

Planteamiento del problema general de predicción

Introducción

Fernando Arias-Rodríguez

Banco Central de Bolivia

29 de agosto de 2024



- ① Premisa
- ② Objetivos y terminología
- ③ Validación, validación cruzada, sobreajuste y regularización.
- ④ Inferencia

1 Premisa

Objetivo del curso

2 Objetivos y terminología

3 Validación, validación cruzada, sobreajuste y regularización.

4 Inferencia

Como en la estadística, esta rama de la econometría busca ampliar la caja de herramientas del investigador, incluyendo metodologías que no dependan necesariamente de modelación de datos (p.e. el supuesto de procesos generadores de datos).

Se proponen técnicas que utilizan *modelos algorítmicos*, donde los mecanismos que generan los datos se suponen desconocidos.

Los modelos algorítmicos demuestran ventaja en predicciones de datos por fuera de muestra. Sin embargo, esto no es suficiente para tener un uso amplio en economía.

Muchas de estas técnicas *NO* permiten hacer inferencia - interpretación de coeficientes, construcción de intervalos de confianza.

También, requieren una fina calibración y adaptación para abordar adecuadamente los problemas que interesan a los economistas.

1 Premisa

Objetivo del curso

2 Objetivos y terminología

3 Validación, validación cruzada, sobreajuste y regularización.

4 Inferencia

El objetivo de este curso es discutir un subconjunto de herramientas que deberían hacer parte del kit de los econométristas enfocados en el estudio empírico de fenómenos económicos.

Se hace énfasis en modelos de pronóstico de series de tiempo macroeconómicas.

1 Premisa

2 Objetivos y terminología

Objetivos

Terminología

3 Validación, validación cruzada, sobreajuste y regularización.

4 Inferencia

1 Premisa

2 Objetivos y terminología

Objetivos

Terminología

3 Validación, validación cruzada, sobreajuste y regularización.

4 Inferencia

El acercamiento *tradicional* en econometría ha sido:

- 1 Especificar parámetros como una forma funcional de la distribución conjunta de los datos.
- 2 Con una muestra aleatoria, se hallan estimadores de los parámetros tales que se ajusten lo mejor posible, usando una función objetivo - función de verosimilitud o de minimización de errores.
- 3 La calidad de los estimadores se mide en términos de eficiencia con muestra grande - inferencia. También, hay interés en construir intervalos de confianza.

La literatura de modelos algorítmicos es construir predicciones sobre algunas variables, dado un conjunto de información.

La diferencia puede verse en estos términos:

- 1 Método tradicional: modelar la distribución condicional de Y_i dado un vector de información X_i : $Y_i|X_i \sim N(\alpha + \beta'X_i, \sigma^2)$.

El objetivo es estimar $\theta = (\alpha, \beta)$ tal que

$$(\hat{\alpha}_{ls}, \hat{\beta}_{ls}) = \min_{\alpha, \beta} \sum_{i=1}^N (Y_i - \alpha - \beta'X_i)^2.$$

- 2 Algorítmico: hacer un pronóstico, Y_{N+1} , a partir de un algoritmo e información de evaluación X_{N+1} :

$\hat{Y}_{N+1} = \hat{\alpha} + \hat{\beta}'X_{N+1}$, para algún $(\hat{\alpha}, \hat{\beta})$ y una función de pérdida $(Y_{N+1} - \hat{Y}_{N+1})^2$.

Bajo el enfoque algorítmico, la pregunta es cómo llegar a $(\hat{\alpha}, \hat{\beta})$ con buenas propiedades, de acuerdo con la función de pérdida.

Una alternativa es Mínimos Cuadrados Ordinarios, pero cuando el espacio de información de los regresores es amplio, hay otras metodologías alternativas que hacen mucho mejor trabajo.

- Reglas de decisión basadas en datos introducen un problema al ejercicio de pronóstico: un balanceo entre sesgo y varianza.
- Existen algoritmos que se enfocan en solucionar el problema de sesgo, con el costo de aumentar el problema de varianza, llevando a sobreidentificación.
- Otros algoritmos - reducción de dimensionalidad, por ejemplo - agregan algún sesgo con la intención de reducir la varianza de pronóstico.

1 Premisa

2 Objetivos y terminología

Objetivos

Terminología

3 Validación, validación cruzada, sobreajuste y regularización.

4 Inferencia

- La literatura de estos temas es antigua, por lo que maneja una terminología que resulta desconocida en econometría.
- Sin embargo, es posible emparejar términos entre ambos enfoques:
 - 1 Trad: muestra de estimación \equiv Alg: muestra de entrenamiento.
 - 2 Trad: el modelo se estima \equiv Alg: el modelo se entrena.
 - 3 Trad: regresores, predictores o covariables \equiv Alg: características (*features*).
 - 4 Trad: estimadores, parámetros de la regresión \equiv Alg: pesos (*weights*).

