Aprendizaje Dirigido

Seminario

Laura Ación
en colaboración con Diana Kelmansky

Instituto de Cálculo
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
lacion@gmail.com

14 de Diciembre de 2015

Problema

- Es común que haya resultados de investigación que no son reproducibles.

Problema

- Es común que haya resultados de investigación que no son reproducibles.
- Cuando los resultados no reproducibles provienen de estudios de investigación grandes y bien diseñados, la falta de reproducibilidad puede derivar en falta de credibilidad en la estadística utilizada para obtener esos resultados.

Problema

- Es común que haya resultados de investigación que no son reproducibles.
- Cuando los resultados no reproducibles provienen de estudios de investigación grandes y bien diseñados, la falta de reproducibilidad puede derivar en falta de credibilidad en la estadística utilizada para obtener esos resultados.
- Estas contradicciones derivan en publicaciones de alto impacto tales como:
  - Contradicted and initially stronger effects in highly cited clinical research. JAMA, 2005.
  - Believe it or not: how much can we rely on published data on potential drug targets? Nat Rev Drug Discov, 2011.
  - Wrong: why experts keep failing us and how to know when not to trust them. Little, Brown & Co, NY, 2010.

Los Sospechosos de Siempre

- **Sesgo en los efectos estimados**, **errores estándar inválidos** y/o **sobreajuste**.

Los Sospechosos de Siempre

- **Sesgo en los efectos estimados**, **errores estándar inválidos** y/o **sobreajuste**.
- El **sesgo en los efectos estimados** suele ser el resultado de imponer modelos mal especificados a los datos cuando, usualmente, el mecanismo que genera los datos es desconocido.

Los Sospechosos de Siempre

- **Sesgo en los efectos estimados**, **errores estándar inválidos** y/o **sobreajuste**.
- El **sesgo en los efectos estimados** suele ser el resultado de imponer modelos mal especificados a los datos cuando, usualmente, el mecanismo que genera los datos es desconocido.
- Para el análisis de datos es usual proponer varios modelos candidatos y elegir un modelo final usando algún algoritmo de selección.

Los Sospechosos de Siempre

- **Sesgo en los efectos estimados**, **errores estándar inválidos** y/o **sobreajuste**.
- El **sesgo en los efectos estimados** suele ser el resultado de imponer modelos mal especificados a los datos cuando, usualmente, el mecanismo que genera los datos es desconocido.
- Para el análisis de datos es usual proponer varios modelos candidatos y elegir un modelo final usando algún algoritmo de selección.
- Cuando se ajusta más de un modelo, habitualmente se reporta un único modelo final.

Los Sospechosos de Siempre

- **Sesgo en los efectos estimados**, **errores estándar inválidos** y/o **sobreajuste**.
- El **sesgo en los efectos estimados** suele ser el resultado de imponer modelos mal especificados a los datos cuando, usualmente, el mecanismo que genera los datos es desconocido.
- Para el análisis de datos es usual proponer varios modelos candidatos y elegir un modelo final usando algún algoritmo de selección.
- Cuando se ajusta más de un modelo, habitualmente se reporta un único modelo final.
- Ningún otro modelo ajustado y puesto a prueba se tiene en consideración en la inferencia final.

Los Sospechosos de Siempre

- **Sesgo en los efectos estimados**, **errores estándar inválidos** y/o **sobreajuste**.
- El **sesgo en los efectos estimados** suele ser el resultado de imponer modelos mal especificados a los datos cuando, usualmente, el mecanismo que genera los datos es desconocido.
- Para el análisis de datos es usual proponer varios modelos candidatos y elegir un modelo final usando algún algoritmo de selección.
- Cuando se ajusta más de un modelo, habitualmente se reporta un único modelo final.
- Ningún otro modelo ajustado y puesto a prueba se tiene en consideración en la inferencia 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 en su versión original) es una herramienta desarrollada por Mark van der Laan y equipo en la última década.

Una Solución: Aprendizaje Dirigido

- Aprendizaje dirigido (AD, targeted learning en su versión original) 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 y aprendizaje automático (machine learning), más común en ciencias de la computación.

Etapas de AD

1. **Datos:** n observaciones iid de la variable $O \sim P_0$.

Etapas de AD

1. **Datos:** n observaciones iid de la variable $O \sim P_0$.

2. **Modelo estadístico:** $\mathcal{M}$, conjunto de posibles distribuciones de O, $P_0 \in \mathcal{M}$.

Etapas de AD

1. **Datos:** n observaciones iid de la variable $O \sim P_0$.

2. **Modelo estadístico:** $\mathcal{M}$, conjunto de posibles distribuciones de O, $P_0 \in \mathcal{M}$.

3. **Parámetro objetivo:** $\Psi(P_0)$ es una característica de P_0, Ψ mapea P_0 en el parámetro de interés.

Etapas de AD

1. **Datos:** n observaciones iid de la variable $O \sim P_0$.

