorden genético complejo en el cual el alelo ε4 de la apolipoproteína E (APOE) es la variante de riesgo genético de mayor influencia.
- Recientemente, estudios que abarcan todo el genoma (GWAS) muestran más de 100 variantes genéticas (SNPs) que modifican, en menor medida que ε4, el riesgo de EAE.



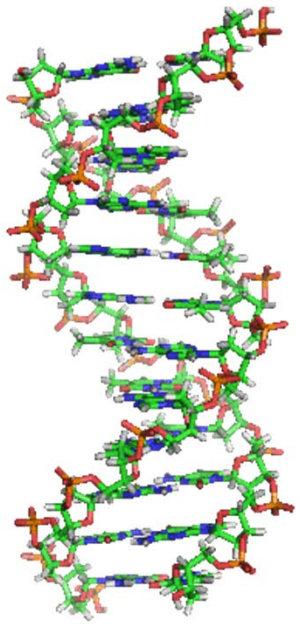
Ad en el Marco de la EAE: Objetivo

- El objetivo de este trabajo es aplicar *Ad* para la estimación del efecto de APOE4 en EAE.

- Se ajustó por género, edad, ancestría y otros marcadores genéticos elegidos por su habilidad para predecir EAE en estudios previos.

Ad en el Marco de la EAE: Métodos

- Usamos los datos de un estudio realizado en población argentina mayor de 65 años (218 casos con diagnóstico clínico de EAE y 175 controles sin deterioro cognitivo).
- Se determinaron los siguientes SNPs:
  - rs429358 (APOE4)
  - rs3764650 (ABCA7)
  - rs3818361 (CR1)
  - rs3851179 (PICALM)
  - rs610932 (MS4A6A)



Ad en EAE: Descripción de la Muestra

	Grupo		p-valor
	EAE n = 218	Control n = 175	
Género, n (%), mujeres	142 (65,1)	117 (66,9)	0,72
Edad, media (DE), años	75,6 (5,4)	77,8 (6,2)	< 0,001
Ancestría, media (DE), PC1	0,8 (0,2)	0,8 (0,2)	0,78
Presencia de Variante:			
rs429358 (APOE4), n (%)	105 (48,2)	31 (17,7)	< 0,0001
rs3764650 (ABCA7), n (%)	56 (25,7)	39 (22,3)	0,43
rs3818361 (CR1), n (%)	77 (35,3)	65 (37,1)	0,71
rs3851179 (PICALM), n (%)	131 (60,1)	102 (58,3)	0,72
rs610932 (MS4A6A), n (%)	149 (68,4)	113 (64,6)	0,43

Ad en el Marco de la EAE: Métodos

- Los datos se analizaron mediante dos métodos:
  - *Regresión Logística Clásica.*
    - Variable Dependiente: Presencia/Ausencia de EAE.
    - Variables Independientes: rs429358 (APOE4), rs3764650 (ABCA7), rs3818361 (CR1), rs3851179 (PICALM), rs610932 (MS4A6A), género, ancestría y edad.
  - *Aprendizaje Dirigido.*
    - Parámetro Objetivo (targeted parameter): OR ajustado para rs429358 (APOE4).
    - Variable Dependiente: Presencia/Ausencia de EAE.
    - Covariables de Ajuste: rs3764650 (ABCA7), rs3818361 (CR1), rs3851179 (PICALM), rs610932 (MS4A6A), género, ancestría y edad.

Ad en el Marco de la EAE: Métodos

- Etapa 4: Súper aprendizaje admite el uso simultáneo de varios algoritmos.
  - Se utilizaron:
    - Regresión logística
    - Regresión logística Bayesiana
    - Máquinas de vectores soporte
    - Redes neuronales artificales
    - Arboles de regresión y partición recursiva
    - Modelos de regresión aditivos

Súper Aprendizaje: Propiedades

- Incorporando una colección rica de métodos y modelos con distintos niveles de sesgo y ajuste, la validación cruzada previene la selección de:
  - Un modelo sobreajustado
  - Un modelo con parámetros sesgados
- Se puede demostrar que SA se desempeña tan bien como el método que minimiza la esperanza de la función de pérdida.
- En simulaciones en que la colección de métodos contiene el modelo que generó los datos, SA lo elige.
- En ese caso, la variabilidad de SA es mayor que la del modelo paramétrico.

