Clase 189 — DoubleML / EconML: Machine Learning para causalidad

Parte: 3 — Estadística Inferencial y Causal · Fuente: Chernozhukov et al. (2018) DML + docs DoubleML + EconML. Duración estimada: 95 min.

Clase 189 — DoubleML / EconML: Machine Learning para causalidad

> *Parte: 3 — Estadística Inferencial y Causal · Fuente: Chernozhukov et al. (2018) DML + docs DoubleML + EconML. Duración estimada: 95 min.*

Objetivo

Aplicar Double Machine Learning (Chernozhukov 2018) y EconML (Microsoft Research) para estimar ATE y CATE (Conditional ATE — heterogeneidad del efecto) usando cualquier ML como nuisance estimator (Random Forest, XGBoost, neural net). Inference válida (CI, p-value) sin asumir linearidad de los confounders.

Resultados de aprendizaje

Al finalizar, el estudiante podrá:

- Explicar Neyman-orthogonal score y por qué DML es doubly robust.
- Aplicar DoubleMLPLR para ATE con ML nuisance.
- Usar CausalForestDML de EconML para CATE personalizado.
- Cross-fitting: K-fold para evitar overfitting del nuisance.
- Inspeccionar policy óptimo: policy_tree para árbol de decisión de tratamiento.

Temas

- Marco PLR (Partially Linear Regression): $Y = \theta T + g(X) + \varepsilon$, $T = m(X) + v$.
- Score orthogonal: derivada respecto a nuisances = 0 en expectation.
- Cross-fitting: estimar nuisances en fold A, evaluar en B.
- CATE: efecto por subgroup.
- Heterogeneity tests.
- Policy learning: decidir a quién tratar.

Definiciones y características

- ATE (Average Treatment Effect): $E[Y(1) - Y(0)]$.
- CATE: $E[Y(1) - Y(0) \mid X=x]$.
- Nuisance functions: $g(X) = E[Y|X]$, $m(X) = E[T|X]$.
- Doubly robust: consistente si o g o m es correctamente especificado.
- Cross-fitting: K folds, estimar nuisance en train fold, evaluar score en test fold.
- CausalForestDML: random forest entrenado para minimizar heterogeneidad del effect (no error de predicción).

Dataset / recursos

- Sintético con efectos conocidos (de econml.tests.dgps).

- Lalonde 1986 (NSW training program) — clásico.
- IHDP (Infant Health and Development).
- Librerías: doubleml, econml, scikit-learn, xgboost.

Ejercicios

1. Sintético: simular Y = 2·T + 3·X1 + X2² + ε, T = P(...|X). ATE verdadero = 2.
2. DML básico: DoubleMLPLR(data, ml_g=RF, ml_m=RF, n_folds=5). Reportar θ̂.
3. Comparar OLS vs DML: OLS lineal con Y ~ T + X1 + X2 sesgado por X2 no lineal. DML lo recupera.
4. CausalForestDML: con dataset heterogéneo, estimar CATE. Mapa por X1.
5. Policy tree: econml.policy.PolicyTree para decidir a quién tratar.

Homework verificable

Estimar efecto causal sobre dataset con confounders no lineales:

1. Generar dataset sintético (econml.dgps.ihdp_surface_B o custom).
2. Estimar ATE con: (a) diferencia ingenua, (b) OLS lineal, (c) DML RF, (d) DML XGBoost.
3. Reportar bias contra ground truth.
4. CATE con CausalForestDML; visualizar heterogeneidad.

Criterio de aceptación: DML recupera ATE con bias < 5 %; OLS lineal sesgado; CATE muestra variación por subgrupo.

Errores comunes

Síntoma / mensaje	Causa y cómo arreglar
DML con n_folds=2 inestable	Fix: 5-10 folds.
Confounders no observados → DML sesgado	DML asume unconfoundedness. Fix: si hay IV
ml_g muy expresivo overfittea	Cross-fitting ayuda pero no inmune. Fix: r
Reportar CATE sin IC	Necesario para inference. Fix: est.effect_
Asumir SUTVA cuando hay interferencia	Fix: ajustar diseño experimental.

Preguntas frecuentes

DML o regresión clásica?

Si confounders few + relación cerca-lineal: regresión OK. Si muchos / no lineales / interacciones: DML.

DoubleML vs EconML?

DoubleML: más académico, claro framework. EconML: más features (instrumental variables, dynamic treatment, policy). En la práctica complementarios.

Random Forest mejor que XGBoost como nuisance?

Suele dar similar. Probar ambos y elegir el mejor en out-of-fold prediction.

Causal Forest = Random Forest?

NO. Causal Forest divide para minimizar heterogeneidad del efecto causal, no error de Y.

Cuántos samples?

Mínimo 1000-5000 para CATE confiable. Para ATE, menos.

Referencias

- Chernozhukov, Chetverikov, Demirer et al. (2018), Double/Debiased Machine Learning, Econometrics Journal.
- Athey & Wager (2019), Estimating Treatment Effects with Causal Forests.
- Imbens (2020), Potential Outcome and Directed Acyclic Graph Approaches to Causality.
- DoubleML docs, EconML docs.

Siguiente clase

Clase 190 — Uplift modeling, DiD (difference-in-differences)

Apéndice: notebook (primer bloque)

El problema: con confounding observado high-dim, regresión naive sesga. Double/Debiased ML (Chernozhukov 2018) usa ML para residualizar Y y T, y aplica Frisch-Waugh-Lovell sobre los residuos → estimador eficiente con CI válido. Requiere: pip install numpy scikit-learn (opcional doubleml).

```
import numpy as np
from sklearn.linear_model import Ridge, LogisticRegression
from sklearn.model_selection import KFold

rng = np.random.default_rng(42)
n, p = 2000, 20
X = rng.normal(0, 1, (n, p))
beta = rng.normal(0, 1, p)
# Tratamiento depende de X (confounding)
logit = X @ beta * 0.3
T = (rng.uniform(0, 1, n) < 1/(1+np.exp(-logit))).astype(float)
# Outcome depende de X y T; ATE verdadero = 2.0
TRUE_ATE = 2.0
Y = X @ beta + TRUE_ATE * T + rng.normal(0, 1, n)
print(f'n={n}  p={p}  P(T=1)={T.mean():.3f}  TRUE ATE = {TRUE_ATE}')
```

Archivos complementarios

- notebook.ipynb