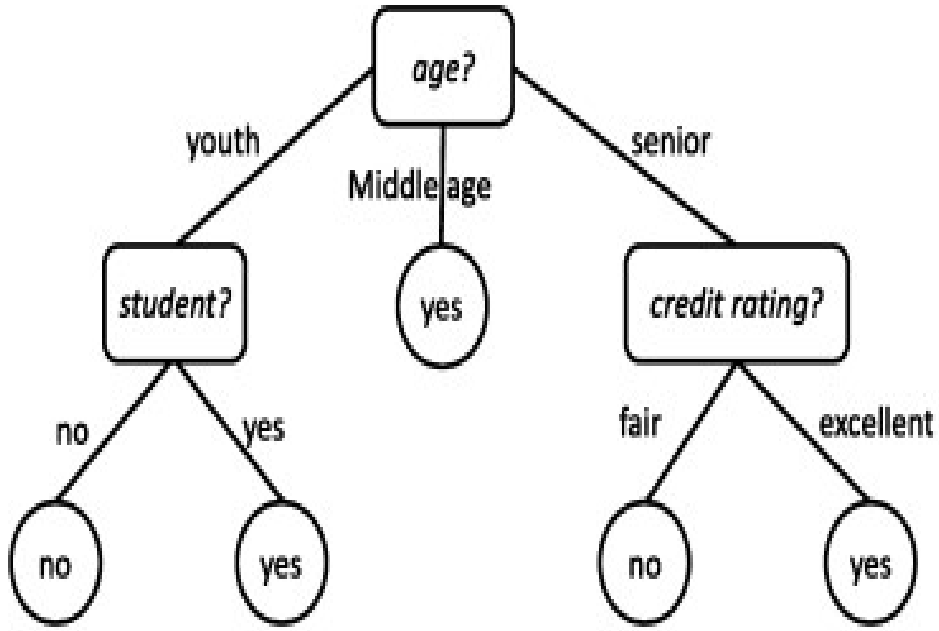
Arboles de decisión y modelos de regresión

Prof. Dr. Jorge Zavaleta
zavaleta.jorge@gmail.com

Arboles de decisión - Definición

- Los árboles de decisión son modelos de **aprendizaje automático** que representan un conjunto de reglas de decisión en una **estructura jerárquica**.
- Se utilizan para resolver problemas de **clasificación** y **regresión**.



Han, J. et al. **Data Mining: Concepts and techniques**. 4ed. Elsevier, 2023

Arboles de decisión - Estructura

- **Nodo raíz**: representa el comienzo del árbol y contiene todos los ejemplos del conjunto de datos.
- **Nodo interno**: Representa una pregunta o prueba sobre un atributo de los datos.
- **Nodo terminal** (hoja): representa una clase o un valor de regresión.
- **Ramas**: Conectan los nodos y representan los diferentes valores posibles para un atributo.

Arboles de decisión - Funcionamiento

- **Entrenamiento**: el árbol de decisiones se construye a partir de un conjunto de datos etiquetados.
  - El algoritmo de construcción de árboles elige el atributo que mejor divide los ejemplos en clases o valores de regresión.
- **Predicción**: para clasificar un nuevo ejemplo, el árbol se recorre desde el nodo raíz hasta un nodo terminal.
  - La clase o valor de regresión del nodo terminal es la predicción para la nueva muestra.

Arboles de decisión - Ventajas

- **Fácil interpretación**: las reglas de decisión se interpretan y visualizan fácilmente.
- **Robustez**: estos son modelos robustos que pueden manejar datos ruidosos y valores faltantes.
- **Versatilidad**: Puede usarse para resolver problemas de **clasificación** y **regresión**. **Baixo Custo Computacional**: A construção e a previsão de árvores de decisão são computacionalmente eficientes. O tempo de treinamento é geralmente menor em comparação com algoritmos mais complexos, como redes neurais ou Support Vector Machines (SVM).