Máxima Verosimilitud Dirigida (Etapa 5)

- Problema: El estimador generado por súper aprendizaje no está dirigido al parámetro de interés y es sesgado.
- Solución: Aplicar máxima verosimilitud dirigida (MVD).
  - MVD no tiene ninguna relación con la estimación por máxima verosimilitud clásica.
- Se puede demostrar que usando MVD se puede reducir el sesgo del estimador generado por súper aprendizaje.
- MVD actualiza el estimador obtenido por SA.
- El estimador resultante tiene un equilibrio óptimo entre sesgo y varianza para el parámetro objetivo.

Ad en el Marco de la EAE: Resultados

Regresión Logística Clásica	Aprendizaje Dirigido
Modelo: $logit(\hat{p}_{EAE}) =$ $-1{,}75 + 0{,}01 * mujer$ $-0{,}33 * edad$ $+0{,}06 * ancestría$ $-0{,}74 * rs429358$ $-0{,}07 * rs3764650$ $+0{,}09 * rs3818361$ $-0{,}10 * rs3851179$ $-0{,}11 * rs610932$	*Modelo que minimiza ECM VC:* combinación lineal de modelos de regresión aditiva, de regresión logística Bayesiana y árboles de regresión y partición recursiva.
OR ajustado: 4,40	*OR ajustado:* 4,00
Intervalo de Confianza (IC): [2,7 – 7,2]	*IC:* [2,6 – 6,5]
Ancho del IC: 4,5	*Ancho del IC:* 3,9

Ad en el Marco de la EAE: Conclusiones

Este ejemplo de uso de *Ad* para estimación en EAE muestra que:

- ***Ad* no elige el modelo de regresión logística clásico** como el mejor modelo para este ejemplo.
- El **OR ajustado para el efecto de APOE4 en EAE** estimado por regresión logística es **casi un 10% mayor** que el estimado por *Ad*.
- **El intervalo de confianza para el OR ajustado** estimado por regresión logística es **más de un 10% más ancho** que el calculado por *Ad*.

Otras Aplicaciones de *Ad*

Ad permite el análisis de varios tipos de datos y diseños estadísticos:

- Ensayos clínicos tradicionales o con diseños de aleatorización adaptativa.
- Estudios caso-control independientes o apareados.
- Diseños con datos de supervivencia o de tiempo hasta un evento con datos censurados.
- Big data genómica y descubrimiento de biomarcadores.
- Datos longitudinales con datos faltantes, y covariables invariantes y variantes con el tiempo.

Ventajas de *Ad*

- Sólida base teórica y empírica a través de simulaciones que demuestran la igualdad y, frecuentemente, superioridad de *Ad* sobre otras alternativas analíticas en lo que a sesgo, variabilidad y sobreajuste se refiere.
- Alta divulgación y número creciente de publicaciones aplicando la técnica a problemas reales.
- Disponible en R gratuitamente.
  - Paquetes SuperLearner y tmle.

Desventajas de *Ad*

- Desde el punto de vista de la estadística clásica, requiere un cambio importante en la forma en que se piensa y realiza el análisis.
- La diversidad de métodos incorporados por *Ad* requiere, como mínimo, un conocimiento básico de una multiplicidad de tipos de análisis que no suelen estar en la caja de herramientas del analista promedio.
- La teoría involucrada en la etapa de máxima verosimilitud dirigida es muy compleja.
- *Ad* puede requerir mayor tiempo computacional que otros métodos analíticos.

Equipo de Investigación

- **Instituto de Cálculo, FCEN, UBA**

  Dra. Diana Kelmansky

- **Fundación Instituto Leloir**

  Dra. Carolina Dalmasso

  Dra. Laura Morelli

- **Instituto de Investigaciones Médicas A. Lanari y CENECON, F. de Medicina, UBA**

