Predicción de Supervivencia del Titanic

Javier Horacio Pérez Ricárdez

Aprendizaje máquina I
Catedrático: Dr. Félix Orlando Martínez

25 de noviembre del 2024

Introducción

Esta aplicación predice la probabilidad de supervivencia de los pasajeros del Titanic usando un modelo de árbol de decisión. A continuación, se describen los pasos del proceso y la matemática detrás de las técnicas utilizadas.

Preprocesamiento de los Datos

Se carga un conjunto de datos que contiene información sobre los pasajeros, como la clase del pasajero (`Pclass`), el sexo (`Sex`), la edad (`Age`), el número de hermanos o esposos a bordo (`SibSp`), el número de padres o hijos a bordo (`Parch`), y la tarifa (`Fare`). A continuación, se realiza el siguiente preprocesamiento:

- Se convierten las categorías de sexo en valores numéricos (0 para mujeres y 1 para hombres).
- Se reemplazan los valores nulos en la columna `Age` con la media de la edad de los pasajeros.

La media de la edad (μ) se calcula como:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} \text{Age}_i$$

donde N es el número total de pasajeros con un valor de `Age` no nulo.

Entrenamiento del Modelo

El modelo utilizado es un **árbol de decisión** con el criterio de entropía y una profundidad máxima de 3. El árbol de decisión es un clasificador basado en dividir el espacio de características en regiones que maximicen la homogeneidad de las clases dentro de cada región. El criterio de selección de la mejor división en cada nodo se calcula usando la *entropía*.

La entropía (H) se calcula como:

$$H(S) = -\sum_{i=1}^{k} p_i \log_2 p_i$$

donde p_i es la proporción de elementos de la clase i en el conjunto S, y k es el número total de clases. El árbol selecciona la característica que minimiza la entropía.

Evaluación del Modelo

Una vez entrenado el modelo, se evalúa utilizando las siguientes métricas:

- **Precisión (Accuracy):** Es la proporción de predicciones correctas sobre el total de predicciones.

$$\text{Precisión} = \frac{TP + TN}{TP + TN + FP + FN}$$

  donde TP son los verdaderos positivos, TN son los verdaderos negativos, FP son los falsos positivos, y FN son los falsos negativos.

- **Precisión (Precision):** Mide la proporción de predicciones positivas correctas.

$$\text{Precisión} = \frac{TP}{TP + FP}$$

- **Exhaustividad (Recall):** Mide la proporción de casos positivos correctamente identificados.

$$\text{Exhaustividad} = \frac{TP}{TP + FN}$$

- **F1-Score:** Es la media armónica entre la precisión y la exhaustividad.

$$F1 = 2 \cdot \frac{\text{Precisión} \cdot \text{Exhaustividad}}{\text{Precisión} + \text{Exhaustividad}}$$

- **ROC-AUC:** La curva ROC (Receiver Operating Characteristic) mide el rendimiento del clasificador a través de la tasa de verdaderos positivos (TPR) frente a la tasa de falsos positivos (FPR). El área bajo la curva (AUC) es una métrica que evalúa la capacidad del modelo para discriminar entre clases.

  La TPR y la FPR se calculan como:

$$\text{TPR} = \frac{TP}{TP + FN}, \quad \text{FPR} = \frac{FP}{FP + TN}$$

Predicción

Una vez entrenado y evaluado el modelo, se puede realizar una predicción para un nuevo pasajero basado en sus características. Si el modelo predice que el valor es 1, significa que la persona sobrevivió, y si es 0, significa que no sobrevivió.

Exportación de Resultados

Los resultados de la predicción para el conjunto de datos de prueba se pueden exportar como un archivo CSV para su posterior análisis.