BOURBAKI

COLEGIO DE MATEMÁTICAS

Causalidad en Machine Learning I

Interpretabilidad, inferencia bayesiana y causalidad

BOURBAKI
COLEGIO DE MATEMÁTICAS

Índice

01. Introducción — pág. **02**

02. ¿Para qué sirve la causalidad? — pág. **03**

- 01. Lectura de referencia: Premio Nobel de Economía 2021 — pág. 05

03. Inferencia bayesiana — pág. **07**

- 01. El cambio de paradigma — pág. 08
- 02. Inferencia Bayesiana según Terence Tao pág. 09
- 03. El problema de la Bella Durmiente pág. 09

04. Interpretabilidad post hoc — pág. **12**

- 01. Ordenamiento de familias de características — pág. 13
- 02. Interpretación de los coeficientes de la regresión logística — pág. 14

01 Introducción

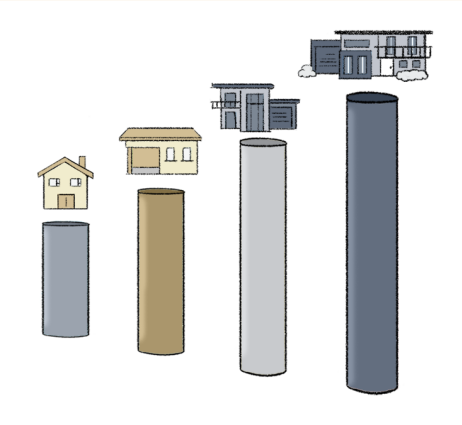
La causalidad es un análogo asimétrico del concepto de correlación, esta última es una de las relaciones más importantes en Machine Learning. En este curso, buscamos familiarizar al Científico de Datos con los fundamentos y las ideas detrás del estudio matemático de la casualidad. Por medio de casos 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 ¿Para qué sirve la causalidad?



Para entender cuál es el papel que juega la causalidad en la solución de problemas por medio de estrategias datadriven podríamos pensar en el siguiente problema perenne en la mente de los científicos de datos o analistas financieros: ¿cuánto cuesta una casa o una acción? Independientemente del precio al que se oferta, nos gustaría conocer el precio "real" con el posible objetivo de aprovechar alguna oportunidad de negocio.

Supongamos que tenemos una base de datos con un histórico que contiene información sobre los bienes, junto a un histórico de los precios en los que se han vendido, es decir una base de datos supervisada.

Machine Learning predictivo. Si entrenamos un modelo de machine learning como una regresión podemos predecir tanto los precios de nuestros bienes como de nuevos registros que nos interese valuar. Esto es muy útil y es el objetivo final en la mayor parte de los trabajos que realiza un científico de

datos. Si confiamos en nuestro modelo y nos ofrecen un bien más barato que lo que predice el modelo, es una buena oportunidad para comprarlo.

Machine Learning interpretable: Una vez que hemos entrenado este modelo, en algunas circunstancias nos gustaría comprender cuáles son las características más importantes que está tomando en cuenta el modelo para hacer estas predicciones. Esta pregunta podría resolverse a nivel global (toda la base de datos) o a nivel local (cada uno de los registros). La razón principal por la que queremos hacer esto es porque nos gustaría comprender las razones que está tomando en cuenta nuestro modelo entrenado. Si por ejemplo una variable como el ID de la propiedad tuviera mucha importancia, esta sería sin lugar a dudas una relación espuria pues no debería de importar.

Causalidad en Machine Learning El papel del estudio causal no es muy distinto al de la interpretabilidad que describimos en el punto anterior sin embargo en la práctica existe una gran diferencia. Supongamos que hemos encontrado un modelo que hace excelentes predicciones para el valor de una casa y después de hacer un estudio de la interpretabilidad global, este modelo toma mucho en cuenta una variable como el número de WC's. No es muy intuitivo este resultado sin embargo con el fin de mejorar las predicciones es razonable dejarla en su lugar en pos de una inferencia de calidad. Un estudio causal nos ayudaría a asegurarnos de que el número de los WC en una propiedad es una causa del precio y no solo estamos siendo víctimas de las relaciones estadísticas entre el resto de las variables.