2. **Modelo estadístico:** $\mathcal{M}$, conjunto de posibles distribuciones de O, $P_0 \in \mathcal{M}$.

3. **Parámetro objetivo:** $\Psi(P_0)$ es una característica de P_0, Ψ mapea P_0 en el parámetro de interés.

4. **Súper aprendizaje (SA):** Estima la parte relevante de P_0, usando aprendizaje automático de conjunto.

Etapas de AD

1. **Datos:** n observaciones iid de la variable $O \sim P_0$.

2. **Modelo estadístico:** $\mathcal{M}$, conjunto de posibles distribuciones de O, $P_0 \in \mathcal{M}$.

3. **Parámetro objetivo:** $\Psi(P_0)$ es una característica de P_0, Ψ mapea P_0 en el parámetro de interés.

4. **Súper aprendizaje (SA):** Estima la parte relevante de P_0, usando aprendizaje automático de conjunto.

5. **Máxima verosimilitud dirigida:** Actualiza el estimador anterior, logra un balance óptimo de sesgo-varianza para $\Psi(P_0)$.

Etapas de AD

1. **Datos:** n observaciones iid de la variable $O \sim P_0$.

2. **Modelo estadístico:** $\mathcal{M}$, conjunto de posibles distribuciones de O, $P_0 \in \mathcal{M}$.

3. **Parámetro objetivo:** $\Psi(P_0)$ es una característica de P_0, Ψ mapea P_0 en el parámetro de interés.

4. **Súper aprendizaje (SA):** Estima la parte relevante de P_0, usando aprendizaje automático de conjunto.

5. **Máxima verosimilitud dirigida:** Actualiza el estimador anterior, logra un balance óptimo de sesgo-varianza para $\Psi(P_0)$.

6. **Inferencia e interpretación:** Se calcula el error estándar para la estimación de $\Psi(P_0)$ y se interpreta los resultados en forma habitual.

Datos y Modelo Estadístico

Siguiendo Van der Laan y Rose (2011), sea O una variable aleatoria tal que $O = (Y, A, W)$ y $O \sim P_0$. Donde:

Datos y Modelo Estadístico

Siguiendo Van der Laan y Rose (2011), sea O una variable aleatoria tal que $O = (Y, A, W)$ y $O \sim P_0$. Donde:

- Y es una característica binaria (ej.: mortalidad dentro de los próximos 5 años cuyos posibles valores son "Sí" y "No").

Datos y Modelo Estadístico

Siguiendo Van der Laan y Rose (2011), sea O una variable aleatoria tal que $O = (Y, A, W)$ y $O \sim P_0$. Donde:

- Y es una característica binaria (ej.: mortalidad dentro de los próximos 5 años cuyos posibles valores son "Sí" y "No").
- A es una exposición binaria (ej.: $A = 1$ si se realiza cierto nivel de actividad física, $A = 0$ si no).

Datos y Modelo Estadístico

Siguiendo Van der Laan y Rose (2011), sea O una variable aleatoria tal que $O = (Y, A, W)$ y $O \sim P_0$. Donde:

- Y es una característica binaria (ej.: mortalidad dentro de los próximos 5 años cuyos posibles valores son "Sí" y "No").
- A es una exposición binaria (ej.: $A = 1$ si se realiza cierto nivel de actividad física, $A = 0$ si no).
- $W = \{W_1, ..., W_k\}$ es el conjunto de k (ej. $k = 3$) potenciales covariables en la relación entre A e Y (ej.: edad, género, presencia/ausencia de enfermedades crónicas).

Datos y Modelo Estadístico

Siguiendo Van der Laan y Rose (2011), sea O una variable aleatoria tal que $O = (Y, A, W)$ y $O \sim P_0$. Donde:

- Y es una característica binaria (ej.: mortalidad dentro de los próximos 5 años cuyos posibles valores son "Sí" y "No").
- A es una exposición binaria (ej.: $A = 1$ si se realiza cierto nivel de actividad física, $A = 0$ si no).
- $W = \{W_1, ..., W_k\}$ es el conjunto de k (ej. $k = 3$) potenciales covariables en la relación entre A e Y (ej.: edad, género, presencia/ausencia de enfermedades crónicas).
- $P_0 \in \mathcal{M}$, con $\mathcal{M}$ un modelo estadístico, conjunto de todas las posibles distribuciones de probabilidad para P_0.

Datos y Modelo Estadístico

Siguiendo Van der Laan y Rose (2011), sea O una variable aleatoria tal que $O = (Y, A, W)$ y $O \sim P_0$. Donde:

- Y es una característica binaria (ej.: mortalidad dentro de los próximos 5 años cuyos posibles valores son "Sí" y "No").
- A es una exposición binaria (ej.: $A = 1$ si se realiza cierto nivel de actividad física, $A = 0$ si no).
- $W = \{W_1, ..., W_k\}$ es el conjunto de k (ej. $k = 3$) potenciales covariables en la relación entre A e Y (ej.: edad, género, presencia/ausencia de enfermedades crónicas).
- $P_0 \in \mathcal{M}$, con $\mathcal{M}$ un modelo estadístico, conjunto de todas las posibles distribuciones de probabilidad para P_0.
- $\mathcal{M}$ puede ser no paramétrico, semi-paramétrico o paramétrico.

