Clase 224 — Métricas de fairness: demographic parity, equalized odds, calibration

Parte: 7 — Ética, Fairness y Privacidad · Fuente: Barocas, Hardt, Narayanan — Fairness and Machine Learning cap. 3 + Hardt, Price, Srebro (NeurIPS 2016) Equality of Opportunity in Supervised Learning. Duración estimada: 75 min.

Clase 224 — Métricas de fairness: demographic parity, equalized odds, calibration

> *Parte: 7 — Ética, Fairness y Privacidad · Fuente: Barocas, Hardt, Narayanan — Fairness and Machine Learning cap. 3 + Hardt, Price, Srebro (NeurIPS 2016) Equality of Opportunity in Supervised Learning. Duración estimada: 75 min.*

Objetivo

Pasar de "el modelo es injusto" a medirlo con un número. Implementar las tres familias de métricas grupales que dominan la literatura — demographic parity, equalized odds, calibration — sobre un dataset binario con atributo protegido, y demostrar numéricamente el teorema de imposibilidad de Kleinberg-Mullainathan-Raghavan / Chouldechova (2017): salvo casos triviales, no se pueden satisfacer las tres a la vez.

Resultados de aprendizaje

Al finalizar, el estudiante podrá:

- Calcular demographic parity gap = $|P(\hat{Y}=1|A=0) - P(\hat{Y}=1|A=1)|$ sobre predicciones de cualquier clasificador binario.
- Calcular equal opportunity (TPR por grupo) y equalized odds (TPR y FPR por grupo) — Hardt, Price, Srebro 2016.
- Verificar calibración por grupo con reliability curves: $P(Y=1|\hat{S}=s, A=a)$ debe ser igual entre grupos para un mismo score s.
- Demostrar el teorema de imposibilidad: ajustar threshold por grupo para forzar demographic parity rompe calibración.
- Aplicar mitigación post-processing con thresholds por grupo (Hardt 2016) y reportar el trade-off accuracy vs fairness gap.

Temas

#	Tema	Por qué importa
1	Atributo protegido A y notación Y / Ŷ / Ŝ	Sin notación común no se discute fairness;
2	Demographic parity (statistical parity)	La métrica más antigua (regla del 80%, US
3	Equal opportunity y equalized odds (Hardt	Condicionan en Y — corrigen el defecto de
4	Calibration por grupo (Chouldechova 2017)	El score debe significar lo mismo en cada
5	Teorema de imposibilidad (KMR / Chouldecho	Si las base-rates difieren, DP + equalized
6	Post-processing: threshold por grupo	Mitigación más simple; revela explícitamen

Definiciones y características

- Atributo protegido A: variable demográfica sensible (sexo, raza, edad). En la práctica nunca está sola — hay proxies (zip code, nombre, historial).

- Demographic parity (DP): P(Ŷ=1 | A=0) = P(Ŷ=1 | A=1). Métrica: DP_gap = |selection_rate(A=0) − selection_rate(A=1)|. Regla del 80%: ratio ≥ 0.8 para que la US EEOC no lo considere discriminación.
- Equal opportunity: P(Ŷ=1 | Y=1, A=0) = P(Ŷ=1 | Y=1, A=1). O sea, TPR igual. Solo a los positivos reales se les exige misma tasa de aceptación.
- Equalized odds (Hardt-Price-Srebro 2016): equal opportunity + FPR igual. EO_gap = max(|TPR_diff|, |FPR_diff|).
- Calibration (predictive parity, Chouldechova 2017): P(Y=1 | Ŝ=s, A=a) es igual entre grupos para todo score s. Un score 0.7 significa "70% chance" tanto para A=0 como para A=1.
- Impossibility theorem (Kleinberg-Mullainathan-Raghavan 2017 + Chouldechova 2017): si P(Y=1|A=0) ≠ P(Y=1|A=1) (base rates diferentes) y el clasificador no es perfecto, no existe clasificador que sea simultáneamente calibrado por grupo y con equalized odds. Es matemático, no técnico — no se resuelve con más datos.
- Accuracy-fairness trade-off: forzar cualquier paridad sobre un clasificador óptimo de Bayes baja accuracy. La pregunta política es cuánta accuracy estamos dispuestos a sacrificar.
- Tooling: fairlearn (Microsoft) — MetricFrame, ThresholdOptimizer. aif360 (IBM) — 70+ métricas y mitigadores. Ambas open source.

Dataset / recursos

- Dataset notebook: sintético binario con un atributo protegido A{0,1} y base rates diferentes (60% vs 40%) — necesario para que el teorema de imposibilidad se active.
- Dataset real recomendado para tarea: Adult / Census Income (UCI) con sex como atributo protegido, o COMPAS (ProPublica) con race.
- Librerías: numpy, pandas, scikit-learn. Opcional: fairlearn.

Ejercicios

1. Selection rate por grupo: entrenar LogisticRegression baseline. Calcular P(Ŷ=1|A=0) y P(Ŷ=1|A=1) y el DP_gap. ¿Cumple regla del 80%?
2. TPR y FPR por grupo: armar confusion_matrix separada por grupo. Calcular equal_opportunity_gap = |TPR_0 − TPR_1| y equalized_odds_gap = max(|TPR_diff|, |FPR_diff|).
3. Calibration curves por grupo: binning de scores en 10 bins. Para cada bin y cada grupo, graficar mean(y_true) vs mean(y_score). ¿Las curvas coinciden?
4. Romper calibración: ajustar threshold por grupo (t_0, t_1) tal que se cumpla DP exacta. Recalcular calibración — debe degradarse. (Demostración numérica del teorema.)
5. Post-processing Hardt: buscar (t_0, t_1) que minimicen equalized_odds_gap y reportar el costo en accuracy global. Tabla: baseline vs DP-fixed vs EO-fixed.

