BOURBAKI

COLEGIO DE MATEMÁTICAS

Causalidad en Machine Learning VI

Ortogonalización y Double Machine Learning

BOURBAKI

COLEGIO DE MATEMÁTICAS

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 aleatorizados: tests A/B, Simulación y RCT
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

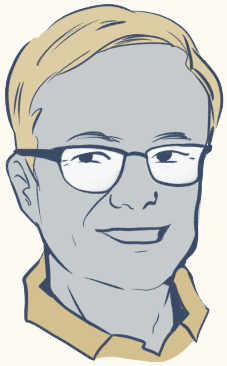
🔗 El repositorio de esta semana está disponible en este link.

02 El Teorema de Frisch-Waugh-Lovell: Causalidad y doble ML

Una de las grandes ventajas que tiene Machine Learning sobre otros métodos para resolver problemas complejos es su capacidad para incluir una enorme cantidad de variables explicativas. Todos los científicos de datos saben que una base de datos con pocas columnas difícilmente es un objeto amable para los métodos tradicionales.

La gran desventaja de este enfoque es la ausencia de explicabilidad y causalidad de estos modelos. Si abordamos un problema de previsión de un desastre natural con Machine Learning, no es suficiente con obtener resultados magníficos en un backtesting o inclusive en las primeras pruebas reales, si el modelo no está prediciendo por las razones correctas entonces no es un modelo confiable.

Por lo anterior es indispensable que se encuentren mejores métodos para detectar relaciones de causalidad en los modelos de machine learning. En este capítulo hablaremos sobre la piedra angular detrás de **Double Machiine Learning**. Una técnica extraordinariamente útil para investigar la causalidad en una base de datos. Estas ideas han sido promovidas por el exitoso economista Victor Chernozhukov así como sus colaboradores.



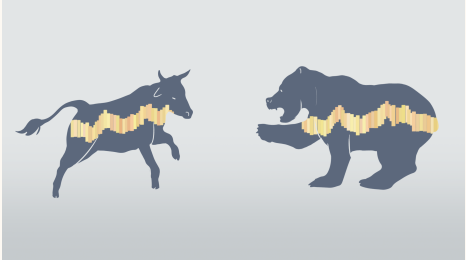
Predecir lluvias usando USD-MXN

Supongamos que estamos intentando predecir el número de litros por metro cuadrado que caerán en distintas zonas en la República Mexicana. Si abordamos este problema como científicos datos pronto notamos que es un problema de regresión, es decir nuestra variable objetivo es numérica. Inmediatamente reconocemos que debemos agregar variables que expliquen el lugar donde estamos prediciendo, por ejemplo si estamos en la playa, en una montaña, etc. A todas estas variables les llamaremos nuestras X, estas podrían ser muchísimas pero supongamos por ahora que todas son numéricas o binarias para poder utilizar un modelo lineal de regresión.

Posiblemente la descripción geográfica del lugar no sea suficiente. Para todos sería razonable incluir variables que describen el momento del año en el que nos gustaría predecir la cantidad de lluvia.

Vamos a imaginar que hemos enloquecido y las variables que añadimos están relacionadas con la economía nacional, por ejemplo las últimas tasas de interés, el tipo de cambio entre pesos mexicanos y dólares americanos en los últimos días, el valor del petróleo también en los últimos días y un gran etc.

A todas estas variables las llamaremos $.



El Teorema de Frisch-Waugh-Lovell

Un buen benchmark para comenzar esta locura podría ser utilizar un modelo de regresión lineal. Como siempre, debemos de comenzar probando con modelos sencillos. Escrito en una ecuación, nos gustaría resolver lo siguiente:

$$Y = AX + B\$ + C \tag{02.1}$$

Con resolver queremos decir que nos gustaría encontrar cuáles son los valores A, B y C que mejor aproximen los litros por metro cuadrado en una zona particular Como estamos suponiendo que tanto X como $ son más de una sola variable, los A,B y C serían vectores. Vamos a suponer por el momento que $ es una sola variable, digamos el cambio peso/dólar. Supongamos que ya lo hemos hecho. El valor de B refleja cómo se utiliza MXN/USD para predecir los litros por metro cuadrado. ąQué horrible se escucha!