Parámetro Objetivo

- Parámetro objetivo (targeted parameter), $\Psi : \mathcal{M} \to \mathbb{R}$.

Parámetro Objetivo

- Parámetro objetivo (targeted parameter), $\Psi : \mathcal{M} \to \mathbb{R}$.
- Por ejemplo, supongamos que queremos estimar la **diferencia de riesgo** de muerte en los próximos 5 años entre adultos mayores que realizan actividad física y aquéllos que no la realizan.

Parámetro Objetivo

- Parámetro objetivo (targeted parameter), $\Psi : \mathcal{M} \to \mathbb{R}$.
- Por ejemplo, supongamos que queremos estimar la **diferencia de riesgo** de muerte en los próximos 5 años entre adultos mayores que realizan actividad física y aquéllos que no la realizan.
- $a_i = 1$ indica que el sujeto i realiza actividad física y $a_i = 0$ indica que un sujeto no realiza actividad física.
- En este caso $\Psi(P_0)$ es:

$$\Psi(P_0) = E_{W,0}[E_0(Y|A = 1, W) - E_0(Y|A = 0, W)].$$

Parámetro Objetivo

- Parámetro objetivo (targeted parameter), $\Psi : \mathcal{M} \to \mathbb{R}$.
- Por ejemplo, supongamos que queremos estimar la **diferencia de riesgo** de muerte en los próximos 5 años entre adultos mayores que realizan actividad física y aquéllos que no la realizan.
- $a_i = 1$ indica que el sujeto i realiza actividad física y $a_i = 0$ indica que un sujeto no realiza actividad física.
- En este caso $\Psi(P_0)$ es:

$$\Psi(P_0) = E_{W,0}[E_0(Y|A = 1, W) - E_0(Y|A = 0, W)].$$

$\Psi(P_0)$ depende de P_0 únicamente a través de:

$$\bar{Q}_0(A, W) = E_0(Y|A, W) \text{ y de } Q_{W,0} = P_0(W = w).$$

¿Por qué un Parámetro Objetivo?

- Se podría estimar la distribución condicional completa de Y en lugar de estimar únicamente la media condicional de Y.

¿Por qué un Parámetro Objetivo?

- Se podría estimar la distribución condicional completa de Y en lugar de estimar únicamente la media condicional de Y.
- De ser así, en la mayoría de los casos se estiman porciones de la densidad que no son necesarias.

¿Por qué un Parámetro Objetivo?

- Se podría estimar la distribución condicional completa de Y en lugar de estimar únicamente la media condicional de Y.
- De ser así, en la mayoría de los casos se estiman porciones de la densidad que no son necesarias.
- En general, estimar únicamente la porción relevante de la distribución de probabilidad provee estimadores con la mayor eficiencia y menor sesgo posibles.

¿Por qué un Parámetro Objetivo?

- Se podría estimar la distribución condicional completa de Y en lugar de estimar únicamente la media condicional de Y.
- De ser así, en la mayoría de los casos se estiman porciones de la densidad que no son necesarias.
- En general, estimar únicamente la porción relevante de la distribución de probabilidad provee estimadores con la mayor eficiencia y menor sesgo posibles.
- **Único supuesto:** Los datos observados pueden ser representados como n observaciones de $O = (Y, A, W)$ iid.

Abordaje General para Estimar $\bar{Q}_0(A, W)$

- Parámetro de interés: $\bar{Q}_0(A, W) = E_0(Y|A, W)$.
- Para evitar estimar $\bar{Q}_0(A, W)$ imponiendo un modelo paramétrico, buscamos un método automático no-paramétrico (o semi-paramétrico) que lo estime.

Abordaje General para Estimar $\bar{Q}_0(A, W)$

- Parámetro de interés: $\bar{Q}_0(A, W) = E_0(Y|A, W)$.
- Para evitar estimar $\bar{Q}_0(A, W)$ imponiendo un modelo paramétrico, buscamos un método automático no-paramétrico (o semi-paramétrico) que lo estime.
- Queremos un estimador que aprenda de los datos.

Abordaje General para Estimar $\bar{Q}_0(A, W)$

- Parámetro de interés: $\bar{Q}_0(A, W) = E_0(Y|A, W)$.
- Para evitar estimar $\bar{Q}_0(A, W)$ imponiendo un modelo paramétrico, buscamos un método automático no-paramétrico (o semi-paramétrico) que lo estime.
- Queremos un estimador que aprenda de los datos.
- La idea es encontrar el mejor estimador del parámetro de interés (o parámetro objetivo) adaptándose a los datos sin sobreajustarlos.