Lectura de referencia: Premio Nobel de Economía 2021

El estudio de la causalidad ha cobrado un papel fundamental en las ciencias sociales y económicas, especialmente cuando los experimentos controlados no son viables. Un ejemplo destacado de su aplicabilidad es el trabajo de David Card, Joshua Angrist y Guido Imbens, galardonado con el Premio Nobel de Economía en 2021. Estos investigadores desarrollaron métodos innovadores para responder preguntas causales a partir de datos observacionales, es decir, sin necesidad de realizar experimentos aleatorios. Por ejemplo, David Card analizó el impacto del aumento del salario mínimo sobre el empleo, utilizando variaciones naturales en políticas regionales como "experimentos cuasi-naturales". Angrist e Imbens, por su parte, establecieron un marco estadístico para interpretar correctamente los efectos causales en este tipo de estudios. Su trabajo ha transformado la forma en que evaluamos políticas públicas, educación, salud y economía laboral, demostrando que, con las herramientas adecuadas, es posible inferir relaciones causales confiables a partir de datos del mundo real.

Uno de los ejemplos más interesantes que se recogen en el texto que premia con el Nobel a David Card, Joshua Angrist y Guido Imbens es el siguiente. Es bastante razonable imaginar que si observamos las variables: años de estudio e ingresos mensuales, observaríamos una correlación positiva. Todos sabemos que esta correlación no necesariamente es una relación causal, por

ejemplo podría ser que la gente que estudia más es porque son aristócratas y en ese caso sus ganancias no serían mayores a quienes estudian menos a causa del tiempo que hayan estudiado. Supongamos que la gente puede abandonar los estudios en el momento que cumplen los 18 años. Si el calendario escolar fuera de enero a Diciembre a elegir a todas las personas que abandonaron la escuela tanto en el primer trimestre como en el último trimestre. Afortunadamente la aristocracia no tiene ninguna preferencia sobre en qué época del año nacer así que al elegir a esta muestra estamos deshaciéndonos de la posibilidad de que los aristócratas sean quienes estudiaron más tiempo. Los investigadores de este experimento notaron una correlación aún mayor entre esta sub-muestra con los ingresos percibidos que en el total de la población. A este tipo de experimento le llaman experimento natural.

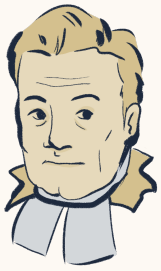
El artículo de referencia, *Answering causal questions using observational data-achievements of the 2021 Nobel Laureates in Economics* [1] se puede encontrar en este link.

03 Inferencia bayesiana

Tanto los modelos matemáticos como el cálculo de probabilidades dependen de las hipótesis que estemos dispuestos a asumir sobre el fenómeno que queremos analizar. Es importante que todas las personas involucradas en el uso de modelos de predicción (ya sean desarrolladores, analistas o usuarios) sean conscientes de esto.

Lo anterior no significa que debamos dejar de lado la evidencia que obtenemos al observar el fenómeno. De hecho, una de las mejores formas de combinar estas dos dimensiones del modelado las suposiciones y la evidencia es a través de la inferencia bayesiana, que hoy en día es una de las herramientas más poderosas para hacer análisis en contextos de incertidumbre.

La inferencia bayesiana, tal como la conocemos hoy, se basa en el uso del Teorema de Bayes para actualizar el cálculo de probabilidades a partir de la evidencia disponible. Este teorema fue demostrado por primera vez por el matemático Thomas Bayes en 1763, en su texto *An essay towards solving a problem in the doctrine of chances.*



La inferencia Bayesiana nos permite afrontar un problema de modelación matemática desde un punto de vista más general que mediante la estadística de los grandes números en la que el objetivo principal es aproximar una familia de parámetros desconocidos. En este caso no supondremos que tales parámetros existe sino que existe una distribución sobre la familia de los parámetros posibles.

