Machine Learning

Selección de Grupo de Control

Martín Dominguez, David Giuliodori, Juan Porta y Alejandro Rodriguez

Instituto de Economía y Finanzas - Facultad de Matemática, Astronomía y Física - Universidad de Talca

2022

Machine Learning: Selección de Grupo de Control

¿En qué consiste una evaluación de políticas?

- En medir el impacto del programa (o tratamiento) sobre un conjunto de variables de resultado
- Las variables de resultado son las variables sobre las cuales se espera que el programa tenga un efecto en los individuos beneficiarios del programa evaluado

El problema de evaluación de impacto consiste en establecer la diferencia entre la variable de resultado del individuo/empresa participante en el programa en presencia del programa y la variable de resultado de ese mismo individuo/empresa en ausencia del programa. Esta diferencia es lo que se conoce como efecto del tratamiento o programa.
Dado que los individuos no pueden participar y no participar al mismo tiempo, la estimación del impacto del programa requiere estimar un contra-fáctico.

Machine Learning: Selección de Grupo de Control

Dicho lo anterior, surge una primera pregunta:

¿por qué no se compara la variable de individuos que participaron del tratamiento contra la variable de aquellos no lo hicieron?

La respuesta a esta pregunta es relativamente simple: la variable de aquellos individuos que participan del tratamiento puede ser diferente a la de los individuos que no participan como resultado sólo del tratamiento, sino por otros factores propios de cada grupo.

Machine Learning: Selección de Grupo de Control

La segunda pregunta que surge es:

¿cómo estimamos el impacto?

Si la participación en el tratamiento se lleva a cabo en forma aleatoria, entonces es posible hacer una comparación entre los dos grupos, tratados y no tratados, porque las características que hacen diferentes a los individuos se distribuyen en forma aleatoria entre los dos grupos.

Machine Learning: Selección de Grupo de Control

Ahora, el problema de esta alternativa, es que en la práctica, casi no existen casos, en donde la asignación se lleva a cabo en forma aleatoria. En la mayorías de los casos, la participación del tratamiento suele depender del mismo individuo y, como consecuencia, termina generando dos grupos que tienen características diferenciadoras que impactan sobre la variable objetivo.

La literatura resuelve el problema de la auto-selección mediante técnicas de medición que se denominan cuasi-experimentales.

La idea es tener un grupo de control que cumpla con ciertos supuestos que permitan inferir que la diferencia en la variable objetivo de ambos grupos sólo será significativa si el tratamiento tuvo su impacto.

Machine Learning: Selección de Grupo de Control

La tercera pregunta que surge es:

¿cómo se obtiene un grupo de control?

Se debe buscar un grupo de control que cumpla con ciertos supuestos

Machine Learning: Selección de Grupo de Control

En términos matemáticos el efecto promedio del programa sobre los individuos beneficiarios, se define como una esperanza condicional (a que el individuo participe) de la diferencia entre el resultado cuando el individuo participa y el resultado que obtendría si no participa, es decir,

$$E(Y_i(1) - Y_i(0)|D = 1) \tag{1}$$

donde D es una variable dicotómica que toma el valor 1 cuando el individuo es beneficiario y 0 cuando no lo es. $Y_i(1)$ es la variable de resultado del individuo i cuando participa del tratamiento, y $Y_i(0)$ cuando ese mismo individuo no participa.

El problema ahora se reduce a la estimación de la respuesta contrafactual promedio de los individuos que fueron beneficiarios.

Machine Learning: Selección de Grupo de Control

Una primera solución es condicionar sobre estas características para controlar esta fuente de sesgo. Si X es un vector características, el estimador del impacto promedio del programa sobre los beneficiarios es:

$$ATT = E(Y_i(1)|D = 1, X) - E(Y_i(0)|D = 1, X) \tag{2}$$

Claramente no es posible observar ambos resultados al mismo tiempo. Sin embargo, sí se puede observar la variable de resultado entre un grupo de individuos elegibles que no participan en el programa (o grupo de control), $E(Y_i(0)|D = 0, X)$. El principal reto de la evaluación de impacto es determinar las condiciones bajo las cuales $E(Y_i(0)|D = 0, X)$ se puede utilizar como una aproximación válida de $E(Y_i(0)|D = 1, X)$.

Machine Learning: Selección de Grupo de Control

En el caso de que se disponga de un panel de individuos, es posible aplicar el método de Diferencias en Diferencias (DID).
La idea general del método DID consiste en estimar el efecto tratamiento considerando las respuestas antes y después de la aplicación del programa en ambos grupos (tratamiento y control) a fin de barrer efectos fijos en el tiempo y luego medir las diferencias entre estas comparaciones.

$$ATT = E(Y_{t_1,i}(1) - Y_{t_0,i}(1)|D = 1, X) - E(Y_{t_1,i}(0) - Y_{t_0,i}(0)|D = 0, X) \quad (3)$$

donde t_0 y t_1 son el periodo anterior y posterior a la aplicación del programa, y X es el vector de características de las empresas.