Abordaje General para Estimar $\bar{Q}_0(A, W)$

- Parámetro de interés: $\bar{Q}_0(A, W) = E_0(Y|A, W)$.
- Para evitar estimar $\bar{Q}_0(A, W)$ imponiendo un modelo paramétrico, buscamos un método automático no-paramétrico (o semi-paramétrico) que lo estime.
- Queremos un estimador que aprenda de los datos.
- La idea es encontrar el mejor estimador del parámetro de interés (o parámetro objetivo) adaptándose a los datos sin sobreajustarlos.
- Ejemplos de métodos que hacen esto: locally weighted regression and scatterplot smoothing (loess), regresión polinomial pesada, funciones de splines, splines con smoothing, redes neuronales, random forest, etc.

¿Cómo Elegir el Método "Correcto"?

- Los métodos pueden diferir en varios factores (covariables que se usan, función de pérdida que se usa para evaluarlo, algoritmo de búsqueda, etc.).

¿Cómo Elegir el Método "Correcto"?

- Los métodos pueden diferir en varios factores (covariables que se usan, función de pérdida que se usa para evaluarlo, algoritmo de búsqueda, etc.).
- El "mejor" método será el que minimice una función de pérdida.

$$L : (O, \bar{Q}) \rightarrow L(O, \bar{Q}) \in \mathbb{R}.$$

¿Cómo Elegir el Método "Correcto"?

- Los métodos pueden diferir en varios factores (covariables que se usan, función de pérdida que se usa para evaluarlo, algoritmo de búsqueda, etc.).
- El "mejor" método será el que minimice una función de pérdida.

$$L : (O, \bar{Q}) \rightarrow L(O, \bar{Q}) \in \mathbb{R}.$$

- En nuestro ejemplo:

$$\bar{Q}_0 = E_0(Y|A, W) = \text{arg min}_{\bar{Q}} E_0(L(O, \bar{Q})), \text{ donde}$$
$$L(O, \bar{Q}) = (Y - \bar{Q}_{A,W})^2.$$

¿Cómo Elegir el Método "Correcto"?

- Los métodos pueden diferir en varios factores (covariables que se usan, función de pérdida que se usa para evaluarlo, algoritmo de búsqueda, etc.).

- El "mejor" método será el que minimice una función de pérdida.

$$L : (O, \bar{Q}) \rightarrow L(O, \bar{Q}) \in \mathbb{R}.$$

- En nuestro ejemplo:

$$\bar{Q}_0 = E_0(Y|A, W) = \arg\min_{\bar{Q}} E_0(L(O, \bar{Q})), \text{ donde}$$
$$L(O, \bar{Q}) = (Y - \bar{Q}_{A,W})^2.$$

- Dadas dos estimaciones $\bar{Q}_n^a$ y $\bar{Q}_n^b$, buscamos elegir aquélla para la cual $E_0(L(O, \bar{Q}_n))$ sea menor.

Súper Aprendizaje

- Súper aprendizaje (SA, super learning en su versión original) es un método de conjunto para aprendizaje automático (ensemble machine learning) que permite usar múltiples métodos para establecer el mejor estimador de $\bar{Q}_0(A, W)$.

Súper Aprendizaje

- Súper aprendizaje (SA, super learning en su versión original) es un método de conjunto para aprendizaje automático (ensemble machine learning) que permite usar múltiples métodos para establecer el mejor estimador de $\bar{Q}_0(A, W)$.
- El uso de varios métodos simultáneamente le permite a SA tener mejor desempeño en la estimación que cualquiera de los métodos individuales.

Súper Aprendizaje

- Súper aprendizaje (SA, super learning en su versión original) es un método de conjunto para aprendizaje automático (ensemble machine learning) que permite usar múltiples métodos para establecer el mejor estimador de $\bar{Q}_0(A, W)$.
- 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).

SA: Ejemplo

- Supongamos que queremos predecir la tasa de mortalidad de personas mayores (Y) usando actividad física (A) controlando por otras covariables como edad (W_1), género (W_2) y presencia de enfermedades crónicas (W_3).

SA: Ejemplo

- Supongamos que queremos predecir la tasa de mortalidad de personas mayores (Y) usando actividad física (A) controlando por otras covariables como edad (W_1), género (W_2) y presencia de enfermedades crónicas (W_3).
- Supongamos que hay tres expertos en la materia y que cada uno propone tres modelos distintos de regresión logística para este análisis:

$$\begin{aligned}
\bar{Q}_n^a(A, W) &= P_n^a(Y = \text{Sí}|A, W) \\
&= \text{expit}(\alpha_{0,n} + \alpha_{1,n}A + \alpha_{2,n}W_1 + \alpha_{3,n}W_2 + \alpha_{4,n}W_3), \\
\bar{Q}_n^b(A, W) &= \text{expit}(\alpha_{0,n} + \alpha_{1,n}A + \alpha_{2,n}W_1 + \alpha_{3,n}W_2 + \alpha_{4,n}W_3 \\
&\quad + \alpha_{5,n}(W_1 \times W_2)) \text{ y} \\
\bar{Q}_n^c(A, W) &= \text{expit}(\alpha_{0,n} + \alpha_{1,n}A + \alpha_{2,n}W_1 + \alpha_{3,n}W_2 + \alpha_{4,n}W_3 \\
&\quad + \alpha_{5,n}(W_1^2)).
\end{aligned}$$