El cambio de paradigma

Sea Y una variable o un vector aleatorio y β una familia de parámetros, podrían ser los parámetros que definen a una distribución o una función.

Example 01.1. *Podríamos suponer que Y son los retornos del precio de Bitcoin en un determinado tiempo y β son los parámetros de una distribución de Laplace con la que nos gustaría aproximar el histograma de Y.*

Example 01.2. *En el caso de una regresión o un problema supervisado modelado de manera paramétrica, Y en realidad es (X, Y). En el caso de las regresiones lineales por ejemplo, si $X \in \mathbb{R}^d$ entonces $\beta \in \mathbb{R}^{d+1}$.*

Definition 01.1. Siguiendo la notación de las líneas anteriores, un **modelado frecuentista** consiste en encontrar el β adecuado que maximice $\mathbb{P}\left(Y|\beta\right)$.

Definition 01.2. Un **modelado bayesiano** consiste en $\mathbb{P}\left(\beta|Y\right)$.

Notemos que gracias al teorema de Bayes ambas formas de modelar están

relacionadas de la siguiente manera:

$$\mathbb{P}\left(\beta|Y\right) = \frac{\mathbb{P}\left(Y|\beta\right)\cdot\mathbb{P}\left(\beta\right)}{\mathbb{P}\left(Y\right)} \tag{03.1}$$

Definition 01.3. Las cantidades de la ecuación anterior reciben los siguientes nombres:

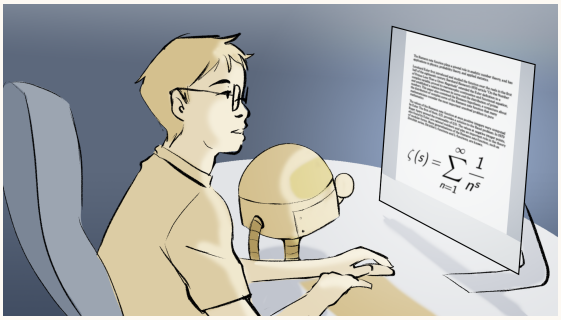
1. A la probabilidad $\mathbb{P}\left(\beta|Y\right)$ se le conoce como el **posterior**.
2. A la probabilidad $\mathbb{P}\left(Y|\beta\right)$ se le conoce como la **verosimilitud**.
3. A la probabilidad $\mathbb{P}\left(\beta\right)$ se le conoce como el **prior**.

La ventaja en este problema de la inferencia bayesiana es que nos permite agregar un conocimiento previo sobre nuestro problema, a saber el comportamiento del prior.

Inferencia Bayesiana según Terence Tao

Terence Tao, uno de los matemáticos más activos y sobresalientes de la actualidad, profesor de UCLA publicó hace algunos años en su blog un diagrama que propone para estructurar sistemáticamente el razonamiento bayesiano.