Machine Learning: Selección de Grupo de Control

Cuando la información temporal disponible de los datos es mayor a uno y además, el método DID es equivalente al estimador within, lo que permite estimar el efecto usando un modelo de regresión simple en datos de panel. Cuando existen más periodos disponibles en la base de datos, el estimador DID es el parámetro γ de la siguiente regresión (estimado por efectos fijos):

$$Y_{it} = \gamma D_{it} + \beta \boldsymbol{x}_{it} + \alpha_i + \lambda_t + \varepsilon_{it}, \tag{4}$$

donde Y_{it} es la variable respuesta, y D_{it} es una variable dicotómica que toma el valor 1 después que el individuo comienza a participar del programa.

Machine Learning: Selección de Grupo de Control

Incluso si los existen componentes no observables invariables en el tiempo y se tienen co-variables observabas sobre los individuos, el método DID puede generar sesgos en la estimación simplemente porque el modelo que se plantea es lineal en las variables, y como uno sabe, el mundo el altamente no lineal.

Para complementar el método de DID, la literatura propone encontrar un grupo de control mediante algún sistema de permita encontrar individuos que sean tan parecidos como sea posible a los que participan del tratamiento.

El método más utilizado, por su simplicidad y robustez, es el Propensity Score Matching.

Machine Learning: Selección de Grupo de Control

Es método Propensity Score Matching se basa en el supuesto de solapamiento, Para cualquier valores de $\boldsymbol{X}_i$ se verifica que

$$0 < P(D_i = 1 \mid X_i) < 1. \tag{5}$$

Lo que implica esta desigualdad es que hay una mezcla dentro de la población de individuos tratados y no tratados.

La idea es emparejar individuos/empresas teniendo en cuenta la probabilidad estimada de participar en el programa, dadas las características observables $P(X)$:

$$P(X_i) = P(D_i = 1 \mid X_i) \tag{6}$$

Es decir, el "clon" adecuado para cada individuo del grupo de tratamiento será aquel del grupo de control con una probabilidad de participación en el programa suficientemente cercana.

Machine Learning: Selección de Grupo de Control

La función $P(X_i)$ se conoce con el nombre de propensity score, o probabilidad de participación.

PSM sólo se puede calcular en la región de soporte común para asegurar que los grupos de tratamiento y control sean muy parecidos. Esta condición se conoce como **Condición de Soporte Común**, esto implica la necesidad del supuesto (5).

La estimación de la probabilidad se suele hacer utilizando un modelo logístico o un modelo probit.

Machine Learning: Selección de Grupo de Control

Problema

Cuando el ingreso o al tratamiento por parte de los individuos no se realizar en un **único instante en el tiempo**, sino que ocurre en forma **secuencial**, los modelos logísticos que se utilizan para poder obtener el score de matching no son capaz, desde el punto de vista de su concepción, gestionar esa probabilidad variable en el tiempo.

Lo que se hace es estimar la probabilidad por camada de ingreso al programa, es decir, juntar grupo de ingresos aquellos individuos que ingresan al tratamiento en temporalidades similares, y se buscan controles para cada una de estas camadas.

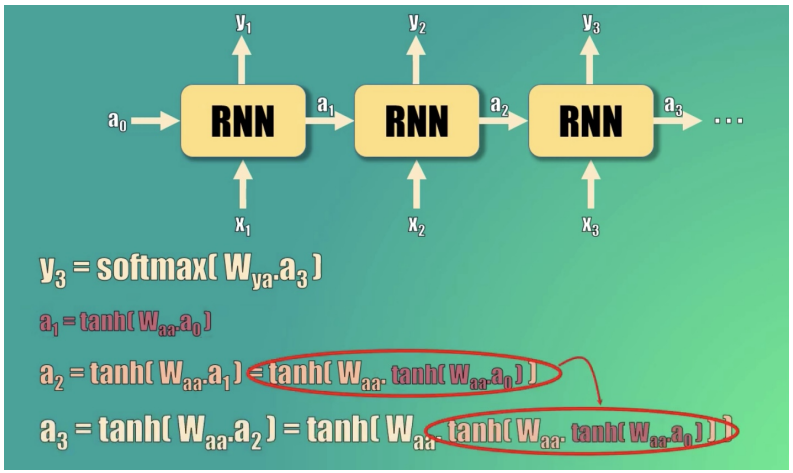
Machine Learning: Selección de Grupo de Control