El Teorema de Frisch Waugh Lovell afirma que también podríamos calcular a

B siguiendo estos pasos:

1. Entrenamos una regresión lineal para predecir los litros por metro cuadrado únicamente utilizando la información geográfica, es decir las X. Una vez que lo hayamos hecho podríamos calcular el error en nuestra base de datos, es decir cuánto se subestimaron o sobreestimaron los metros cúbicos de lluvia en cada lugar. A esta variable le llamaremos e.

2. Ahora vamos a entrenar una regresión lineal para predecir a $ utilizando información geográfica. Este paso es una locura, pues estamos intentando utilizar características geográficas de algún lugar para predecir cuánto cuesta un dólar americano en cierto instante. Al error de estas predicciones le llamaremos E para reflejar nuestro sentir con este paso.

3. Por último, vamos a entrenar una regresión para predecir la columna e utilizando la columna E. Esta es una regresión univariada que tendrá un único peso asociado a E al que llamaremos B'.

La primera afirmación que hace el Teorema de FrischWaughLovell es que B y B' son idénticos.

Formalización del Teorema FWL

En esta sección desarrollaremos teóricamente el Teorema de Firsch - Waugh - Lovell.

Consideremos un problema con variables como las que hemos utilizado hasta en el curso: la variable del tratamiento X, la variable objetivo Y y covariables Z. Consideremos también un modelo lineal $f(X,Z) = \hat{Y}$ con coeficientes como a continuación.

$$\hat{Y} = \beta_X \cdot X + \beta_Z \cdot Z. \tag{02.2}$$

Recordemos que las técnicas comunes de interpretabilidad pueden confundirnos en un estudio causal. El Teorema de Firsh-Waugh-Lovell proporciona una manera alternativa para obtener el coeficiente β_X.

Consideremos las siguientes regresiones parciales, junto con sus respectivos errores:

(I) $\hat{Y}_Z = \beta_Z^Y \cdot Z$, con error $\epsilon_Y = \hat{Y}_Z - Y$.

(II) $\hat{X}_Z = \beta_Z^X \cdot Z$, con error $\epsilon_X = \hat{X}_Z - X$.

Example 02.1. *Imaginemos un problema de valuación en donde X son los metros cuadrados de un inmueble y Z es el año de construcción. Quisiéramos saber qué tanto influyen los metros cuadrados en el precio de la propiedad, Y. En este ejemplo, (I) es la predicción del precio solo considerando el año de*

construcción de la propiedad. El error de (I) es el error que se comete en el precio cuando no se consideran los años de antigüedad.

(II) corresponde a la predicción de los metros cuadrados dado su año de construcción.

Tomemos las variables aleatorias dada por los errores.

ϵ_X	ϵ_Y
$\vdots$	$\vdots$
$\epsilon_{X,40}$	$\epsilon_{X,40}$
$\vdots$	$\vdots$

Calculemos una regresión con ellas.

$$\hat{\epsilon}_Y = \beta'_X \cdot \epsilon_X \tag{02.3}$$

Notemos que solamente hemos realizado regresiones univariadas para calcular todos los ϵ hasta ahora. Esto por supuesto, es muy valioso computacionalmente incluso si la cantidad de covariables es grande.

