BOURBAKI

COLEGIO DE MATEMÁTICAS

Causalidad en Machine Learning II

Pruebas Controladas Aleatorizadas & Variables contrafactuales

Índice

01 Introducción

La causalidad es un análogo asimétrico del concepto de correlación, esta última es una de las relaciones más importantes en Machine Learning. En este curso, buscamos familiarizar al Científico de Datos con los fundamentos y las ideas detrás del estudio matemático de la casualidad. Por medio de casos de uso reales practicaremos el uso de modelos causales para el desarrollo de modelos más confiables.

El contenido del curso se divide en seis módulos:

I. Interpretabilidad, Inferencia Bayesiana y el operador Do
II. Ensayos controlados aleatorizados y variables contrafactuales
III. Modelos Causales, D separación y los Axiomas de la Causalidad
IV. El cálculo Do y la solución a la paradoja de Simpson: el criterio Back Door
V. Propensity Score: ventajas y controversias en el Do-Calculus
VI. Ortogonalización y Doble Machine Learning

🔗 El repositorio de esta semana está disponible en este Link.

02 Repaso de probabilidad

En este corto capítulo vamos a repasar algunas nociones generales sobre la probabilidad condicional y la independencia estadística, a los lectores que no estén familiarizados con este contenido se les invita a repasar con detalle pues a lo largo del curso se utilizarán estos conceptos recurrentemente.

- **Probabilidad condicional en una base de datos.** Supongamos que tenemos una base de datos $S = \{(x_i, y_i)\}_i$. Para estudiar la probabilidad condicional $\mathbb{P}(Y = y|X = x)$, lo más conveniente es observar todos los registros que tienen $X = x$ y luego, contar todos aquellos registros con $X = x$ en donde $Y = y$, para hacer el cociente con los registros que contienen $X = x$.

- La ecuación $\mathbb{P}(Y = y|X = x) = \mathbb{P}(Y = y)$ denota **independencia** entre las variables.

  Supongamos que y es un cliente nos abandona. Cuando condicionamos a X y ocurre que $\mathbb{P}(Y = y|X = x) = \mathbb{P}(Y = y)$, observamos que algo muy particular pasa con la variable X para que no haya ninguna relación con la otra variable. Dicho de otra manera, X es muy rara.

- Para obtener $\mathbb{P}(X = x, Y = y)$, debemos contar todos los registros en los que ocurre (x, y), y luego dividirla entre el total de registros de la base de datos.

- Otra manera de denotar independencia es cuando se cumple que

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y)$$

Notation 1. *Utilizaremos la notación $X \perp_Z Y$ para denotar que X es independiente de Y condicionado a Z.*

$$X \perp_Z Y \Leftrightarrow P(Y|Z) = P(Y|X, Z) \tag{02.1}$$

La igualdad de la derecha se puede ilustrar con la igualdad de las siguientes bases de datos.

Z	X	Y
z		y

Z	X	Y
z	x	y

Supongamos que tenemos un modelo lineal $\beta_X x + \beta_Z Z + \beta_0 = Y$ y además sabemos que $X \perp_Z Y$, entonces podríamos reemplazar el modelo anterior por $\beta_Z Z + \beta_0 = Y$.

03 Ensayos controlados aleatorizados

En este segundo módulo del curso vamos a presentar dos posibles acercamientos que nos permitirán estudiar la relación causal entre dos variables: las pruebas A/B y los modelos generativos controlados. Antes de introducir estos conceptos y las hipótesis necesarias será muy importante introducir las notaciones adecuadas para las variables contrafactuales que se utilizarán a lo largo del curso.

El estudio matemático de la causalidad proporciona un modelo para afirmaciones del tipo X es causa de Y. Eventualmente estudiaremos los axiomas de la causlidad y los diagramas DAG, es decir, la causalidad á la Pearl. Pero por ahora trabajaremos en un modelo llamado **Potencial Outcomes**.

Variables contrafactuales

Supongamos que tenemos dos variables binarias X y Y. Para mayor comodidad $\Omega_X = \{0,2\}$ y $\Omega_Y = \{-1,+1\}$.