Homework verificable

Notebook con:

1. Cargar Adult Census (UCI). Atributo protegido: sex. Target: income > 50K.
2. Entrenar baseline LogisticRegression + reportar accuracy, AUC.
3. Calcular las tres métricas: DP_gap, equal_opportunity_gap, equalized_odds_gap y calibration_gap (max |calibration(A=0) − calibration(A=1)| sobre bins).
4. Implementar post-processing con threshold por grupo que minimice EO_gap sujeto a accuracy_drop ≤ 3pp.

5. Tabla final: 3 modelos (baseline, DP-mitigated, EO-mitigated) × 5 métricas (accuracy, AUC, DP_gap, EO_gap, calibration_gap). Discutir cuál elegirías y por qué — no hay respuesta universal.

Criterio de aceptación: las 4 métricas implementadas a mano (no llamando a fairlearn), el EO_gap mitigado < 0.05, y un párrafo justificando la elección de métrica en función del dominio de aplicación (crédito vs admisión universitaria vs medicina).

Errores comunes

Síntoma / mensaje	Causa y cómo arreglar
DP_gap = 0 pero el modelo es claramente in	DP ignora el ground truth — si las base ra
Calibration "perfecta" pero TPR diferente	Es el resultado natural cuando las base ra
Quitar el atributo protegido del input "so	Fairness through unawareness — no funciona
Threshold único 0.5 para todos los grupos	Asume que los scores significan lo mismo e
Usar accuracy global para comparar fairnes	Accuracy global puede subir y empeorar al
MetricFrame de fairlearn devuelve nan	Algún grupo tiene 0 positivos predichos o

Preguntas frecuentes

¿Cuál de las tres métricas uso?

Depende del dominio. Crédito o contratación: equal opportunity (no negarle empleo a quien sí calificaría). Justicia penal / riesgo de reincidencia: calibration + FPR equal (el debate ProPublica vs COMPAS giró exactamente sobre cuál priorizar). Publicidad: DP suele alcanzar. Pero la elección es política, no técnica.

¿Demographic parity no es siempre lo que queremos?

No. Si la base rate real difiere entre grupos (ej. distribución de ingresos), DP puede forzar al modelo a aceptar candidatos peores de un grupo y rechazar mejores de otro. Equal opportunity suele ser más defendible.

¿El teorema de imposibilidad significa que fairness es imposible?

No — significa que no se pueden satisfacer las tres simultáneamente cuando las base rates difieren. Hay que elegir cuál priorizar, y documentarlo. El paper de Chouldechova (2017) es lectura obligada.

¿fairlearn o aif360?

fairlearn para empezar (API más limpia, integrada con sklearn). aif360 cuando se necesite catálogo amplio de mitigadores (pre/in/post-processing) y métricas exóticas. Ninguna sustituye entender las definiciones.

¿Y la fairness individual?

Otra familia (Dwork et al. 2012): "individuos similares deben recibir predicciones similares". Más difícil de operacionalizar (requiere métrica de similaridad justificable) y queda fuera del alcance de esta clase. Ver Clase 225-227 para privacy y Clase 228 para causal fairness.

Referencias

- Hardt, M., Price, E., Srebro, N. Equality of Opportunity in Supervised Learning, NeurIPS 2016. https://arxiv.org/abs/1610.02413
- Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction

instruments, Big Data 2017. https://arxiv.org/abs/1610.07524
- Kleinberg, J., Mullainathan, S., Raghavan, M. Inherent Trade-Offs in the Fair Determination of Risk Scores, ITCS 2017. https://arxiv.org/abs/1609.05807
- Barocas, S., Hardt, M., Narayanan, A. Fairness and Machine Learning (fairmlbook.org), cap. 3 — Classification.
- fairlearn documentation — Microsoft, MIT license.
- AIF360 documentation — IBM, Apache 2.0.

Siguiente clase

Clase 225 — Privacidad diferencial: intro

Apéndice: notebook (primer bloque)

Dataset sintético binario con atributo protegido A{0,1} y base rates diferentes (necesario para activar el teorema de imposibilidad). Requiere: pip install numpy pandas scikit-learn matplotlib. fairlearn opcional — implementamos todo a mano.

```
import numpy as np, pandas as pd
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, roc_auc_score, confusion_matrix
import matplotlib.pyplot as plt

rng = np.random.default_rng(42)
n = 10_000

# Atributo protegido (50/50)
A = rng.integers(0, 2, n)

# Features con distribución que depende de A (proxy realista)
x1 = rng.normal(loc=0.5 * A, scale=1.0, size=n)
x2 = rng.normal(loc=-0.3 * A, scale=1.0, size=n)
x3 = rng.normal(loc=0.0, scale=1.0, size=n)

# Base rates DIFERENTES → activa impossibility
logits = 1.2  x1 - 0.8  x2 + 0.5 * x3 + np.where(A == 0, 0.4, -0.4)
p = 1 / (1 + np.exp(-logits))
y = (rng.random(n) < p).astype(int)

X = np.column_stack([x1, x2, x3, A])
print(f'n={n} | base rate A=0: {y[A==0].mean():.3f} | base rate A=1: {y[A==1].mean():.3f}')
```

Archivos complementarios

- notebook.ipynb