Clase 076 — Calibración de probabilidades: Platt, isotonic, temperature scaling

Parte: 1 — Machine Learning Clásico · Fuente: Platt (1999) + Niculescu-Mizil & Caruana (2005) + Guo et al. (2017). Duración estimada: 75 min.

Clase 076 — Calibración de probabilidades: Platt, isotonic, temperature scaling

> *Parte: 1 — Machine Learning Clásico · Fuente: Platt (1999) + Niculescu-Mizil & Caruana (2005) + Guo et al. (2017). Duración estimada: 75 min.*

Objetivo

Saber cuándo las probabilidades que devuelve predict_proba son calibradas — es decir, si el modelo dice "70 %" para un grupo, ¿realmente el 70 % es positivo? Modelos como Random Forest y SVM suelen estar mal calibrados; XGBoost mejor. Aplicar Platt scaling (sigmoid) y isotonic regression para corregir. Evaluar con Brier score, ECE (Expected Calibration Error) y reliability diagrams.

Resultados de aprendizaje

Al finalizar, el estudiante podrá:

- Generar un reliability diagram: agrupar predicciones por bin de probabilidad y plotear "predicho vs real".
- Calcular Brier score = mean((p - y)²). Más bajo = mejor calibración.
- Calcular ECE: Σ (n_b/N) · |acc_b - conf_b|.
- Aplicar sklearn.calibration.CalibratedClassifierCV(estimator, method='sigmoid' | 'isotonic', cv=5).
- Decidir: Platt cuando muestra chica (n < 1000), isotonic cuando hay datos.

Temas

- ¿Por qué importa? Decisiones que dependen de threshold ≠ 0.5 requieren probs reales.
- Reliability diagram: predicción vs frecuencia real.
- Platt scaling: ajustar σ(A·logit + B) con MLE sobre val set.
- Isotonic regression: monótono pero más flexible (puede sobreajustar).
- Temperature scaling: solo divide logits por T aprendido — para multiclase, eficiente.
- Modelos típicamente calibrados (logistic regression) vs no (RF, SVM).

Definiciones y características

- Calibración: P(y=1 | p̂=p) = p para todo p.
- Reliability diagram: histograma de calibración.
- Brier score: (1/N) Σ (p_i - y_i)². Combina calibración + accuracy.
- ECE: weighted gap entre confidence y accuracy por bin.
- Platt: sigmoid-fit; parametric, 2 params (A, B). Bueno con n chico.
- Isotonic: monotonic non-parametric; flexible pero needs más data.
- Temperature scaling: softmax(z / T) con T learnable. Para multiclase.

Dataset / recursos

- fetch_openml('credit-g') o load_breast_cancer.

- Librerías: sklearn.calibration, matplotlib.

Ejercicios

1. Reliability diagram: entrenar RandomForest, generar predict_proba, bin en 10 grupos, plotear curva calibration vs y=x. Suele desviar.
2. Brier + ECE: implementar ambas y comparar entre RF (mala calibración) y LogReg (buena).
3. CalibratedClassifierCV: CalibratedClassifierCV(RF, method='sigmoid', cv=5).fit(X, y). Re-evaluar Brier.
4. Isotonic vs Platt: comparar ambos sobre el mismo modelo. Con n=10_000 ambos similares; con n=300 Platt mejor.
5. Threshold tuning post-calibración: con probs calibradas, F1 vs threshold es más interpretable.

Homework verificable

Sobre credit-g:

1. RandomForest + reliability diagram + Brier + ECE.
2. Calibrar con Platt y con isotonic (CV).
3. Comparar Brier/ECE pre y post.
4. Reliability diagrams superpuestos.

Criterio de aceptación: calibración reduce ECE en ≥ 30 %; reliability diagram post se acerca a la diagonal.

Errores comunes

Síntoma / mensaje	Causa y cómo arreglar
Calibrar sobre train (overfit)	Leak. Fix: usar CV o held-out calibration
Isotonic con < 500 ejemplos por clase	Overfit. Fix: Platt.
Reportar accuracy post-calibración como mé	No mide calibración. Fix: Brier o ECE.
Asumir LogReg está calibrado siempre	Con regularización fuerte o features extra
Calibración en multiclase con sigmoid	Sigmoid es binario. Fix: temperature scali

Preguntas frecuentes

¿Calibración cambia accuracy?

No (mucho). Reordena las probs pero no cambia el argmax típicamente. La accuracy queda igual; las probs son más interpretables.

¿Cuándo importa?

Cuando hacés decisiones threshold ≠ 0.5 (e.g., "alertar si P > 0.8"), reportes de "confianza" al usuario, ensemble por probabilidades.

¿RF tan mal calibrado?

Sí, hacia probs intermedias (0.2-0.8). Boosting (XGBoost) suele estar mejor.

¿DL necesita calibración?

Sí — DL modernos tienden a sobre-confiar. Temperature scaling (Guo 2017) es el estándar.

¿En producción cómo monitoreo calibración?

Logueá (prob_predicha, y_real). Cada N días, calculá Brier y ECE. Si drift, re-calibrar.

Referencias

- Platt (1999), Probabilistic Outputs for Support Vector Machines.
- Niculescu-Mizil & Caruana (2005), Predicting Good Probabilities With Supervised Learning, ICML.
- Guo et al. (2017), On Calibration of Modern Neural Networks, ICML.
- sklearn calibration docs.

Siguiente clase

Clase 077 — SVM lineal

Apéndice: notebook (primer bloque)

GBM out-of-the-box suele estar mal calibrado. Reliability curve + Brier + ECE, después CalibratedClassifierCV con sigmoid e isotonic.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.datasets import make_classification
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.model_selection import train_test_split
from sklearn.calibration import CalibratedClassifierCV
from sklearn.metrics import brier_score_loss, log_loss

np.random.seed(42)
```

Archivos complementarios

- notebook.ipynb