Equipo 3

Boosting Algorithms

Introducción

Definición

Concepto

- Técnica de ensemble
- Para tareas de clasificación y regresión.

Cómo funciona

- Los modelos se entrenan secuencialmente.
- Cada modelo intenta corregir los errores del modelo anterior.
- Se asignan pesos a las observaciones: mal clasificadas , mayor peso.

Introducción

Contexto Historico

Origen

- 1990 por Robert Schapire.

Importancia

- Usado en: detección de fraude, diagnóstico médico y análisis financiero.
- Dió lugar a XGBoost, LightGBM y CatBoost.

Relevancia

- Manejo de datos complejos
- Flexible.

Motivación y Fundamentos

Problemas que resuelve

- **Overfitting:** Modelos muy complejos que ajustan demasiado los datos de entrenamiento.
- **Underfitting:** Modelos que son demasiado simples y no capturan la complejidad del problema.
- **Imbalance de datos:** Puede mejorar la clasificación en conjuntos de datos desbalanceados, donde una clase es dominante.
- **Errores residuales:** Al entrenar sucesivamente modelos en los errores de predicciones anteriores, Boosting reduce el error general.

Motivación y Fundamentos

Concepto de aprendices débiles y fuertes

- **Aprendices débiles:** Son modelos simples que tienen un rendimiento apenas mejor que el azar (por ejemplo, un árbol de decisión muy poco profundo).
  - **Característica clave:** Pueden cometer errores, pero son eficientes y fáciles de ajustar.

- **Aprendices fuertes:** Son combinaciones de muchos aprendices débiles para crear un modelo más potente y preciso.
  - Boosting convierte un conjunto de aprendices débiles en un aprendiz fuerte, mejorando sustancialmente la precisión.

Motivación y Fundamentos

Cómo el Boosting combina modelos simples para mejorar el rendimiento

- Boosting trabaja construyendo modelos secuenciales, donde cada nuevo modelo trata de corregir los errores de los modelos anteriores.

- **Ejemplo:** El algoritmo AdaBoost asigna más peso a los ejemplos mal clasificados, enfocándose en ellos en cada iteración.

- **Ventaja:** Al combinar muchos modelos simples, Boosting puede producir un modelo final que es más robusto y preciso que los modelos individuales, sin sobreajustar los datos.

Concepto de Gradient Boosting

Gradient Boosting es una técnica de aprendizaje supervisado que crea un modelo fuerte combinando múltiples modelos débiles (usualmente de árboles de decisión). Y en cada paso, un nuevo modelo se entrena para corregir los errores residuales del modelo anterior, minimizando la función de pérdida mediante el gradiente descendente.

Diferencias clave con AdaBoost

- **Manejo de errores**: GBM ajusta los errores residuales usando gradiente descendente; AdaBoost ajusta las ponderaciones de los datos mal clasificados.

- **Función de pérdida**: GBM minimiza una función de pérdida diferenciable (como el error cuadrático), mientras que AdaBoost utiliza una función basada en errores de clasificación.

- **Flexibilidad**: GBM puede optimizar diferentes funciones de pérdida, AdaBoost está enfocado en clasificación binaria.

Aplicaciones comunes y ventajas

- Aplicaciones: Detección de fraudes, predicción de series temporales, clasificación y regresión en datos tabulares.

- Ventajas: Alta precisión, flexibilidad para manejar diferentes funciones de pérdida y capacidad de reducir el sobreajuste si se regulariza adecuadamente.

Variantes modernas

XGBoost, conocido por su alta precisión y escalabilidad, su capacidad para manejar datos dispersos y ofrecer control fino a través de la regularización lo hace muy popular, aunque puede ser más lento y complejo en datos grandes. Es ideal para predicción de riesgo crediticio, segmentación de clientes, biomedicina, mercados financieros, etc.

LightGBM es un algoritmo basado en histogramas que destaca por su velocidad de entrenamiento y eficiencia en grandes conjuntos de datos. Su enfoque leaf-wise lo hace muy rápido y eficiente en memoria, pero es más propenso al sobreajuste y sensible en datos pequeños. Este consiste en expandir siempre la hoja (nodo) del árbol que reduce más el error, en lugar de expandir los nodos de todos los niveles de manera equilibrada (level-wise, que es más común).

CatBoost es un algoritmo de boosting diseñado para manejar nativamente características categóricas, lo que simplifica el preprocesamiento. Es robusto contra el sobreajuste y funciona bien con datos desbalanceados, aunque puede ser menos conocido que otros algoritmos. Es usado normalmente en riesgo crediticio y seguros.

Ventajas y Limitaciones del Boosting

- Incrementa la precisión al enfocarse en corregir errores secuenciales generando modelos altamente ajustados a los datos.

- Mayor susceptibilidad al sobreajuste en conjuntos de datos pequeños o ruidosos.
  La secuencialidad del modelo puede hacer que se ajuste demasiado a errores menores.

- Requiere más tiempo de entrenamiento debido al enfoque secuencial de los modelos.

- El ajuste fino de múltiples hiper parámetros puede aumentar la complejidad y el tiempo de procesamiento.

Concusión

- Bastante bueno para la precisión predictiva.

- Hay que considerar los recursos computacionales, sobre ajuste e interpretación.

Gracias