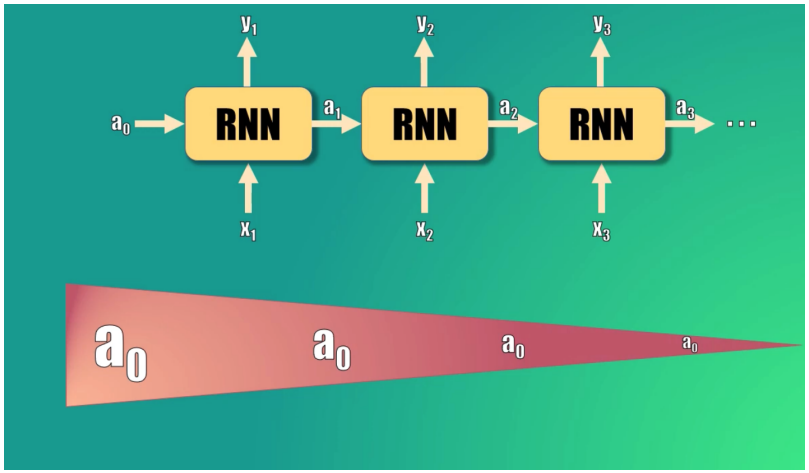
Los modelos de Machine Learling (ML) son un buen competidor de los modelos logísticos o probit para la estimación de la probabilidad de participación en el programa. Algunos de estos métodos tienen el mismo problema que los modelos logístico.

Sin embargo, otros métodos dentro ML, como son las redes neuronales permiten incorporar la dimensión temporal de los datos para obtener grupos de control para cada individuo tratado teniendo en cuenta el ingreso secuencial.

Machine Learning: Selección de Grupo de Control

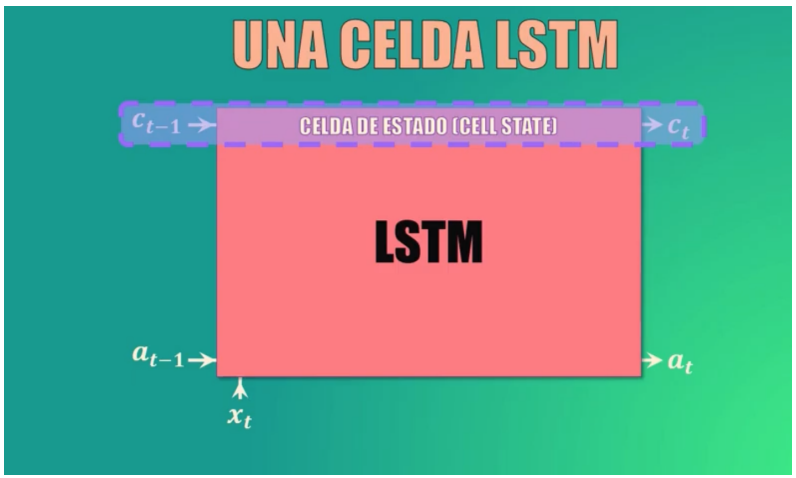
Las Recurrent Neural Nets (RNN) son una familia de redes en las que se comparten características entre nodos, si los nodos los trabajamos como entradas temporales, entonces tenemos dinámica en la red.