Denotamos $Y^*(0)$ y $Y^*(2)$ a las variables aleatorias que corresponden al resultado de la variable Y al cambiar cada registro de X por el valor opuesto. Notemos que al suponer que nuestro espacio de probabilidad $\Omega_X = \{0,2\}$ es

la realización de una base de datos real, cada uno de estos casos hipotéticos tiene algún resultado en $\{-1,+1\}$. Evidentemente no podemos suponer que lo conocemos por el momento.

X	Y	$Y^*(0)$	$Y^*(2)$
0	-1	-1	?
	+1	+1	
2	-1	?	-1
	+1		+1

Observemos que en la tabla, es obvio que $Y^*(0)$ coincide exactamente con Y.

La variable Y es la suma de dos variables que no podemos ver, pero de las cuales depende, la siguiente fórmula nos da alguna esperanza para despejar la ecuación anterior y recuperar información contrafactual de información observada.

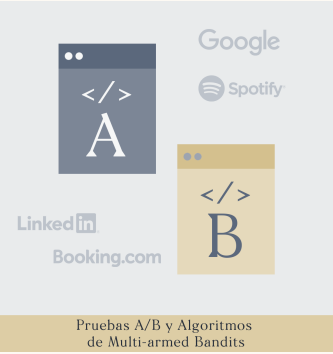
$$Y = \frac{X}{2} Y^*(2) + \left(1 - \frac{X}{2}\right) \cdot Y^*(0) \tag{03.1}$$

Example 01.1. *Supongamos que tenemos la base de datos $S = \{(x_i, y_i)\}_i$ de clientes de un servicio de streaming.*

- $X = 0$ *clientes a los que no le dimos una promoción.*
- $X = 2$ *son clientes a los que sí les dimos una promoción.*
- $Y = -1$ *Clientes que abandonaron el servicio*
- $Y = +1$ *Clientes que continúan teniendo el servicio*
- $Y^*(2)$ *Corresponde al resultado de intervenir $X = 0$, respondiendo la pregunta ¿Qué pasaría con el churn rate si a los clientes a los que no les*

habíamos ofrecido una promoción, de pronto la ponemos a su alcance mientras que a los clientes que ya les habíamos ofrecido la promoción también se las mantenemos?

A/B testing



En el mismo caso del problema anterior supongamos que hemos observado la siguiente cantidad no causal:

$$\mathbb{P}(Y=1|X=2) > \mathbb{P}(Y=1|X=0) \tag{03.2}$$

¿Es enviarles la promoción a los clientes la causa por la que dejan de abandonar el servicio? Nos encantaría contestar positivamente a esa pregunta sin embargo esto solo nos lo podría confirmar las variables contrafactuales.

Vamos a suponer el siguiente caso en el que por arte de magia conocemos a la variable $Y^*(0)$, hemos señalado con otro color a los últimos registros imaginando que son clientes que están extremadamente convencidos de nuestro

producto:

X	Y	$Y^*(0)$
0	-1	-1
0	-1	-1
0	1	1
2	-1	-1
2	1	-1
2	1	1
2	1	1
2	1	1

Inmediatamente notamos que se cumple lo que habíamos supuesto sobre la variable observada:

$$\mathbb{P}(Y=1|X=0) = 1/3 < \mathbb{P}(Y=1)|X=2) = 4/5$$

Otro aspecto a considerar es que por alguna razón que desconocemos se les ha enviado un correo a más personas de las que no se les ha enviado un correo. ¿Podría ser que el equipo de ventas está atrasado en cumplir sus metas mensuales?

Gracias a que tenemos acceso a la variable contrafactual también podemos observar que los tres registros del final están completamente enganchados al producto y no nos dejarán.

