BOURBAKI

COLEGIO DE MATEMÁTICAS

Causalidad en Machine Learning V

Aplicaciones de la causalidad en sistemas de recomendación

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: test 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 Preliminares sobre los sistemas de recomendación

Aunque los primeros sistemas de recomendación que utilizan técnicas de collaborative filtering se entrenan por medio de la factorización de matrices, existe otros métodos ligeramente distintos.

Sea S el conjunto de parejas (u, p) tales que sí existe una calificación $R(u, p)$ dentro de nuestra base de datos. En el caso de la base de datos de netflix este valor corresponde precisamente a la entrada $X(u, p)$ de la matriz de calificaciones.

Es posible considerar la siguiente métrica para cualesquiera vectores (más adelante diremos cómo se podrían entrenar) $\hat{X}_u, \hat{X}_p \in \mathbb{R}^K$:

$$\sum_{(u,p)\in S} \left(R(u,p) - \hat{x}_u \cdot \hat{X}_p\right)^2$$

Comúnmente se agregará a la ecuación anterior un factor de regularización como en Ridge o Lasso de la forma $\lambda\left(||\hat{x}_u||_2^2 + ||\hat{X}_u||_2^2\right)$, también es posible agregar ordenadas al origen entrenables para los vectores en dimensión K correspondientes con usuarios y películas respectivamente.

Definition 00.1. Comenzamos con vectores al azar $\hat{x}_u(0), \hat{X}_p(0)$ y fijamos $\alpha > 0$ una tasa de aprendizaje. Definimos el algoritmo del gradiente descendente de la siguiente manera:

1. $\hat{x}_u(t+1) = \hat{x}_u(t) - \alpha(\epsilon_{u,p}(t)\hat{X}_p(t) - \lambda\hat{x}_u(t))$
2. $\hat{X}_p(t+1) = \hat{X}_u(t) - \alpha(\epsilon_{u,p}(t)\hat{x}_u(t) - \lambda\hat{X}_p(t))$

Donde $\epsilon_{u,p}(t) = R(u,p) - \hat{x}_u(t) \cdot \hat{X}_p(t)$ es el error que comete en el tiempo t al hacer la recomendación. La intención de las fórmulas anteriores es incorporar la información de una película en las coordenadas de un usuario cuando en la iteración anterior el error haya sido grande y en caso de que el error sea pequeño, no modificar las coordenadas ni del usuario ni de la película. El valor α será la tasa de aprendizaje de estas actualizaciones. Cuando el algoritmo busque regularizar entonces a la información de la película que se le desea agregar al usuario se le restará la información del usuario.

00.1 Nuevos usuarios o nuevas películas

Es importante mencionar que si recibimos un nuevo usuario sin ninguna información sobre sus preferencias por ejemplo u^*, el algoritmo anterior podría inicializarlo con un valor aleatorio $x_{u^*}(0)$ y con ese ya hacerle recomendaciones. Con el feedback que reciba el algoritmo mediante algunos pocos ratings ya será posible comenzar a hacer actualizaciones en el vector $x_{u^*}(t)$. Es muy difícil imaginar cómo podría funcionar la solución exacta de SVD para incorporar la información de un nuevo usuario sin embargo la naturaleza iterativa de este algoritmo por medio del gradiente nos permite hacer estas predicciones sin mayor problema.

03 Inverse Probability Weighting

En el módulo pasado estudiamos cómo bajo algunas hipótesis sobre una red bayesiana es posible posible aproximar probabilísticamente a una variable objetivo Y bajo la hipótesis de la intervención (operador Do) sobre algún conjunto de variables X en la presencia de otro conjunto de covariables Z utilizando únicamente información estadística sobre la base de datos (Z, X, Y).

La siguiente fórmula es conocida como la función de ajuste y es la conclusión del teorema del criterio de Back-Door.

$$\mathbb{P}(Y|do(X), Z_1) = \mathbb{P}(Y|X, Z_1)\mathbb{P}(Z_1) + \ldots + \mathbb{P}(Y|X, Z_K)\mathbb{P}(Z_K)$$

Notemos que el lado derecho de la ecuación puede calcularse con observaciones empíricas sin la necesidad de intervenir. Solo para aligerar la notación supongamos que tenemos solo una covariable $Z_{1=K}$, la ecuación anterior se puede re-escribir se la siguiente manera gracias a la definición de la probabilidad condicional:

$$\mathbb{P}(Y|do(X), Z_1) = \mathbb{P}(Y|X, Z_1)\mathbb{P}(Z_1) = \frac{\mathbb{P}(Z_1, X, Y) \cdot \mathbb{P}(Z_1)}{\mathbb{P}(X, Z_1)}$$

