Aprendizaje de

Máquina Supervisado

Sesión 8

¿Qué es el Boosting?

1 Concepto Fundamental

El **boosting** es una técnica de aprendizaje automático perteneciente a la familia de métodos de **ensemble learning**. Su idea central es combinar múltiples modelos débiles para crear un modelo fuerte y más preciso.

2 Entrenamiento Secuencial

A diferencia del **bagging**, donde los modelos se entrenan de manera independiente y en paralelo, en el boosting los modelos se entrenan de forma **secuencial**, con cada nuevo modelo intentando corregir los errores cometidos por los anteriores.

3 Enfoque en Errores

El boosting asigna mayor peso a las instancias mal clasificadas, permitiendo que los nuevos modelos se concentren en corregir estos errores específicos, mejorando gradualmente la precisión del conjunto.

Características Clave del Boosting

Modelos Débiles

Utiliza modelos con rendimiento ligeramente mejor que el azar (51% vs 50%). Suelen ser simples, como árboles de decisión con pocas divisiones (stumps).

Entrenamiento Secuencial

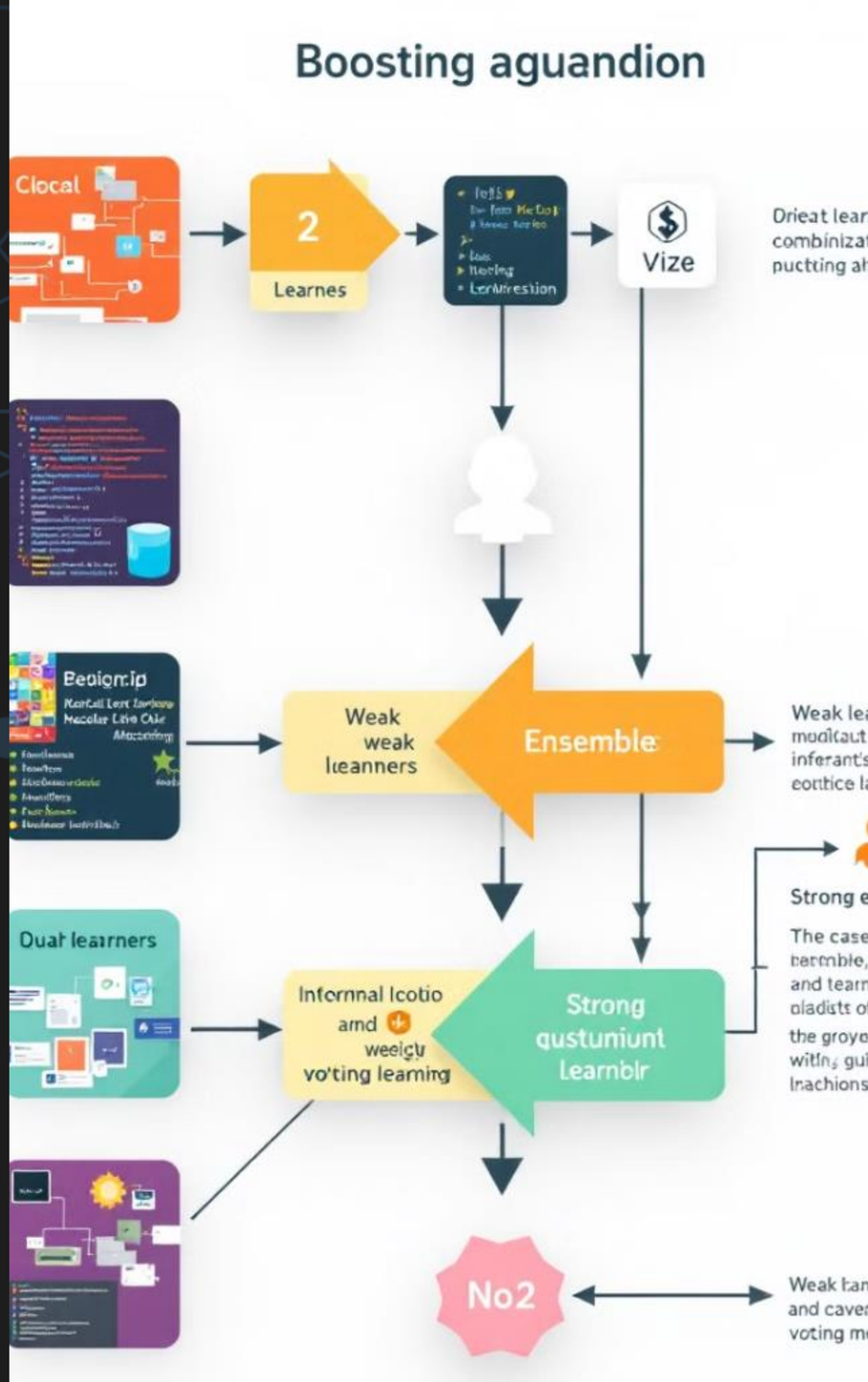
Los modelos se entrenan uno tras otro, enfocándose en las instancias mal clasificadas por los modelos anteriores mediante la asignación de mayor peso a estas instancias.