El **teorema de Fircsh - Waugh - Lovell** asegura que el coeficioente de la regresión 02.2, es exactamente el mismo coeficiente de la regresión de los errores 02.3:

$$\beta'_X = \beta_X \tag{02.4}$$

03 Double Machine Learning

Consideremos de nnuevo un problema de regresión lineal con una variable objetivo Y, una variable de tratamiento X y esta vez con d covariables $Z_1, Z_2, \ldots, Z_d$ y un modelo lineal clásico

$$\hat{Y} = \beta_X \cdot X + \beta_1 \cdot Z_1 + \cdots + \beta_d \cdot Z_d$$

En el capítulo anterior requerimos tres regresiones:

(I) La predicción de Y solamente considerando las covariables $\hat{Y}_Z = \beta^Y_{Z_1} \cdot Z_1 + \beta^Y_{Z\,1} \cdot Z_1 + \cdots + \beta^Y_{Z\,d} \cdot Z_d$, junto con su error $\epsilon_Y = \hat{Y}_Z - Y$.

(II) La predicción del tratamiento respecto de las covariables $\hat{X}_Z = \beta^X_{Z_1} \cdot Z_1 + \cdot\beta^X_{Z_d} \cdot Z_d$ con su respectivo error $\epsilon_X = \hat{X}_Z - X$.

(III) La regresión del error de la variable explicativa respecto al error del tratamiento $\hat{\epsilon}_Y = \beta'_X \cdot \epsilon_X$.

El término **ortogonal** se refiere a que el error de una regresión, ϵ, es perpendicular al espacio de las variables explicativas. En el contexto de causalidad, se refiere a ser independientes.

El error ϵ_Y ya no se puede explicar mediante las covariables Z. En ese sentido, podemos decir que ϵ_Y es la parte de Y que ya no tiene nada que ver con Z.

Example 00.1. *Intuitivamente, si no conocemos ninguna otra característica,*

podemos pensar que entre más metros cuadrados tiene una casa, es más costosa que una casa con menos área de construcción. Toda la información que ***no*** *está relacionada con los metros cuadrados, se encuentra codificada por el error* ϵ_Y*, por lo tanto no está relacionada con las variables explicativas* Z*, es decir:* $\epsilon_Y \perp Z$.

Lo mismo ocurre con ϵ_X. No podemos asegurar que X es independiente de las variables Z, pero sí sabemos que ϵ_X es la parte de X que es independiente de Z.

Para observar una relación causal entre X y Y, necesitamos la capacidad de intervenir X. La parte de X que está siendo influida por Z, no es un experimento aleatorizado por Z. Quien sí representa un experimento aleatorio es ϵ_X. Por tanto ϵ_X es sencilla de intervenir.

En resumen:

ϵ_Y es la nueva variable objetivo que no depende de $Z_1, Z_2, \ldots, Z_d$.

ϵ_X es la nueva variable tratamiento que no depende de $Z_1, Z_2, \ldots, Z_d$.

Podemos pensar a ϵ_Y y ϵ_X como un preprocesamiento de Y y X, para lo que haremos a continuación.

Average Treatment Effect

Recordemos que en un estudio contrafactual en el que $Y \in \{-1, +1\}$ y $X \in \{0, 1\}$, la hipótesis de ignorabilidad consiste en la independencia de los resultados potenciales $Y^*(0)$ y $Y^*(1)$ del tratamiento X, es decir:

$$\mathbb{P}\left(Y^*(0)\right) = \mathbb{P}\left(Y^*(0) \mid X = 0\right) = \mathbb{P}\left(Y^*(0) \mid X = 1\right)$$

$$\mathbb{P}\left(Y^*(1)\right) = \mathbb{P}\left(Y^*(1) \mid X = 0\right) = \mathbb{P}\left(Y^*(1) \mid X = 1\right)$$

Esta independencia indica que el tratamiento es verdaderamente un experimento aleatorizado.

