

L U M B R E



Sobreajuste y subajuste

Grado: 5

Materia: Fundamentos de Machine Learning · Machine Learning

11/10/2026

Antes de empezar

Todo modelo falla de una de dos formas: o es demasiado tonto para capturar el patrón, o tan listo que memoriza el ruido. Aprender a distinguir subajuste de sobreajuste con dos curvas es el gesto diagnóstico que salva proyectos enteros.

QUÉ APRENDERÁS

- Diagnosticar el sobreajuste y el subajuste comparando el error de entrenamiento y de validación, y elegir la complejidad, los datos o la regularización adecuados.

Sobreajuste y subajuste

Dos fracasos, dos diagnósticos

Un modelo subajustado es un estudiante que no estudió: falla igual en el examen y en la vida. Un modelo sobreajustado es quien memorizó las respuestas del simulacro: brilla en casa y se cae en el real. Los dos erran, pero con síntomas opuestos.

VIDEO

Sobreajuste y subajuste en machine learning

Vídeo en español que muestra con gráficos cómo se ven el sobreajuste y el subajuste, y por qué sus síntomas son opuestos.



Escanea para verlo en Lumbre

https://www.youtube.com/watch?v=NLp2yhoR_HE

Subajuste: demasiado simple

El subajuste ocurre cuando el modelo no tiene capacidad suficiente para capturar la estructura real. Ajusta una recta a datos con forma de curva. Falla tanto en

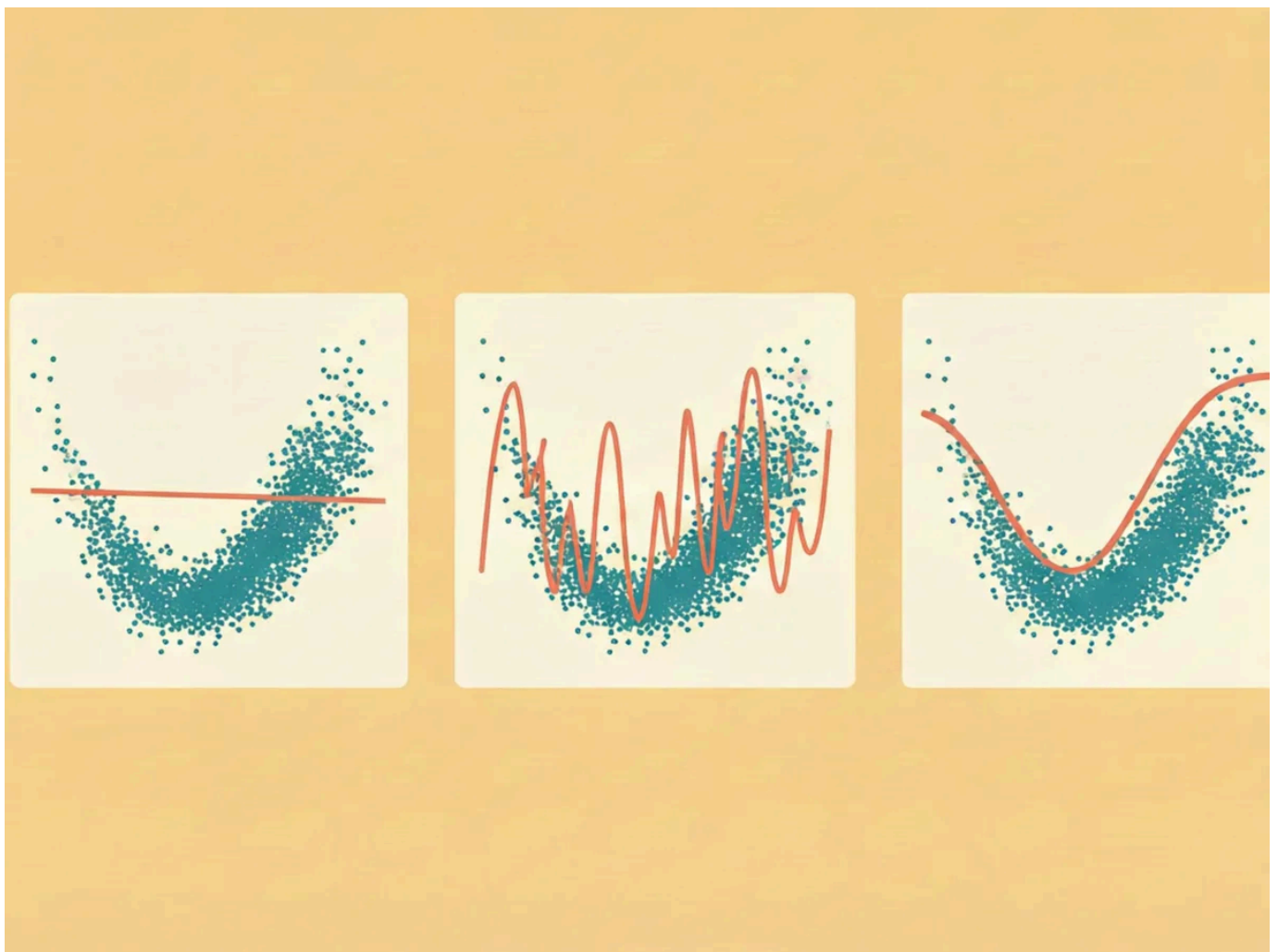
entrenamiento como en validación: su error es alto por doquier. La solución pasa por un modelo más expresivo o mejores características.

Sobreajuste: demasiado flexible

El sobreajuste aparece cuando el modelo se adapta a los datos de entrenamiento con su ruido incluido. Acierta casi perfecto en entrenamiento pero se desploma en validación. La señal y el azar quedaron indistinguibles para él; memorizó en vez de aprender.