SA: VC y Combinación Lineal Óptima

- Se ajustan todos los modelos usando validación cruzada con, por ejemplo, 90% de la muestra para el ajuste y un 10% de la muestra para evaluar cada ajuste. Se realizan 10 iteraciones.

SA: VC y Combinación Lineal Óptima

- Se ajustan todos los modelos usando validación cruzada con, por ejemplo, 90% de la muestra para el ajuste y un 10% de la muestra para evaluar cada ajuste. Se realizan 10 iteraciones.
- Se calcula el riesgo de validación cruzada para cada modelo j (ej. $j = \{1, 2, 3\}$) en toda la muestra usando, por ejemplo:

$$\text{VC ECM}_j = \frac{\sum_{i=1}^{n}(Y_i - D_{j,i})^2}{n},$$

donde $D_{j,i}$ es la probabilidad predictiva de cada modelo para cada observación.

SA: VC y Combinación Lineal Óptima

- Se ajustan todos los modelos usando validación cruzada con, por ejemplo, 90% de la muestra para el ajuste y un 10% de la muestra para evaluar cada ajuste. Se realizan 10 iteraciones.
- Se calcula el riesgo de validación cruzada para cada modelo j (ej. $j = \{1, 2, 3\}$) en toda la muestra usando, por ejemplo:

$$\text{VC ECM}_j = \frac{\sum_{i=1}^{n}(Y_i - D_{j,i})^2}{n},$$

  donde $D_{j,i}$ es la probabilidad predictiva de cada modelo para cada observación.
- SA mejora calculando la combinación pesada óptima de todos los métodos utilizados indexados por un vector de pesos p con coordenadas ≥ 0 que sumadas dan 1.

SA: VC y Combinación Lineal Óptima

- Se ajustan todos los modelos usando validación cruzada con, por ejemplo, 90% de la muestra para el ajuste y un 10% de la muestra para evaluar cada ajuste. Se realizan 10 iteraciones.
- Se calcula el riesgo de validación cruzada para cada modelo j (ej. $j = \{1, 2, 3\}$) en toda la muestra usando, por ejemplo:

$$\text{VC ECM}_j = \frac{\sum_{i=1}^{n}(Y_i - D_{j,i})^2}{n},$$

  donde $D_{j,i}$ es la probabilidad predictiva de cada modelo para cada observación.
- SA mejora calculando la combinación pesada óptima de todos los métodos utilizados indexados por un vector de pesos p con coordenadas ≥ 0 que sumadas dan 1.
- En el ejemplo se genera la siguiente función de predicción:

$$\bar{Q}_n^0 = 0.461\bar{Q}_{modelo_1,n} + 0.539\bar{Q}_{modelo_2,n} + 0.000\bar{Q}_{modelo_3,n}.$$

Métodos Incluidos en el Ejemplo Real de SA

Método	Función en SuperLearner
Modelo Lineal	glm
Modelo Lineal Bayesiano	bayesglm
Regresión Polinomial con Splines	polymars
Random Forest	randomForest
Red Elástica	glmnet, $\alpha = 0.25$
	glmnet, $\alpha = 0.50$
	glmnet, $\alpha = 0.75$
	glmnet, $\alpha = 1.00$
Modelo Aditivo Generalizado	gam, degree=2
	gam, degree=3
	gam, degree=4
	gam, degree=5
Red Neuronal	nnet, size=2
	nnet, size=4

SA: Propiedades

- Incorporando una colección rica de métodos y modelos con distintos niveles de sesgo y ajuste, la validación cruzada previene el sobreajuste y también previene la selección de un modelo muy sesgado.

SA: Propiedades

- Incorporando una colección rica de métodos y modelos con distintos niveles de sesgo y ajuste, la validación cruzada previene el sobreajuste y también previene la selección de un modelo muy sesgado.
- La colección de métodos incluidos puede ser grande e incluir todo tipo de modelos paramétricos, semi-paramétricos y no paramétricos.

SA: Propiedades

- Incorporando una colección rica de métodos y modelos con distintos niveles de sesgo y ajuste, la validación cruzada previene el sobreajuste y también previene la selección de un modelo muy sesgado.
- La colección de métodos incluidos puede ser grande e incluir todo tipo de modelos paramétricos, semi-paramétricos y no paramétricos.
- Se puede demostrar que SA se desempeña tan bien como el método que minimiza la esperanza de la función de pérdida.