Un aspecto fundamental de esta base de datos relacionada con el desbalance en la manera como se enviaron las promociones es la siguiente ecuación:

$$\mathbb{P}\left(Y^*(0)=1|X=0\right) = 1/3 < \mathbb{P}\left(Y^*(0)=1|X=2\right) = 3/5 \qquad (03.3)$$

En un ensayo aleatorizado nos gustaría que ambas cantidades fueran muy parecidas, un poco más adelante hablaremos del porqué. Una manera como podríamos lograr que ambas cantidades sean parecidas es por ejemplo modificando nuestras encuestas y no enviándole al último registro un correo electrónico como lo muestra en la siguiente tabla. Esta nueva encuesta a la que denotaremos por X^* podría perfectamente ser el resultado de una prueba A/B en la que de la manera más aleatoria posible se les enviará o no una promoción a los clientes (sin tomar en cuenta el sesgo ocasionado por la prisa del equipo de ventas). Fijémonos en el último registro al que en esta nueva encuesta ya no le enviamos la promoción.

X^*	Y	$Y^*(0)$
0	-1	-1
0	-1	-1
0	1	1
2	-1	-1
2	1	-1
2	1	1
2	1	1
0	1	1

Notemos que ahora la última desigualdad ha cambiado:

$$\mathbb{P}(Y^*(0) = 1|X = 0) = 2/4 = \mathbb{P}(Y^*(0) = 1|X = 2) = 2/4$$

Ahora la muestra que tenemos en nuestra base de datos se ha hecho de tal manera que tanto el grupo de los que han recibido la promoción como el grupo de los que no la han recibido tienen la misma probabilidad de aceptar en las variables contrafactuales. Es muy importante resaltar que este balance

entre las variables contrafactuales no necesariamente implica que haya balance entre las variables observadas pues si lo hubiera entonces no serviría de nada el experimento, en nuestro caso:

$$\mathbb{P}(Y=1|X=0) = 2/4 < \mathbb{P}(Y=1)|X=2) = 3/4$$

Lo anterior tiene más posibilidades de indicar que efectivamente las promociones están ayudando a que los clientes sigan siéndolo.

La hipótesis de independencia y el teorema fundamental

De acuerdo al razonamiento mostrado en la sección anterior, si fuéramos capaces de elegir de la manera más aleatoria posible a quiénes se les envían los correos, habrán más posibilidades de que se cumpla la siguiente ecuación sobre las variables contrafactuales:

$$\mathbb{P}(Y^*(0)|X=0) = \mathbb{P}(Y^*(0)|X=2), \mathbb{P}(Y^*(2)|X=0) = \mathbb{P}(Y^*(2)|X=2) \quad (03.4)$$

Esta ecuación es una consecuencia del siguiente enunciado conocido en la teoría de la causalidad como la **hipótesis de independencia**, también cono-

cida como **hipótesis de ignorabilidad**. Ésta hipótesis requiere que los resultados potenciales $Y^*(0)$ y $Y^*(2)$ sean independientes del tratamiento recibido, es decir, de X:

$$Y^*(0), Y^*(2) \perp X \tag{03.5}$$

Efectivamente gracias a la independencia podemos deducir las dos igualdades anteriores:

$$\mathbb{P}(Y^*(0)|X=0) = \mathbb{P}(Y^*(0)) = \mathbb{P}(Y^*(0)|X=2)$$

$$\mathbb{P}(Y^*(2)|X=0) = \mathbb{P}(Y^*(2)) = \mathbb{P}(Y^*(2)|X=2)$$

Es decir, los clientes que tendrán un resultado potencial $Y^* = +1$ tienen la misma probabilidad de estar en el grupo $X = 0$ o en $X = 2$, que los clientes que tendrían un resultado potencial opuesto $Y^* = -1$.

Los lectores no deberían de confundir esta hipótesis con la afirmación $Y \perp X$ pues si esto ocurriera entonces estaríamos suponiendo que no existe ni siquiera una relación estadística entre las variables observadas lo cual de ninguna manera podemos suponer.