La distancia entre errores

La firma del sobreajuste es la brecha entre un error de entrenamiento bajo y uno de validación alto. Si ambos son altos y parecidos, hay subajuste. Mirar la distancia entre las dos curvas —no solo su altura— es el diagnóstico esencial.



Los tres diagnósticos en una imagen: subajuste (la curva ignora la forma real), sobreajuste (persigue hasta el ruido) y ajuste equilibrado (captura la tendencia sin memorizar los detalles).

Complejidad y U del error

Al aumentar la complejidad, el error de entrenamiento cae monótonamente, pero el de validación desciende hasta un óptimo y luego vuelve a subir: una U. El punto mínimo de la curva de validación marca el equilibrio deseado entre subajustar y sobreajustar.

El papel de los datos

El sobreajuste se combate sobre todo con más datos: con suficientes ejemplos, el ruido promedio se cancela y resulta más difícil memorizarlo. Añadir experiencia vale a menudo más que retocar el modelo, y es la cura más subestimada.

Regularizar

La regularización penaliza la complejidad —coeficientes grandes, árboles profundos—, frenando el sobreajuste sin reducir el modelo. Es un dial entre ajustar los datos vistos y mantener la solución simple y generalizable. Ridge y lasso son sus formas clásicas.

Curvas de aprendizaje

Graficar error frente al tamaño del conjunto de entrenamiento revela el diagnóstico: si ambas curvas están juntas y altas, hace falta un modelo mejor; si la de validación sube hacia la de entrenamiento al añadir datos, ganarás con más ejemplos.

Podar y early stopping

En árboles, podar recorta ramas que solo capturan ruido. En métodos iterativos, detener el entrenamiento cuando la validación deja de mejorar —early stopping— corta el sobreajuste de raíz. Son regularizaciones disfrazadas de criterios de parada.

Errores frecuentes

- Alegrarse por un error de entrenamiento próximo a cero ignorando la validación.
- Añadir más complejidad cuando el problema era de subajuste por datos pobres en características.

Ejemplo resuelto: curar un modelo sobreajustado

Un clasificador marca casi cero por ciento de error al entrenar y treinta y cinco al validar. El protocolo de cura:

1. Confirmar el diagnóstico: brecha grande entre entrenamiento y validación, señal de sobreajuste.

2. Añadir datos: es la cura más eficaz, el ruido de cada muestra se promedia y deja de memorizarse.
3. Regularizar: penalizar coeficientes grandes o podar el árbol para frenar la flexibilidad.
4. Reducir complejidad si sigue: probar una familia más simple y comparar de nuevo las dos curvas.
5. Releer las curvas tras cada cambio: la meta es achicar la brecha sin disparar el error de validación.

¿Para qué sirve en la realidad?

El costo de no diagnosticar bien es enorme: se despliega un modelo que brilló en el laboratorio y falla en producción, o se descarta un buen enfoque creyendo que hacía falta más inteligencia donde solo hacían falta más datos. Leer dos curvas ahorra meses de esfuerzo mal dirigido.

1. Un modelo tiene un error de entrenamiento muy bajo pero uno de validación muy alto. El diagnóstico más probable es:

- ☐ Subajuste: el modelo es demasiado simple
- ☐ Sobreajuste: memorizó el ruido del entrenamiento
- ☐ Falta de variables categóricas
- ☐ Un error de cálculo en la pérdida

2. Con suficiente complejidad de modelo, el error de validación seguirá bajando sin límite.

- ☐ Verdadero
- ☐ Falso

El juicio del practicante

No existe un ajuste perfecto: se elige un punto en la frontera entre fiabilidad y flexibilidad. Un buen científico de datos tolera un ligero sesgo para ganar estabilidad, consciente de que predecir un poco peor en media, pero de forma robusta, casi siempre vale más.

ACTIVIDAD INTERACTIVA DISPONIBLE EN LUMBRE

3. Ajuste



4. Ordena el protocolo de diagnóstico por las curvas de error:

Elementos:

- entrenar un modelo y medir su error en ambos conjuntos
- leer la altura de la curva de entrenamiento
- leer la distancia hasta la curva de validación
- actuar: más modelo si subajuste, más datos o regularización si sobreajuste

Entrenar	. . .
Comparar	. . .
Decidir	. . .

Explained <https://explained.ai/Representaciones intuitivas del equilibrio sesgo-varianza y el sobreajuste>.

5. Une cada síntoma de las curvas con su diagnóstico o su cura.

5.	error alto en entrenamiento y en validación	-----
5.	entrenamiento casi perfecto, validación pobre	-----
5.	penalizar la complejidad del modelo	-----
5.	detener el entrenamiento cuando la validación deja de mejorar	-----

Opciones: subajuste · sobreajuste · regularización · early stopping

Dos fracasos

Sobre y subajuste

Subajuste

modelo rígido

error alto por doquier

Sobreajuste

memoriza el ruido

gran brecha

Curas

más datos

regularizar

podar

early stopping

Diagnóstico

altura de curvas

distancia entre ellas

El mapa mental interactivo es interactivo en la versión online.



Overfitting vs. underfitting <https://www.ibm.com/mx-es/think/topics/overfitting-vs-underfitting> Explicación de referencia en español sobre ambos modos de fracaso y cómo corregirlos.

6. Tu validación mejora al añadir datos, pero el error de entrenamiento es ya bajísimo. ¿Tienes sobreajuste o subajuste? ¿Invertirías en más modelo o en más datos?

Referencia curricular

Grado:	5
Materia:	Fundamentos de Machine Learning · Machine Learning

Versión interactiva completa (con actividades, videos y autoevaluación):



<https://univ.lumbre.lat/object/cdia-mlf-004-sobreajuste-y-subajuste>

Material educativo generado desde Lumbre · lumbre.lat