SA: Propiedades

- Incorporando una colección rica de métodos y modelos con distintos niveles de sesgo y ajuste, la validación cruzada previene el sobreajuste y también previene la selección de un modelo muy sesgado.
- La colección de métodos incluidos puede ser grande e incluir todo tipo de modelos paramétricos, semi-paramétricos y no paramétricos.
- Se puede demostrar que SA se desempeña tan bien como el método que minimiza la esperanza de la función de pérdida.
- Incluso cuando la colección de métodos contiene un modelo paramétrico correctamente especificado, los resultados de SA y del modelo verdadero son similares. En este caso la variabilidad de SA es mayor que la del modelo paramétrico.

SA: Problema

- Para cada sujeto en la muestra, uno podría evaluar la diferencia entre $\bar{Q}_n^0(A = 1, W_i)$ y $\bar{Q}_n^0(A = 0, W_i)$ y calcular el promedio de estas diferencias.

SA: Problema

- Para cada sujeto en la muestra, uno podría evaluar la diferencia entre $\bar{Q}_n^0(A = 1, W_i)$ y $\bar{Q}_n^0(A = 0, W_i)$ y calcular el promedio de estas diferencias.
- **Problema:** Este estimador no está dirigido al parámetro de interés y es sesgado.

SA: Problema

- Para cada sujeto en la muestra, uno podría evaluar la diferencia entre $\bar{Q}_n^0(A = 1, W_i)$ y $\bar{Q}_n^0(A = 0, W_i)$ y calcular el promedio de estas diferencias.
- **Problema:** Este estimador no está dirigido al parámetro de interés y es sesgado.
- **Solución:** Máxima verosimilitud dirigida (MVD).

SA: Problema

- Para cada sujeto en la muestra, uno podría evaluar la diferencia entre $\bar{Q}_n^0(A = 1, W_i)$ y $\bar{Q}_n^0(A = 0, W_i)$ y calcular el promedio de estas diferencias.
- **Problema:** Este estimador no está dirigido al parámetro de interés y es sesgado.
- **Solución:** Máxima verosimilitud dirigida (MVD).
- Usando estimación basada en una función de máxima verosimilitud dirigida (targeted maximum likelihood estimation o TMLE) se puede reducir este sesgo.

SA: Problema

- Para cada sujeto en la muestra, uno podría evaluar la diferencia entre $\bar{Q}_n^0(A = 1, W_i)$ y $\bar{Q}_n^0(A = 0, W_i)$ y calcular el promedio de estas diferencias.
- **Problema:** Este estimador no está dirigido al parámetro de interés y es sesgado.
- **Solución:** Máxima verosimilitud dirigida (MVD).
- Usando estimación basada en una función de máxima verosimilitud dirigida (targeted maximum likelihood estimation o TMLE) se puede reducir este sesgo.
- MVD es la segunda etapa del algoritmo de AD.

SA: Problema

- Para cada sujeto en la muestra, uno podría evaluar la diferencia entre $\bar{Q}_n^0(A = 1, W_i)$ y $\bar{Q}_n^0(A = 0, W_i)$ y calcular el promedio de estas diferencias.
- **Problema:** Este estimador no está dirigido al parámetro de interés y es sesgado.
- **Solución:** Máxima verosimilitud dirigida (MVD).
- Usando estimación basada en una función de máxima verosimilitud dirigida (targeted maximum likelihood estimation o TMLE) se puede reducir este sesgo.
- MVD es la segunda etapa del algoritmo de AD.
- MVD actualiza el estimador obtenido por SA de manera tal de obtener un equilibrio óptimo entre sesgo y varianza para la estimación del parámetro objetivo, $\Psi(Q_0)$.

Máxima Verosimilitud Dirigida

- Se estima $g_0 = P_0(A|W)$ con g_n usando SA y se agrega a la matriz de datos.

Máxima Verosimilitud Dirigida

- Se estima $g_0 = P_0(A|W)$ con g_n usando SA y se agrega a la matriz de datos.
- Se actualiza $\bar{Q}_n^0$ a $\bar{Q}_n^1$ de la siguiente manera:

$$\text{logit}\bar{Q}_n^1(A, W) = \text{logit}\bar{Q}_n^0(A, W) + \epsilon_n H_n^*(A, W).$$

Máxima Verosimilitud Dirigida

- Se estima $g_0 = P_0(A|W)$ con g_n usando SA y se agrega a la matriz de datos.
- Se actualiza $\bar{Q}_n^0$ a $\bar{Q}_n^1$ de la siguiente manera:

$$\text{logit}\bar{Q}_n^1(A, W) = \text{logit}\bar{Q}_n^0(A, W) + \epsilon_n H_n^*(A, W).$$

- Este modelo paramétrico de trabajo incorpora información de $g_0 = P_0(A|W)$ a través de $H_n^*(A, W)$.