La consecuencia más importante de la hipótesis de independencia es poder calcular efectos causales únicamente utilizando cantidades observables como lo vamos a mostrar a continuación. Comencemos calculando la siguiente cantidad estadística causal que mide en promedio la diferencia entre las dos

variables contrafactuales:

$$\mathbb{E}\left[Y^*(2) - Y^*(0)\right] = \mathbb{E}\left[Y^*(2)\right] - \mathbb{E}\left[-Y^*(0)\right] = \mathbb{E}\left[Y^*(2)|X=2\right] - \mathbb{E}\left[Y^*(0)|X=0\right]$$

La primera igualdad es gracias a la aditividad del valor esperado mientras que la segunda es gracias a la hipótesis de independencia. Además de lo anterior recordemos que las variables contrafactuales que intervienen con el valor 0 o 2 son idénticas a la variable observada Y en el subconjunto de los registros donde $X = 0$ y $X = 2$ respectivamente (fuera de esos conjuntos no las conocemos). Gracias a este argumento y a la última ecuación podemos deducir que:

$$\mathbb{E}\left[Y^*(2) - Y^*(0)\right] = \mathbb{E}\left[Y|X=2\right] - \mathbb{E}\left[Y|X=0\right]$$

Dicho de otra manera:

Theorem 03.1. *Si las dos variables contrafactuales satisfacen la hipótesis de independencia con respecto a la variable del tratamiento X entonces la siguiente estadística causal puede calcularse utilizando únicamente datos observados en nuestra base de datos (X, Y). A esta cantidad le llamaremos el Efecto del Tratamiento Promedio (ATE)*

$$\mathbb{E}\left[Y^*(2) - Y^*(0)\right]$$

Las hipótesis para el caso multi-variado

En el resto del curso vamos a continuar estudiando el caso anterior cuando estamos en presencia de más variables explicativas y no solo la variable X que utilizamos en esta sección. La notación que utilizaremos en el resto del curso para las otras variables será el vector Z, en el siguiente capítulo daremos un ejemplo de este caso.

La hipótesis de independencia en este caso multivariado se escribie de la siguiente manera:

- $Y^* \perp_Z X$

Es decir las variables contrafactuales siguen siendo independientes del tratamiento a pesar de que condicionemos a las distintas sub-poblaciones determinadas por las variables Z. Nuevamente es posible demostrar un teorema parecido al de la sección anterior solo que en este caso requerirá un poco más de trabajo.

Otra hipótesis importante que agregaremos en este caso y se estudiará en el caso de uso de esta semana es la siguiente:

- $0 < \mathbb{P}(X|Z) < 1$

A esta hipótesis se le conoce como la hipótesis de positividad y esencialmente está diciendo que a pesar de condicionar con las variables de Z, la proporción de los registros en X sigue siendo no cero, es decir que siguen existiendo

registros a los que se les aplica y no el tratamiento.

04 Primeros modelos generativos

En el capítulo anterior propusimos un método conocido comúnmente como A/B testing para garantizar que las estadísticas causales se podrían deducir utilizando únicamente los datos observados. Es muy importante insistir en que para lograr obtener una muestra con un A/B testing es necesario rehacer el muestreo (X, Y) por lo cual estamos hablando más bien del diseño de nuestro experimento.

En esta sección hablaremos de algunos ejemplos de modelos matemáticos que incluyen funciones deterministas y distribuciones los cuales también nos permitirán deducir cantidades causales utilizando únicamente este modelo. Estos modelos serán la base de los modelos causales que presentaremos en el siguiente módulo del curso.

En este capítulo nuevamente utilizaremos la misma notación para las variables contrafactuales.

Modelo I

Supongamos que tenemos dos variables aleatorias de Bernoulli e independientes $X \perp Y$. A continuación mostramos una simulación casi perfecta de las

hipótesis anteriores.

X	Y
0	-1
0	1
2	-1
2	1

Con las hipótesis del problema de las promociones y los clientes churn del capítulo anterior nos podríamos hacer la siguiente pregunta contrafactual:

- Supongamos que un cliente ha recibido la promoción pero a pesar de ello ha decidido abandonar su suscripción. ¿Qué hubiera pasado si ese cliente no hubiera recibido la promoción? En particular: £si ese cliente no hubiera recibido la promoción nos habría abandonado?