Arboles de decisión - Desventajas

- **Sesgo**: los árboles de decisión pueden estar sesgados si el conjunto de datos de entrenamiento no está bien equilibrado.
- **Sobreajuste**: los árboles de decisión pueden ser demasiado complejos y funcionar bien en el conjunto de entrenamiento, pero no con datos nuevos.
- **Inestabilidad**: Pequeñas variaciones en los datos de entrenamiento pueden dar como resultado modelos de árbol completamente diferentes. Esto hace que los árboles de decisión sean **inestables** y menos confiables para realizar predicciones sólidas. Para mejorar la estabilidad, es recomendable utilizar técnicas de conjunto como **Random Forests** o **Gradient Boosting**.

Algoritmos de árboles de decisión

- **ID3**: Algoritmo clásico que utiliza información de ganancia de información para elegir el atributo que mejor divide los ejemplos.
- **C4.5**: Mejora de ID3 que utiliza la razón de ganancia para elegir el atributo y permite tratar valores faltantes.
- **CART**: Algoritmo que utiliza el criterio de Gini para elegir el atributo y puede usarse para problemas de regresión.
- **ID3**, **C4.5** y **CART** adoptan un enfoque codicioso (sin retroceso) en el que los árboles de decisión se construyen de una manera recursiva de arriba hacia abajo de divide y vencerás.

Ejemplos de aplicación

- **Diagnóstico médico**: Clasificar a los pacientes según sus síntomas.
- **Aprobación de crédito**: predecir si un cliente pagará un préstamo.
- **Marketing**: Segmentar clientes para campañas de marketing.
- **Detección de fraude**: Las empresas financieras utilizan árboles de decisión para identificar transacciones sospechosas. Según los patrones de gasto, la ubicación y otros factores, el árbol puede clasificar las transacciones como legítimas o fraudulentas.
- **Clasificación de Riesgo de Crédito**: Las instituciones financieras utilizan árboles de decisión para evaluar el riesgo de conceder préstamos. Según las características del cliente (como el historial de pagos, los ingresos, etc.), el árbol clasifica si el cliente tiene un riesgo de incumplimiento bajo, medio o alto.

Modelos de regresión - Definición

- Un modelo de regresión establece una **relación funcional** entre una **variable dependiente** (o respuesta) y una o más **variables independientes (o predictores)**. El objetivo es **predecir** o **explicar** el valor de la variable dependiente en función de las variables independientes.
- **Un modelo de regresión es una herramienta matemática que se utiliza para describir la relación entre variables**.
- Se utilizan para resolver problemas de **aprendizaje supervisado**, donde se utiliza un conjunto de datos con ejemplos etiquetados para entrenar el modelo.

Modelos de regresión - Funcionamiento

- **Entrenamiento**: el modelo de regresión se entrena con un conjunto de datos etiquetados, llamado conjunto de entrenamiento.
  - Cada ejemplo del conjunto de entrenamiento tiene un conjunto de atributos (características) y un valor de respuesta (variable dependiente).
- **Predicción**: después del entrenamiento, el modelo puede predecir el valor de la variable dependiente para nuevos ejemplos que no se utilizaron en el entrenamiento.

Modelos de regresión - Tipos

- **Regresión lineal:** Modelo simple que utiliza una línea recta para estimar el valor de la variable dependiente.
- **Regresión polinómica**: Modelo que utiliza una curva polinómica para estimar el valor de la variable dependiente.
- **Regresión Logística**: Modelo probabilístico que utiliza la función logística para estimar la probabilidad de que ocurra un evento.
- **Regresión KNN**: Algoritmo que utiliza el promedio de los valores de la variable dependiente de los K ejemplos más cercanos a ella en el espacio de características.