Máxima Verosimilitud Dirigida

- Se estima $g_0 = P_0(A|W)$ con g_n usando SA y se agrega a la matriz de datos.
- Se actualiza $\bar{Q}_n^0$ a $\bar{Q}_n^1$ de la siguiente manera:

$$\text{logit}\bar{Q}_n^1(A, W) = \text{logit}\bar{Q}_n^0(A, W) + \epsilon_n H_n^*(A, W).$$

- Este modelo paramétrico de trabajo incorpora información de $g_0 = P_0(A|W)$ a través de $H_n^*(A, W)$.
- Este paso de actualización se repite hasta que $\epsilon_n = 0$.

Máxima Verosimilitud Dirigida

- Se estima $g_0 = P_0(A|W)$ con g_n usando SA y se agrega a la matriz de datos.
- Se actualiza $\bar{Q}_n^0$ a $\bar{Q}_n^1$ de la siguiente manera:

$$\text{logit}\bar{Q}_n^1(A, W) = \text{logit}\bar{Q}_n^0(A, W) + \epsilon_n H_n^*(A, W).$$

- Este modelo paramétrico de trabajo incorpora información de $g_0 = P_0(A|W)$ a través de $H_n^*(A, W)$.
- Este paso de actualización se repite hasta que $\epsilon_n = 0$.
- En el ejemplo de mortalidad, $\epsilon_n = 0$ luego de la primera iteración.

Máxima Verosimilitud Dirigida

- Se estima $g_0 = P_0(A|W)$ con g_n usando SA y se agrega a la matriz de datos.
- Se actualiza $\bar{Q}_n^0$ a $\bar{Q}_n^1$ de la siguiente manera:

$$\text{logit}\bar{Q}_n^1(A, W) = \text{logit}\bar{Q}_n^0(A, W) + \epsilon_n H_n^*(A, W).$$

- Este modelo paramétrico de trabajo incorpora información de $g_0 = P_0(A|W)$ a través de $H_n^*(A, W)$.
- Este paso de actualización se repite hasta que $\epsilon_n = 0$.
- En el ejemplo de mortalidad, $\epsilon_n = 0$ luego de la primera iteración.
- El estimador de MVD del parámetro objetivo es:

$$\psi_{MVD,n} = \Psi(Q_n^*) = \frac{1}{n}\sum_{i=1}^{n}\{\bar{Q}_n^1(1, W_i) - \bar{Q}_n^1(0, W_i)\}.$$

Máxima Verosimilitud Dirigida

- Se estima $g_0 = P_0(A|W)$ con g_n usando SA y se agrega a la matriz de datos.
- Se actualiza $\bar{Q}_n^0$ a $\bar{Q}_n^1$ de la siguiente manera:

$$\text{logit}\bar{Q}_n^1(A, W) = \text{logit}\bar{Q}_n^0(A, W) + \epsilon_n H_n^*(A, W).$$

- Este modelo paramétrico de trabajo incorpora información de $g_0 = P_0(A|W)$ a través de $H_n^*(A, W)$.
- Este paso de actualización se repite hasta que $\epsilon_n = 0$.
- En el ejemplo de mortalidad, $\epsilon_n = 0$ luego de la primera iteración.
- El estimador de MVD del parámetro objetivo es:

$$\psi_{MVD,n} = \Psi(Q_n^*) = \frac{1}{n}\sum_{i=1}^{n}\{\bar{Q}_n^1(1, W_i) - \bar{Q}_n^1(0, W_i)\}.$$

- En el ejemplo de mortalidad, $\psi_{MVD,n} = -0,055$.

Inferencia: Curva de Influencia

Para calcular el error estándar se usa la siguiente curva de influencia (CI):

$$\begin{aligned} CI_n(O_i) = & H_n^*(A_i, W_i)(Y - \bar{Q}_n^1(A_i, W_i)) \\ & + \bar{Q}_n^1(1, W_i) - \bar{Q}_n^1(0, W_i) - \psi_{MVD,n}. \end{aligned}$$

Inferencia: Curva de Influencia

Para calcular el error estándar se usa la siguiente curva de influencia (CI):

$$CI_n(O_i) = H_n^*(A_i, W_i)(Y - \bar{Q}_n^1(A_i, W_i)) + \bar{Q}_n^1(1, W_i) - \bar{Q}_n^1(0, W_i) - \psi_{MVD,n}.$$

Media muestral de CI: $\bar{CI}_n = \frac{1}{n}\sum_{i=1}^{n} CI_n(o_i)$.

Varianza muestral de CI: $S^2(CI_n) = \frac{1}{n}\sum_{i=1}^{n}(CI_n(o_i) - \bar{CI}_n)^2$.

Error estándar del estimador: $S_n = \sqrt{\frac{S^2(IC_n)}{n}} = 0,012$.