  La siguiente cantidad nos podría ayudar a contestar esa pregunta:

$$\mathbb{P}(Y^*(0) = +1 | X = 2, Y = -1)$$

Utilizando el modelo generativo anterior es bastante sencillo responder a la pregunta anterior siguiendo el siguiente razonamiento:

1. Los registros en los que $X = 2, Y = -1$ son una cuarta parte de la base de datos, están en el tercer renglón de nuestra base de datos.

2. Si modificamos la variable X para que ahora valga 0, ya que las variables son independientes, nada cambiaría en la variable Y lo cual significa que la probabilidad que queremos calcular es igual a 0.

Modelo II

Para este segundo modelo vamos a suponer que tenemos dos variables aleatorias de Bernoulli e independientes $X \perp Z$. En el proceso generativo también supondremos que existe una tercera variable Y que cumple la siguiente relación funcional donde la función "sign" envía a los números positivos al $+1$ y a los números negativos al -1:

$$Y = sign(XZ + (1-X)(1-Z))$$

Al simular este modelo generativo obtenemos la siguiente base de datos:

X	Z	Y
0	0	1
0	2	-1
2	0	-1
2	2	1
0	0	1
0	2	-1
2	0	-1
2	2	1

Es interesante notar que en ambos modelos las variables observadas (X, Y) son idénticas lo cual significa que esto es una tarea complicada para el científico de datos.

Nuevamente nos gustaría responder la siguiente pregunta contrafactual (la misma que en el modelo anterior).

- Supongamos que un cliente ha recibido la promoción pero a pesar de

ello ha decidido abandonar su suscripción. £Qué hubiera pasado si ese cliente no hubiera recibido la promoción? En particular: £si ese cliente no hubiera recibido la promoción nos habría abandonado?

La siguiente cantidad nos podría ayudar a contestar esa pregunta:

$$\mathbb{P}(Y^*(0) = +1 | X = 2, Y = -1)$$

Utilizando el modelo generativo anterior también podemos calcular esta cantidad siguiendo el siguiente razonamiento:

1. Los registros en los que $X = 2, Y = -1$ son una cuarta parte de la base de datos, están en el tercer renglón y en el séptimo de la base de datos.

2. La observación importante es que de acuerdo a este modelo, aquellos registros que abandonaron y sí recibieron la promoción son exactamente aquellos para los que la variable $Z = 0$.

3. Ahora supongamos que intervenimos la variable X cambiándola por un 0, gracias la observación anterior la muestra que ahora nos interesa son aquellos registros donde $X = 0, Z = 0$ y en este caso la variable Y siempre será igual a $+1$. Si lo desean verificar son el primer y quinto registro en la base de datos.

4. Por lo anterior, en este segundo modelo podemos deducir que $\mathbb{P}(Y^*(0) = +1 | X = 2, Y = -1) = 1$, precisamente lo contrario que en el Modelo I.

05 Lectura de referencia: Comparación con A/B testing

La lectura de referencia de esta semana es un trabajo de un estudiante de Judea Pearl en el que propone un método alternativo a los famosos A/B tests que hemos mencionado hasta ahora de tal manera que maximice aquellos registros que renuevan gracias a que les hemos enviado una publicidad y que al mismo tiempo minimice las siguientes cantidades:

- Quienes abandonarán aunque les enviemos la publicidad.
- Quienes abandonarán aunque no les enviemos la publicidad.
- Quienes no abandonarán aunque no les enviemos la publicidad.

La intención final es elegir un sector de la población $Z = c$ para la que se cumplen los requisitos anteriores. Es importante mencionar que la notación es ligeramente distinta en el artículo pues y_x denota lo que nosotros llamamos $Y^*(0) = -1$ y ${y'}_{x'}$ a $Y^*(2) = +1$.

El artículo *Unit Selection with Nonbinary Treatment and Effect,* se puede encontrar aquí.

colegio-bourbaki.com
+52 56 2141 7850
info@colegio-bourbaki.com