Los problemas de predicción se dividen en dos grandes grupos:

- 1 Problemas de aprendizaje supervisado (*supervised learning problems*), donde se observan tanto los predictores (*features*), X_i y el producto o variable de salida Y_i .
- 2 Problemas de aprendizaje no supervisado (*Unsupervised learning problems*), donde solo observamos las X_i y tratamos de agruparlas en clústeres o alguna otra medida con la que se pueda estimar su distribución conjunta.

Este curso desarrollará dos grupos de técnicas de estimación, las cuales aprovechan la mayor disponibilidad de información y la evolución de los computadores.

- Modelos lineales de alta dimensionalidad.
- Modelos que pretenden capturar no linealidades e interacciones de orden superior entre los regresores.

- 1 Premisa
- 2 Objetivos y terminología
- 3 Validación, validación cruzada, sobreajuste y regularización.
 - Validación y validación cruzada
 - Sobreajuste y regularización
- 4 Inferencia

- 1 Premisa
- 2 Objetivos y terminología
- 3 Validación, validación cruzada, sobreajuste y regularización.
Validación y validación cruzada
Sobreajuste y regularización
- 4 Inferencia

- Las técnicas econométricas tradicionales parten de un modelo teórico. Así, la *validación* consiste en implementar pruebas de hipótesis para contrastar si lo que se estima es coherente con la teoría económica.
- En el enfoque algorítmico, cada variable define su permanencia a partir de si mejora o no el poder predictivo del modelo. Esto es lo que se conoce como *validación cruzada*, la cual se funda en dos elementos clave:
 - El fin es tener el mejor poder predictivo. No importa cuántos o cuáles sean los parámetros o la forma funcional que hay detrás.
 - La validación cruzada utiliza comparaciones por fuera de muestra y no medidas de bondad de ajuste por dentro de muestra.

- 1 Premisa
- 2 Objetivos y terminología
- 3 Validación, validación cruzada, sobreajuste y regularización.
Validación y validación cruzada
Sobreajuste y regularización
- 4 Inferencia

El sobreajuste es una preocupación en el enfoque algorítmico, especialmente en *Machine Learning*. La intención es desarrollar modelos flexibles que ajusten bien, pero sin comprometer la predicción por fuera de muestra.

En lugar de optimizar una función objetivo, se añade un término a la función objetivo para penalizar la complejidad del modelo. Esto se conoce como *regularización*.

La regularización moderna es muy basada en datos, por lo que esta se determina explícitamente por el desempeño del pronóstico por fuera de muestra.

- *Sparsity* o escasez puede asociarse con el término parsimonia en econometría tradicional.
- Este principio busca implementar un algoritmo de pronóstico con el número exacto de variables incumbentes en el mismo, es decir, hallando el subconjunto óptimo de variables tal que el pronóstico sea el mejor posible.
- A diferencia de la econometría clásica, donde las variables entran en un modelo a partir de consideraciones teóricas, en el enfoque algorítmico se propone incluir variables en función de los datos.

- 1 Premisa
- 2 Objetivos y terminología
- 3 Validación, validación cruzada, sobreajuste y regularización.
- 4 Inferencia**
 - Ensamblaje de modelos
 - Inferencia

- 1 Premisa
- 2 Objetivos y terminología
- 3 Validación, validación cruzada, sobreajuste y regularización.
- 4 Inferencia**
 - Ensamblaje de modelos
 - Inferencia

Ensamblaje de modelos

- En muchos casos, un solo algoritmo no suele desempeñarse bien.
- Una solución es combinar diferentes modelos, promediados con ponderadores (o *votes*).
- A diferencia del acercamiento tradicional, los modelos involucrados pueden ser bien diferentes. Los pesos de la combinación se obtienen a partir de la optimización del poder predictivo por fuera muestra y no por bondad de ajuste.

- 1 Premisa
- 2 Objetivos y terminología
- 3 Validación, validación cruzada, sobreajuste y regularización.
- 4 Inferencia**
 - Ensamblaje de modelos
 - Inferencia

- El enfoque algorítmico se enfoca en desempeño del pronóstico por fuera de muestra, a expensas de la inferencia estadística - intervalos de confianza, análisis de estimadores -. Para varios métodos, es actualmente imposible construir intervalos de confianza válidos.
- Así, estos métodos suelen ser usados cuando el pronóstico es el más importante de los objetivos.
- Aun en casos donde se pueda hacer inferencia, esta viene con el costo de empeorar el pronóstico.