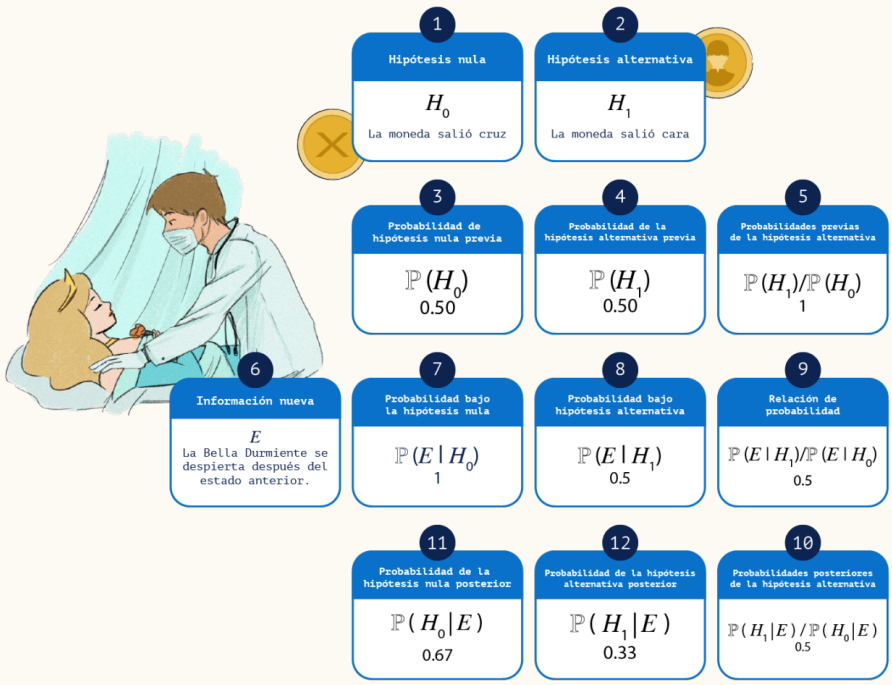
El problema de la Bella Durmiente



A continuación presentamos un ejemplo de un acertijo en teoría de la decisión que fue propuesto originalmente por Arnold Zuboff. Supongamos que la bella durmiente participará en el siguiente experimento que comienza el día domingo:

1. El día domingo por la tarde se anestesiará a la bella durmiente.
2. Inmediatamente después se lanzará una moneda justa (50 % cara y 50 % cruz) para decidir si se despertará solo el lunes o tanto lunes y martes a la bella durmiente.
3. Después despertar a la bella durmiente (ya sea en lunes o martes) se volverá a anestesiar y se le aplicará un medicamento que la hace olvidar que ha sido despertada.
4. El día miércoles termina el experimento y se despertará a la bella durmiente.
5. En cualquier momento que despertemos a la bella durmiente le preguntaremos si cree que la moneda cayó cara o cruz.

En este caso la evidencia E que ella tiene es que ha sido despertada y utilizan-



do el Teorema de Bayes se calculan las probabilidades posteriores (sujetas a la evidencia) de que el resultado de la moneda haya sido cara o cruz. Notemos que la hipótesis alternativa tiene una menor probabilidad de lo que hubiéramos imaginado.

04 Interpretabilidad post hoc

En este capítulo presentaremos tres aspectos fundamentales de la interpretabilidad post hoc. Es muy importante el adjetivo que estamos utilizando para la interpretabilidad pues nuestro énfasis será en técnicas que permita interpretar modelos ya entrenados. Este problema de la interpretabilidad es muy importante en machine learning desde un punto de vista práctico pues está relacionado con la herencia de modelos entre científicos de datos, la auditoría o simplemente la evaluación. Distinguiremos dos grandes clases de técnicas de interpretabilidad, aquellas globales y aquellas locales. Las técnicas globales son las que nos permiten interpretar a la función f en función de sus variables mientras que las locales interpretarán una predicción particular $f(x)$. Sobre las técnicas globales hemos incluido dos capítulos en los que hablaremos sobre el orden entre familias de características y otro sobre la traducción de variables latentes. Existe una relación estrecha entre las siguientes tres secciones y los tres enemigos de machine leanring que presentamos en el capítulo anterior. La relación detallada es la siguiente:

- El sobreajuste se puede mitigar ordenando las características de acuerdo a la importancia.
- El subajuste se puede reducir agregando variables latentes y no siempre es posible interpretar su significado.

- El ruido de un registro se puede detectar si comprendemos las razones por las que nuestro modelo hizo alguna predicción.

Ordenamiento de familias de características

Uno de los objetivos más importantes de la interpretabilidad es la capacidad de comparar la importancia entre dos variables X_j, X_k, más aún es muy importante poder comparar la importancia entre dos familia de variables $X_{j_1,\dots,j_n}, X_{k_1,\dots,k_n}$.