Combinación de Modelos

Las predicciones se combinan mediante votación ponderada (clasificación) o suma ponderada (regresión), dando mayor peso a los modelos con mejor rendimiento.

Reducción del Sesgo

El objetivo principal es reducir el **sesgo** (bias) del modelo, corrigiendo los errores que se producen debido a suposiciones simplistas, evitando así el subajuste (underfitting).



Ensemble Learning, Bagging y Boosting

Ensemble Learning

Técnica que combina múltiples modelos para mejorar el rendimiento general. La diversidad de modelos es clave, ya que deben cometer errores diferentes para que la combinación sea efectiva.

Las predicciones se combinan mediante promedio (regresión), votación mayoritaria (clasificación) o ponderación, reduciendo así la varianza o el sesgo según la técnica utilizada.

Bagging

Utiliza muestreo con reemplazo (bootstrap) para generar múltiples subconjuntos de datos. Cada modelo se entrena de manera independiente en uno de estos subconjuntos.

Las predicciones se combinan mediante promedio o votación mayoritaria. Es especialmente útil para reducir la **varianza** y evitar el sobreajuste (overfitting).

Boosting

Entrena modelos secuencialmente, asignando mayor peso a las instancias mal clasificadas. Cada nuevo modelo se enfoca en corregir los errores de los anteriores.

Las predicciones se combinan mediante votación o suma ponderada. Es especialmente útil para reducir el **sesgo** y evitar el subajuste (underfitting).

Comparativa: Bagging vs Boosting

Característica	Bagging	Boosting
Entrenamiento	Modelos entrenados en paralelo	Modelos entrenados secuencialmente
Enfoque	Reduce la varianza (overfitting)	Reduce el sesgo (underfitting)
Pesos en instancias	No se asignan pesos	Se asignan pesos a instancias
Combinación	Promedio o votación mayoritaria	Votación ponderada o suma ponderada
Ejemplo de algoritmo	Random Forest	Gradient Boosting, AdaBoost
Coste computacional	Menor (paralelización)	Mayor (entrenamiento secuencial)
Interpretabilidad	Menos interpretable	Menos interpretable

El Algoritmo Gradient Boosting

Concepto Fundamental

Gradient Boosting construye un modelo fuerte de manera secuencial, donde cada nuevo modelo intenta corregir los errores (residuos) de los modelos anteriores utilizando el descenso de gradiente para minimizar una función de pérdida.

Proceso de Entrenamiento

Se inicia con un modelo simple que predice un valor constante. En cada iteración, se calculan los residuos entre predicciones y valores reales, se entrena un nuevo modelo para predecir estos residuos, y se actualizan las predicciones.

Función de Pérdida

Utiliza funciones como Error Cuadrático Medio (MSE) para regresión o Log Loss para clasificación. El descenso de gradiente minimiza esta función en cada iteración, ajustando el modelo para reducir el error.

Predicción Final

La predicción final es la suma de las predicciones de todos los modelos entrenados secuencialmente, ponderadas por el factor de aprendizaje (learning rate).

Ventajas del Gradient Boosting

Alta Precisión

Produce modelos muy precisos, especialmente en problemas de regresión y clasificación. Al combinar múltiples modelos débiles secuencialmente, captura patrones complejos que otros métodos podrían pasar por alto.

Flexibilidad

Maneja diferentes tipos de datos (numéricos, categóricos, mixtos) y es compatible con diversas funciones de pérdida, adaptándose a una amplia variedad de problemas como regresión, clasificación y ranking.

Manejo de Relaciones No Lineales

Modela eficazmente relaciones no lineales entre variables de entrada y objetivo, siendo adecuado para problemas donde la relación no es lineal, superando las limitaciones de modelos más simples.

Desventajas del Gradient Boosting

Riesgo de Sobreajuste

Uno de los principales riesgos es el sobreajuste, especialmente con demasiadas iteraciones o un learning rate alto. El modelo puede ajustarse excesivamente a los datos de entrenamiento y perder capacidad de generalización.

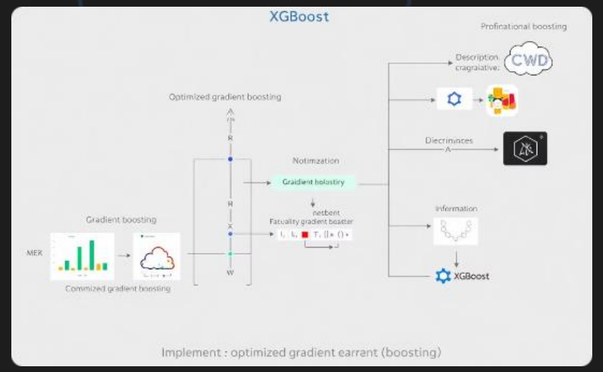
Coste Computacional

El entrenamiento secuencial de múltiples modelos puede ser lento y costoso en términos de recursos, especialmente con grandes conjuntos de datos, siendo más lento que métodos como Random Forest.