El average treatment effect es la diferencia entre uno y otro de los tratamientos: $\mathbb{E}[Y^*(1) - Y^*(0)]$. Por la linealidad del valor esperado tenemos:

$$\begin{aligned} ATE &= \mathbb{E}\left[Y^*(1) - Y^*(0)\right] = \mathbb{E}\left[Y^*(1)\right] - \mathbb{E}\left[Y^*(0)\right] \\ &= \left(\mathbb{E}\left[Y^*(1) \mid X = 1\right] + \mathbb{E}\left[Y^*(1) \mid X = 0\right]\right) - \left(\mathbb{E}\left[Y^*(0) \mid X = 1\right] + \mathbb{E}\left[Y^*(0) \mid X = 0\right]\right) \qquad (03.1) \\ &= \left(\mathbb{E}\left[Y^*(1) \mid X = 1\right] + \mathbb{E}\left[Y^*(1) \mid X = 1\right]\right) - \left(\mathbb{E}\left[Y^*(0) \mid X = 0\right] + \mathbb{E}\left[Y^*(0) \mid X = 0\right]\right) \qquad (03.2) \\ &= 2 \cdot \mathbb{E}[Y \mid X = 1] - 2 \cdot \mathbb{E}[Y \mid X = 0] \qquad (03.3) \end{aligned}$$

En la igualdad (03.1) utilizamos las hipótesis de ignorabilidad y en 03.2 el he-

cho de que la variable contrafacutal $Y^*(k)$ es idéntica a Y cuando $X = k$.

El valor esperado $\mathbb{E}[Y \mid X]$ es exactamente la regresión de Y sobre X, digamos $\mathbb{E}[Y \mid X] = \beta'_X \cdot X$. Como consecuencia, el average treatment effect (03.3) resulta en:

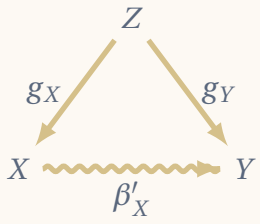
$$ATE = 2 \cdot \left(\beta'_X \cdot (1) - \beta'_X \cdot (0)\right) = 2 \cdot \beta'_X \tag{03.4}$$

Esta representación es importante porque hemos reducido el cálculo del averagre treatment effect a obtener el coeficiente de una regresión lineal.

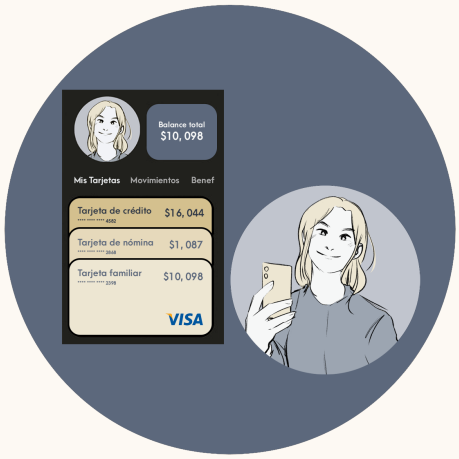
Double Machine Learning mediante FWL

Supongamos que tenemos un modelo $\hat{Y} = \beta_X \cdot X + f(Z)$, donde $f(Z)$ no necesariamente es un modelo lineal que toma en cuenta las covariables Z.

La técnica de Double Machine Learning consiste en construir dos modelos, uno que intente predecir ϵ_X y otro que intente predecir ϵ_Y, digamos $g_y(Z) = \vec{\epsilon}_X$ y $g_X(Z) = \vec{\epsilon}_Y$; para después medir el efecto causal de la variable X sobre Y entrenando una regresión lineal que aproxime $\hat{\epsilon}_Y = \beta'_X \cdot \vec{\epsilon}_X$.



04 Lectura de referencia: Inferencia Causal en IA



La lectura de referencia se pude encontrar en este link.

En este trabajo se presentan dos estudios causales. En el primero BBVA estudia el efecto de la variable X, su herramientas de salud financiera, en el ahorro de las personas.

En el segundo ejemplo muestran un estudio de cómo las recomendaciones que el banco hace en su aplicación de manera personalizada, tienen un efecto en la venta de sus productos financieros.

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