

Sobreajuste, Sesgo y Técnicas de Control

Minería de Datos (CC3074) - 2026

Sobreajuste, Sesgo y Técnicas de Control

Semestre 01, 2026

Problemática

Un modelo puede ajustarse bien para los datos de prueba pero fallar en datos nuevos.

El objetivo no es un ajuste perfecto, sino generalizar.

Sobreajuste (Overfitting)

- El modelo aprende demasiado bien los datos de entrenamiento
- Captura ruido además de señal
- Tiene bajo error en train
- Alto error en test

Memoriza en lugar de aprender ☹️.

Ejemplo

Modelo lineal simple vs polinomio de grado 10.

El polinomio:

- Ajusta todos los puntos
- Pero oscila violentamente
- Mala generalización

Síntomas

- Error de entrenamiento muy bajo
- Error de prueba alto
- Modelo muy complejo
- Demasiados parámetros

Subajuste (Underfitting)

- El modelo es demasiado simple
- No captura la estructura real
- Alto error en train
- Alto error en test

No aprende lo suficiente $\uparrow \sigma^2$.

Ejemplo

Intentar ajustar una recta cuando la relación real es cuadrática.

La recta siempre falla.

Sesgo-Varianza

Error total = Sesgo² + Varianza + Error irreducible

Siempre hay un compromiso.

Sesgo

- Error por simplificación excesiva
- Modelo demasiado rígido
- Subajuste

Alta simplificación → alto sesgo

Varianza

- Sensibilidad a pequeñas variaciones en los datos
- Modelo demasiado flexible
- Sobreajuste

Alta complejidad → alta varianza

Generalización

Modelo simple → alto sesgo, baja varianza

Modelo complejo → bajo sesgo, alta varianza

El punto óptimo está en equilibrio.

Validación Cruzada (k-fold)

- Dividir datos en k partes

- Entrenar k veces
- Promediar desempeño

Reduce dependencia de una sola partición.

Ventajas

- Mejor estimación del error real
- Reduce varianza en evaluación
- Ideal con datasets pequeños

Bootstrap

- Muestreo con reemplazo
- Genera múltiples datasets artificiales
- Estima distribución del estimador

Usos

- Estimar incertidumbre
- Calcular intervalos de confianza
- Medir estabilidad del modelo

Diferencia CV vs Bootstrap

Validación Cruzada:

- Evalúa desempeño predictivo

Bootstrap:

- Evalúa estabilidad e incertidumbre

Ambos ayudan a controlar sobreajuste.

Estrategias adicionales

- Reducir complejidad del modelo
- Regularización (Ridge, Lasso)
- Más datos
- Selección de variables