Dificultad de Interpretación

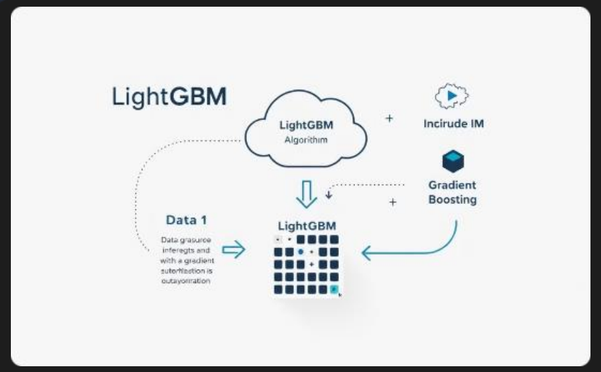
Aunque los modelos base son interpretables, el modelo final es una combinación de muchos modelos, dificultando su interpretación, lo cual puede ser problemático en aplicaciones donde la transparencia es crucial.

Implementaciones Populares



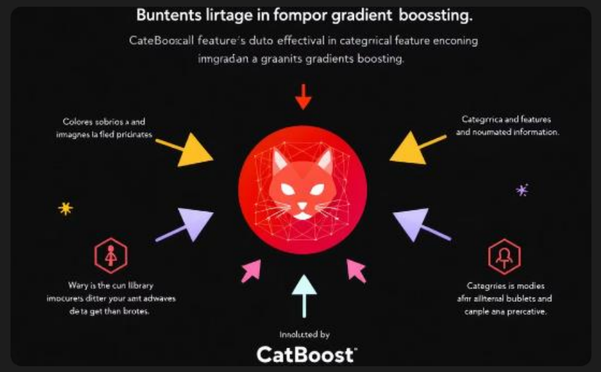
XGBoost

Implementación optimizada que incluye técnicas adicionales como regularización y manejo eficiente de datos faltantes. Destaca por su velocidad y precisión, siendo ampliamente utilizada en competiciones de ciencia de datos.



LightGBM

Diseñado para ser más eficiente en términos de memoria y tiempo de entrenamiento. Utiliza técnicas como "Gradient-based One-Side Sampling" (GOSS) para acelerar significativamente el proceso de entrenamiento.



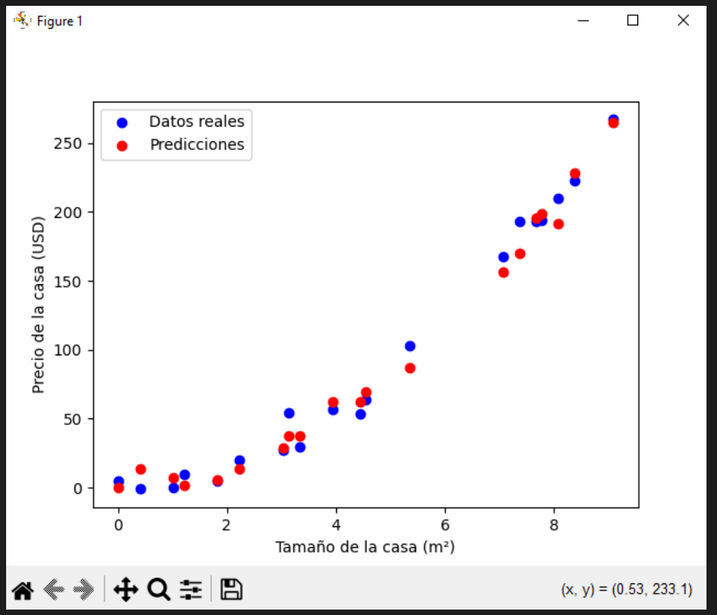
CatBoost

Especialmente diseñado para manejar datos categóricos de manera eficiente. Incluye técnicas avanzadas para evitar el sobreajuste y mejorar la generalización, con un rendimiento destacado en datos con variables categóricas.

Actividad Práctica Guiada

Requisitos:

1. Importa las librearías.
2. Generar datos sintéticos
3. Dividir los datos en conjuntos de entrenamiento y prueba
4. Crear y entrenar el modelo de Gradient Boosting
5. Evaluar el modelo
6. Visualizar las predicciones
7. Mejorar el modelo



El detalle de la actividad se encuentra en guía de estudio de la sesión

Preguntas

Sección de preguntas

Aprendizaje de

Máquina Supervisado

Continúe con las actividades