Actividad de Machine Learning: K-Nearest Neighbors (K-NN)

Gloria

Marzo 2025

Introducción

El algoritmo **K-Nearest Neighbors (K-NN)** es un método supervisado de clasificación que asigna una categoría a una nueva observación basándose en la clase mayoritaria de sus k vecinos más cercanos. Es útil por su simplicidad y efectividad cuando se tiene un conjunto de datos estructurado.

Metodología

Se utilizó el archivo `reviews_sentiment.csv`, el cual contiene reseñas de una aplicación. Se seleccionaron dos variables predictoras: la cantidad de palabras por comentario (`wordcount`) y el valor de sentimiento (`sentimentValue`). La variable objetivo fue el número de estrellas otorgadas (`Star Rating`).

Primero se dividió el conjunto de datos en entrenamiento y prueba utilizando `train_test_split`. Luego, se normalizaron los datos mediante `MinMaxScaler` para asegurar que ambas variables tuvieran el mismo rango. Posteriormente, se entrenó el modelo con `KNeighborsClassifier` de Scikit-Learn usando `k = 7`, valor determinado de forma empírica. Finalmente, se evaluó el modelo con precisión y se realizaron predicciones con nuevas muestras.

Fragmentos clave del código

```
df = pd.read_csv("reviews_sentiment.csv", sep=';')
X = df[['wordcount','sentimentValue']].values
y = df['Star Rating'].values
X_train, X_test, y_train, y_test = train_test_split(X, y,
    ↪ random_state=0)
scaler = MinMaxScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
knn = KNeighborsClassifier(n_neighbors=7)
knn.fit(X_train, y_train)
```

Resultados

El modelo alcanzó una precisión del **89.6%** en los datos de entrenamiento y **86.2%** en los de prueba.

```
print(knn.predict([[20, 0.0]]))
print(knn.predict_proba([[20, 0.0]]))
```

La predicción para una reseña con 20 palabras y sentimiento neutral fue **3 estrellas**, con probabilidad más alta en esa clase.

Conclusión

Esta actividad permitió comprender cómo K-NN puede clasificar datos con base en la proximidad. Se destacó la importancia de escalar las variables y de elegir un valor de `k` adecuado para lograr buenos resultados. El modelo obtuvo un desempeño sólido en un problema real de clasificación de reseñas.