Modelos de regresión – Métricas

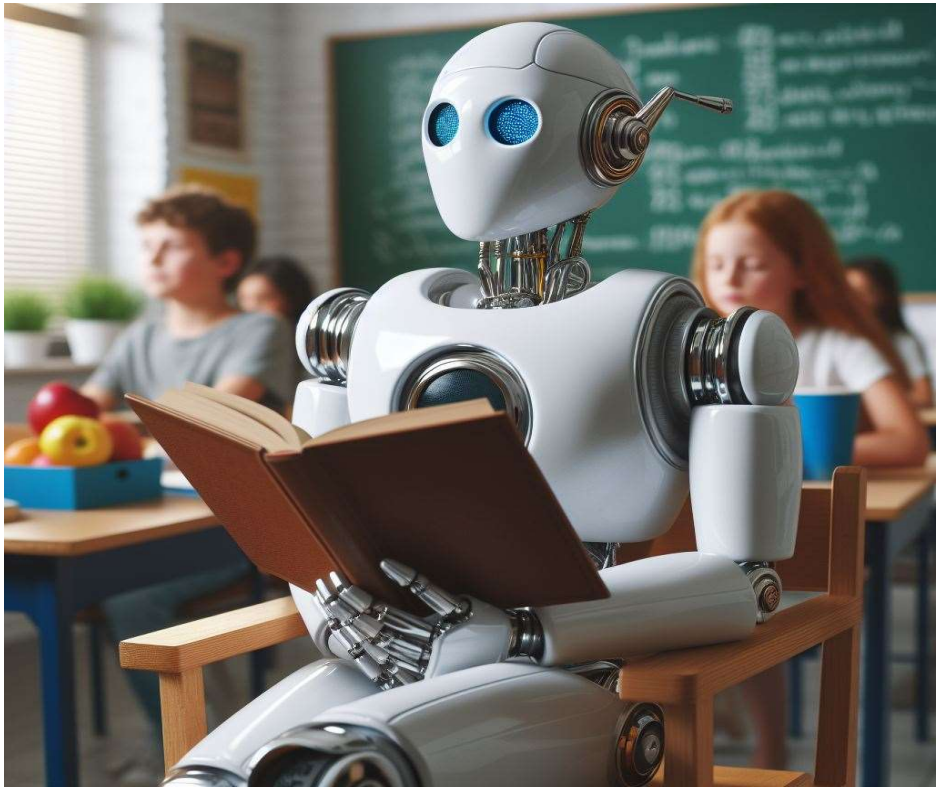
- **Error cuadrático medio** (**MSE**): Promedio de los cuadrados de las diferencias entre los valores reales y los valores predichos.
- **Raíz del Error cuadrático medio** (**RMSE**): Raíz cuadrada del **MSE**.
- **Coeficiente de determinación** (**R^2**): Proporción de la variabilidad de la variable dependiente explicada por el modelo.
- **Ajuste R^2**: **R^2** ajustado para penalizar la adición de variables irrelevantes al modelo.
- **MAE** (**Error absoluto medio**): MAE calcula el promedio de las diferencias absolutas entre los valores reales y previstos. Es útil entender el error medio en términos absolutos sin considerar la dirección de la diferencia.

Modelos de regresión – Consideraciones

- **Elección del modelo**: Depende del tipo de problema, datos disponibles y recursos computacionales.
- **Preprocesamiento de datos**: Limpiar, transformar y normalizar los datos es fundamental para el buen rendimiento de los modelos.
- **Interpretabilidad vs. Rendimiento**: los modelos más complejos pueden tener un mejor rendimiento, pero ser menos interpretables.
- **Monitoreo y actualización de modelos**: es importante monitorear el desempeño de los modelos a lo largo del tiempo y volver a entrenarlos cuando sea necesario.

Modelos de regresión – Conclusiones

- Los modelos de regresión son herramientas poderosas para analizar e interpretar datos, con aplicaciones en varias áreas.
- Elegir el modelo ideal, preprocesar los datos y evaluar el desempeño son pasos importantes para garantizar la calidad de los resultados.



@zavaleta, 2024