  Dr. Luis I. Brusco

  Dra. Carolina Muchnik

  Dra. Natividad Olivar

- **Harvard Medical School**

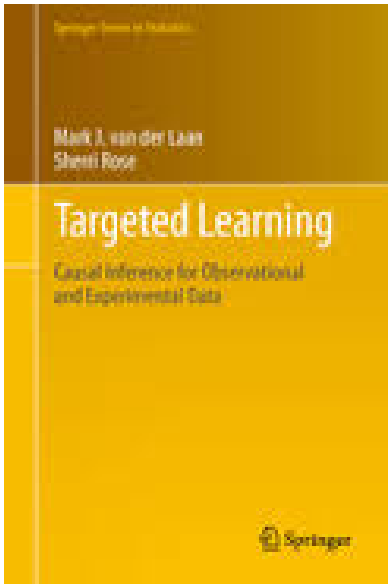
  Dra. Sherri Rose

Referencias

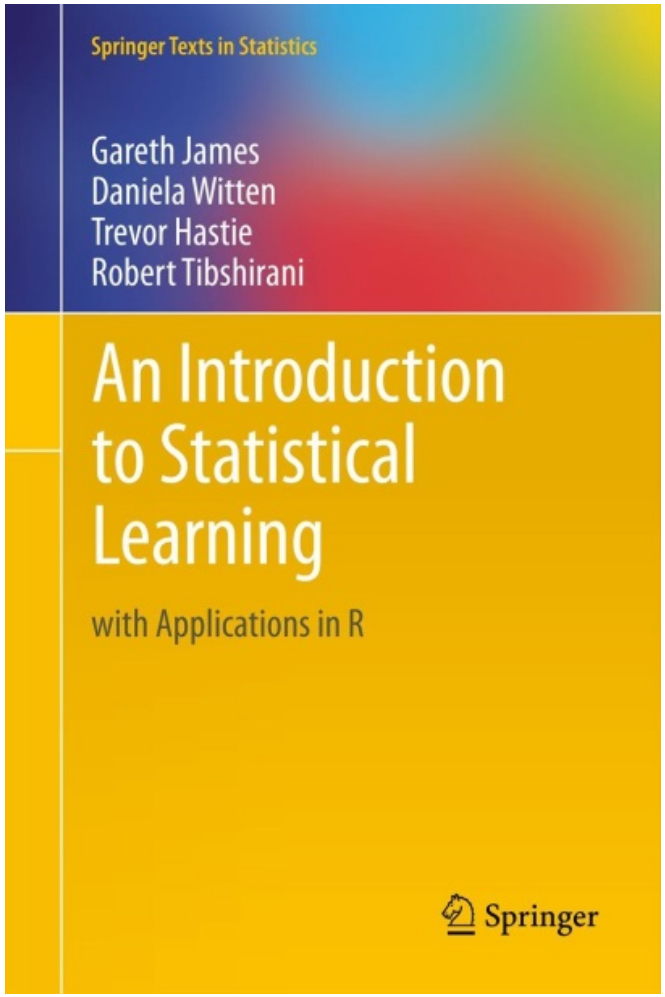
1. Van der Laan M, Rose S. Targeted learning: causal inference for observational and experimental data: Springer Science and Business Media; 2011.

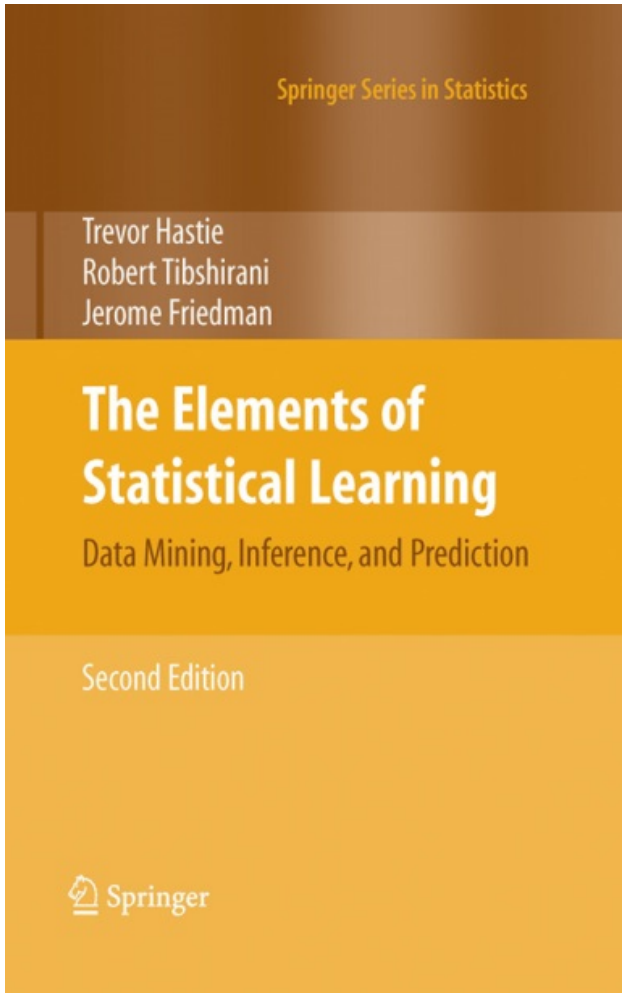
2. Polley E, van der Laan M. SuperLearner: Super Learner Prediction, Package Version 2.0-4. Vienna: R Foundation for Statistical Computing. 2011.

3. Gruber S, van der Laan M. tmle: an R package for targeted maximum likelihood estimation. 2011.



Otras Referencias Útiles






GRACIAS
MAKE GIFS AT GIFSOUP.COM