Inferencia: Intervalo de Confianza y P-Valor

Intervalo de confianza para $\psi_{MDV,n}$:

$$\psi_{MDV,n} \pm z_{0,975} \frac{S_n}{\sqrt{n}},$$

donde z_α es el cuantil α de la distribución normal estándar.

Inferencia: Intervalo de Confianza y P-Valor

Intervalo de confianza para $\psi_{MDV,n}$:

$$\psi_{MDV,n} \pm z_{0,975} \frac{S_n}{\sqrt{n}},$$

donde z_α es el cuantil α de la distribución normal estándar.

P-valor para $\psi_{MDV,n}$:

$$2\left[1 - \Phi\left(\left|\frac{\psi_{MDV,n}}{\frac{S_n}{\sqrt{n}}}\right|\right)\right],$$

donde Φ es la función de distribución acumulada de la distribución normal estándar.

Interpretación

- La interpretación de $\psi_{MVD,n} = -0,055$ es que, luego de controlar por covariables de interés, existe una asociación entre realizar actividad física a un cierto nivel y la mortalidad a 5 años en adultos mayores.
- Realizar cierto nivel de actividad física reduciría un $5,5\%$ el riesgo de mortalidad a 5 años en esta población.
- Esta asociación es estadísticamente significativa ($p < 0,001$, 95% IC $= [-0,078, -0,033]$).
- Esta asociación no implica causalidad.

Aplicaciones

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

Aplicaciones

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

- Sólida base teórica y empírica a través de simulaciones que demuestran la igualdad y, frecuentemente, superioridad del aprendizaje dirigido sobre otras alternativas analíticas en lo que a sesgo y sobreajuste se refiere.

Ventajas

- Sólida base teórica y empírica a través de simulaciones que demuestran la igualdad y, frecuentemente, superioridad del aprendizaje dirigido sobre otras alternativas analíticas en lo que a sesgo y sobreajuste se refiere.
- Alta divulgación y número creciente de publicaciones aplicando la técnica a problemas reales.

Ventajas

- Sólida base teórica y empírica a través de simulaciones que demuestran la igualdad y, frecuentemente, superioridad del aprendizaje dirigido sobre otras alternativas analíticas en lo que a sesgo y sobreajuste se refiere.
- Alta divulgación y número creciente de publicaciones aplicando la técnica a problemas reales.
- Disponible en R y en desarrollo para SAS.

Desventajas

- Desde el punto de vista de la estadística clásica, requiere un cambio importante en la forma en que se piensa y realiza el análisis.

Desventajas

- 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 el súper aprendizaje requiere, como mínimo, un conocimiento básico de una multiplicidad de tipos de análisis que no suelen estar en la caja de herramientas del estadístico aplicado promedio.

Desventajas

- 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 el súper aprendizaje requiere, como mínimo, un conocimiento básico de una multiplicidad de tipos de análisis que no suelen estar en la caja de herramientas del estadístico aplicado promedio.
- AD requiere mayor tiempo computacional que otros métodos analíticos.

Conclusiones y Próximos Pasos

- A pesar de ser novedosa, AD es una herramienta útil y poderosa que combina el aprendizaje automático con la estadística clásica.

Conclusiones y Próximos Pasos

- A pesar de ser novedosa, AD es una herramienta útil y poderosa que combina el aprendizaje automático con la estadística clásica.
- Agregando los supuestos no testeables necesarios, AD puede ser utilizado también para realizar inferencia causal.

Conclusiones y Próximos Pasos

- A pesar de ser novedosa, AD es una herramienta útil y poderosa que combina el aprendizaje automático con la estadística clásica.
- Agregando los supuestos no testeables necesarios, AD puede ser utilizado también para realizar inferencia causal.
- Estamos estudiando AD para implementarla en datos propios.

Conclusiones y Próximos Pasos

- A pesar de ser novedosa, AD es una herramienta útil y poderosa que combina el aprendizaje automático con la estadística clásica.
- Agregando los supuestos no testeables necesarios, AD puede ser utilizado también para realizar inferencia causal.
- Estamos estudiando AD para implementarla en datos propios.
- Esta experiencia en primera persona nos permitirá entender más en profundidad la técnica, sus ventajas y desventajas.

Conclusiones y Próximos Pasos

- A pesar de ser novedosa, AD es una herramienta útil y poderosa que combina el aprendizaje automático con la estadística clásica.
- Agregando los supuestos no testeables necesarios, AD puede ser utilizado también para realizar inferencia causal.
- Estamos estudiando AD para implementarla en datos propios.
- Esta experiencia en primera persona nos permitirá entender más en profundidad la técnica, sus ventajas y desventajas.
- No se pierda el próximo capítulo de "Súper Aprendizaje y sus Amigos" a la misma bati-hora y por este mismo bati-canal.

Referencias

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

2. Rose S. Mortality risk score prediction in an elderly population using machine learning. American Journal of Epidemiology 177, 443-452.

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

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

¡Muchas Gracias!