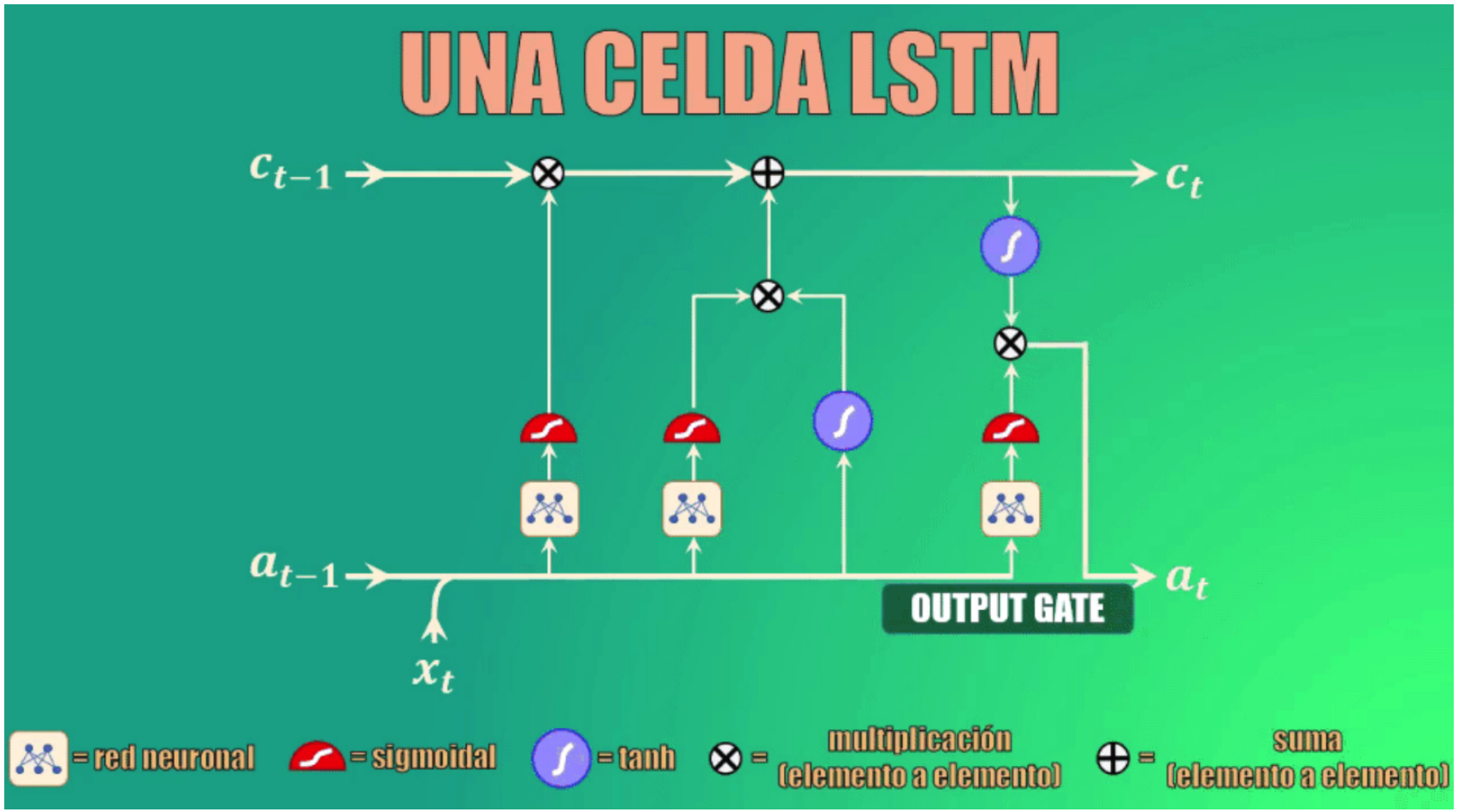
Machine Learning: Selección de Grupo de Control

Mientras las redes recurrentes estándar pueden modelar dependencias a corto plazo, las LSTM (Long Short Term Memory) pueden aprender dependencias largas, por lo que se podría decir que tienen una memoria a más largo plazo.



Machine Learning: Selección de Grupo de Control

Funcionamiento de una celda LSTM.



Machine Learning: Selección de Grupo de Control

Evolución de la memoria de largo plazo en una red LSTM.

Machine Learning: Selección de Grupo de Control

Simulación de Monte Carlo

Se analizaron una serie de combinaciones de procesos generadores de datos (PGD).

- Entrada secuencial desde el período 4 un un total de 10
- Entrada secuencial desde el período 7 un un total de 15
- Diferentes niveles de persistencia en la variable resultado - Proceso AR(1) - ($\phi = 0.5, 0.85, 0.9$)
- Solapamiento de la distribución de los grupos de control/tratados versus los NiNi
- Relación entre variable resultado y la participación del tratamiento:
  - Lineal
  - No lineal
  - Intensidad de la relación

Machine Learning: Selección de Grupo de Control

Se realizaron 36 escenario distintos con 100 simulaciones por escenario, con 1000 individuos en cada simulación.
Se compararon las siguientes metodologías:

- PSM + Logit por cohorte cuando existe entrada secuencial
- XGBoost por cohorte cuando existe entrada secuencial
- LightGBoost por cohorte cuando existe entrada secuencial
- LSTM (como incorpora la temporalidad no hace falta estimar por cohorte)

Las métricas objetivos estuvieron a analizar fueron:

- Sesgo
- MSE

Machine Learning: Selección de Grupo de Control

Primeros Resultados y Algunas Conclusiones Preliminares

- Cuando el programa se implementa en un mismo instante de tiempo y no hay dependencia temporal, PSM es el mejor estimador
- Cuando el programa se implementa secuencialmente, pero no existe dependencia temporal en el período pre-tratamiento, las simulaciones tienen menor sesgo y MSE en la LSTM
- Cuando la dependencia temporal es baja, PSM por cohorte y LSTM tienen un desempeño similar
- Cuando la dependencia temporal es alta, LSTM tiene un mejor desempeño y el sesgo es bajo en comparación con los otros métodos
- XGBoost y LightXBoost no parecen mejor significativamente en ningún escenario el sesgo ni el MSE

Machine Learning: Selección de Grupo de Control

Próximos Pasos

- Incorporar variables exógenas a los modelos
- Estimar el impacto en programas ya previamente evaluados con otras técnicas, para ello se usarán datos administrativos