Proposition 01.1. *La correlación de Pearson nos permite comparar cualesquiera dos variables X_j, X_k siempre y cuando solo nos interesen modelos f univariados, es decir que si solo deseamos agregar alguna de las características X_j podemos elegir cuál nos conviene más.*

Proposition 01.2. *Si se estandarizan los valores de S, los pesos en los modelos lineales nos permiten comparar cualesquiera dos variables X_j, X_k inclusive en el caso multivariado.*

Proposition 01.3. *Los árboles de decisión nos permiten comparar cualesquiera dos variables X_j, X_k inclusive en el caso multivariado. La manera de hacerlo es comparando la posición de los nodos correspondientes dentro de los árboles.*

Proposition 01.4. *Existen resultados positivos para el caso de dos subconjuntos de variables.*

Definition 01.1. Si un modelo incluye un orden ya sea para variables individuales o para subconjuntos de variables diremos que el modelo es interpretable en X.

Interpretación de los coeficientes de la regresión logística

Utilizando la hipótesis **??** que satisfacen los datos en la regresión logística es posible utilizar la siguiente ecuación para interpretar los valores de los pesos β_i que obtenemos al entrenarla.

Recordemos que salvo un error i.i.d. nuestros datos satisfacen:

$$\log\left(\frac{\mathbb{P}\left(y=1|x\right)}{\mathbb{P}\left(y=0|x\right)}\right)=\langle\beta,x\rangle$$

Normalmente en las regresiones lineales se puede interpretar la importancia de una de las variables explicativas en función del tamaño y signo de β_i. En este ejemplo podríamos hacer algo similar pues el signo y el tamaño de β_i afectará a la cantidad $\log\left(\frac{\mathbb{P}\left(y=1|x\right)}{\mathbb{P}\left(y=0|x\right)}\right)$.

Desafortunadamente esto podría no ser muy práctico. Para remediar este problema vamos a desarrollar algebraicamente la ecuación anterior para que sea más sencillo entender la importancia de los pesos.

Al aplicar la función e^X de ambos lados de la ecuación obtenemos

$$\frac{\mathbb{P}\left(y=1|x\right)}{\mathbb{P}\left(y=0|x\right)} = e^{\langle \beta, x\rangle}$$

La cantidad anterior puede utilizarse para conocer la importancia de una cordenada j en la clasificación de un ejemplo $x = (x_1, \ldots, x_d)$ utilizando el siguiente procedimiento: supongamos que aumentamos artificialmente el valor de $x_i + 1$, comparando el aumento con el valor original y dividiendo obtenemos:

$$\frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_i(x_i+1)+\ldots+\beta_d x_d}}{e^{\langle \beta, x\rangle}} = e^{\beta_i(x_i+1)-\beta_i(x_i)} = e^{\beta_i}$$

Es decir

$$\frac{\dfrac{\mathbb{P}\left(y=1|\,(x_1,\ldots,x_i+1,\ldots,x_d)\right)}{\mathbb{P}\left(y=0|\,(x_1,\ldots,x_i+1,\ldots,x_d)\right)}}{\dfrac{\mathbb{P}\left(y=1|x\right)}{\mathbb{P}\left(y=0|x\right)}} = e^{\beta_i}$$

Esta última ecuación nos permite explicar la importancia de β_i en la predicción de y.

Exercise 02.1. *Supongamos que tenemos una sola variable explicativa x que además supondremos binaria, en el caso del problema de las reseñas de hoteles esto significaría que solo tenemos una palabra que puede o no aparecer. Si $y = 1$ significa que la reseña es falsa, explicar qué significa un alto valor de β y qué significa un pequeño valor de β.*

Bibliografía

[1] Zoltan Hermann, Hedvig Horváth, and Attila Lindner. Answering causal questions using observational data-achievements of the 2021 nobel laureates in economics. *Financial and Economic Review*, 21(1):141–163, 2022.

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