Ahora notemos que $\frac{\mathbb{P}(Z_1)}{\mathbb{P}(X,Z_1)} = \frac{1}{\mathbb{P}(X|Z_1)}$ por lo que podemos re-escribir la fórmu-

la anterior de la siguiente manera:

$$\mathbb{P}(Y|do(X), Z_1) = \frac{\mathbb{P}(Z_1, X, Y)}{\mathbb{P}(X|Z_1)}$$

A esta cantidad que aparece como el denominador $\mathbb{P}(X|Z_1)$ la llamaremos a partir de ahora el Propensity Score y a su inverso multiplicativo $\frac{1}{\mathbb{P}(X|Z_1)}$ el inverse probability weighting. Es muy importante notar que todo lo anterior es cierto únicamente cuando es válido el criterio Back-Door pues de otra forma no podríamos concluir lo mismo, el siguiente ejemplo nos puede ayudar a comprender esta imposibilidad:

Example 00.1. *Deseamos calcular la relación causal entre el año X de los vehículos y el precio Y. Supongamos que en presencia de Z, los valores de X son o grandes o pequeños pero rara vez intermedios. Por ejemplo podríamos pensar que $Z = 1$ es una página de internet CARS que vende automóviles, X es el año de los automóviles y la hipótesis significa que la mayor parte de los automóviles que venden en este sitio web son o muy antiguos o muy recientes pero difícilmente venden otros modelos. Si por alguna razón hubiera un sesgo dentro de nuestra base de datos para vehículos ofertados en CARS, la mayor parte de los registros satisfacen $(Z = 1, X, Y)$ por lo cual la mayor parte de los elementos que nos ayudan a calcular $\mathbb{P}(Y|do(X), Z)$ son divididos por $\mathbb{P}(X|Z = 1)$ el cual por nuestra hipótesis será un número grande para X grande y pequeña, por el otro lado será pequeño cuando X sea un año intermedio. Gracias a lo anterior, nos quedaremos con la mayor parte de registros donde X es interme-*

dio lo cual es muy problemático pues si queremos calcular una regresión entre X e Y en nuestra nueva base de datos tendríamos muy pocos registros.

Aplicaciones a los sistemas de recomendación

Cuando pensamos en un problema de un sistema de recomendación inmediatamente surge una tensión causal entre las siguientes variables cuando Y representa si el usuario ha visto o le ha gustado algún producto p:

- Las características X del producto recomendado, por ejemplo en el caso de Netflix podría ser si es o no un documental.
- La información Z de si el sistema de recomendación le ha propuesto o no al usuario ver ese producto p.

De acuerdo a lo anterior podríamos utilizar inferencia causal y en particular los IPW para aproximar la relación causal $\mathbb{P}(Y|do(X), Z)$.

Supondremos en este ejercicio que tenemos acceso a dos matrices distintas $R(u,p)$ y $E(u,p)$ donde la primera representa el ranking que le ha dado el usuario u al producto p (de la misma forma que lo enseñamos en el capítulo anterior), mientras que la nueva matriz E calcula si en algún momento el usuario u ha tenido la oportunidad o no de ver a p.

La idea general es ponderar los datos observados en la base de datos R mediante algún modelo probabilista asociado a la matriz E. Por ejemplo su-

pongamos un modelo muy simple en el que la exposición E solo dependiera de la popularidad que tiene la película p, es decir que la columna $E(u,p) \sim Bernoulli(\rho_p)$. En este caso podríamos ponderar por el inverso del propensity score $\rho_p = \mathbb{P}\left(X = u | Z = p\right)$. En general utilizando a la matriz E podemos calcular el propensity score.

04 Lectura de referencia: Inferencia Causal y los Valores Faltantes

- En la lectura de referencia de esta semana, que se puede encontrar en este link; los autores nos invitan a repensar todo el paradigma de resultados potenciales como un problema de datos faltantes, convirtiendo así la imposibilidad de observar conttrafactuales en un terreno más familiar para los científicos de datos.

  Los autores abordan analogías entre los métodos clásicos para el análisis de datos faltantes como la imputación, inverse probability weighting y métodos doblemente robustos (que combinan los dos anteriores); con métodos de inferencia causal bajo cada uno de los modos de inferencia: Frecuentista, bayesiana y fisheriana

05 Caso de uso: sistemas de recomendación

En el artículo **Causal inference for recomendation**, los autores desarrollan proponen una técnica basada en inferencia causal para superar algunas de las limitaciones que ofrece el enfoque clásico de sistemas de recomendación, pues este último solo considera los datos de valoración del usuario; mientras que el enfoque causal también considera el sesgo causado por la exposición al producto. Para ello se presentan dos fuentes de información: los productos que decidió ver cada usuario (mecanismo de exposición) y cuáles de ellos rankeó positivamente.

Los autores emplean **inverse probability weighting** para ajustar los datos observados para que parezcan como si los productos hubieran sido ofrecidos de forma aleatoria, flexibilizando el modelo de exposición y mejorando la capacidad de recomendar nuevos items.

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