Comparación de modelos supervisados y no supervisados

Abstract

Este trabajo presenta una comparación breve entre enfoques supervisados y no supervisados de aprendizaje automático aplicados a dos conjuntos de datos de referencia. Para el caso supervisado, se evalúa un modelo de *Random Forest Regressor* sobre el dataset *California Housing*, contrastándolo con un modelo base de Regresión Lineal. En el ámbito no supervisado, se explora la estructura del dataset *Iris* mediante la combinación de técnicas de reducción de dimensionalidad (*PCA* y *t-SNE*) con algoritmos de agrupamiento (*K-Means* y *GMM*)

Keywords

Random Forest, Gaussian Mixture Model, K-Means, Clustering

1 Introducción

El aprendizaje automático ofrece técnicas tanto supervisadas como no supervisadas para abordar problemas de predicción y descubrimiento de patrones. En este trabajo se comparan dos enfoques representativos: un modelo supervisado de regresión basado en *Random Forest* aplicado al dataset *California Housing*, y un modelo no supervisado de agrupamiento con *Gaussian Mixture Models* (GMM) sobre el dataset *Iris*, contrastado con un baseline *K-Means*. Se utilizan métricas específicas a cada contexto para evaluar el desempeño. Los resultados muestran que los modelos de ensamble capturan mejor relaciones complejas en regresión, mientras que la combinación *t-SNE + GMM* logra una representación de clusters más cercana a la partición real de las clases.

2 Modelos supervisados

2.1 Datos y preprocesamiento

Para los experimentos de los modelos supervisados se utilizó el conjunto de datos `California Housing`, disponible en la librería `scikit-learn`. Este dataset contiene información sobre precios de viviendas en California en función de diferentes características sociodemográficas y geográficas, con un total de 20,640 instancias y 8 atributos.

Antes del entrenamiento, los datos fueron divididos en dos conjuntos, utilizando un 80% para entrenamiento y un 20% para testing. Adicionalmente, las características predictivas fueron estandarizadas empleando la técnica de **`StandardScaler`** para mejorar la estabilidad numérica del modelo.

2.2 Algoritmos y parámetros

Se empleó el algoritmo **Random Forest Regressor**, que pertenece a la familia de métodos de *ensemble learning*. Este modelo construye múltiples árboles de decisión y combina sus predicciones para mejorar la generalización y reducir el riesgo de *overfitting*.

Los hiperparámetros principales considerados en los experimentos fueron:

- `n_estimators`: número de árboles en el bosque.
- `max_depth`: profundidad máxima de cada árbol.
- `min_samples_split`: número mínimo de muestras necesarias para hacer split
- `min_samples_leaf`: número mínimo de muestras requeridas en una hoja.

Se utilizó `GridSearchCV` para realizar una búsqueda exhaustiva de hiperparámetros con validación cruzada de 5 pliegues. El rango de valores explorado fue el siguiente:

- `n_estimators`: [100, 200, 300, 500]
- `max_depth`: [None, 10, 20, 30]
- `min_samples_split`: [2, 5, 10]
- `min_samples_leaf`: [1, 2, 4]
- `max_features`. ['sqrt', 'log2']

Para medir el rendimiento del modelo se utilizó el **MSE**.

2.3 Mejores Hiperparámetros

Tras aplicar GridSearch, se obtuvo que la mejor combinación para el modelo *Random Forest Regressor* fue:

- **n_estimators = 300:** Un mayor número de estimadores permite reducir la varianza del modelo, ya que cada árbol extra contribuye a un promedio más estable. Aunque incrementa el costo computacional, con 300 se logró un equilibrio favorable entre desempeño y tiempo de ejecución.
- **min_samples_split = 5:** Este valor evita splits demasiado agresivos que podrían llevar a *overfitting*, mejorando la generalización.
- **min_samples_leaf = 1:** Este valor permite capturar relaciones locales muy específicas, lo que resulta útil en datos heterogéneos como California Housing.
- **max_features = log2:** Se considera un subconjunto de tamaño $\log_2(M)$ en cada división, lo que reduce la correlación entre árboles y mejora la capacidad de ensamblado del bosque.
- **max_depth = None:** Permite que los árboles crezcan sin restricción, capturando relaciones complejas sin overfitting gracias al control del resto de hiperparámetros.

con un **mejor puntaje de validación cruzada** dado por:

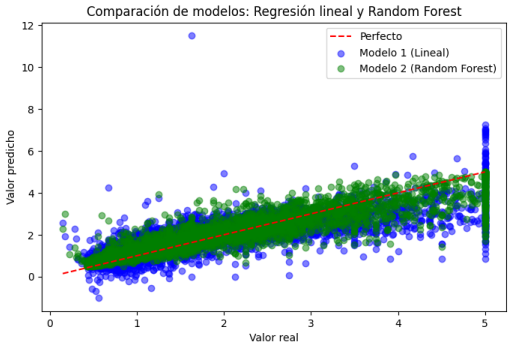
$$R^2 = 0.24408$$

2.4 Comparación

Métrica	Linear Regression	Random Forest
MSE	0.5559	0.2407
RMSE	0.7456	0.4906
MAE	0.5332	0.3202
R^2	0.5758	0.8163

Table 1: Comparación de métricas entre Linear Regression y Random Forest

Los resultados obtenidos permiten comparar el desempeño de dos modelos: **Regresión Lineal** y **Random Forest**. A partir de los valores de las métricas de evaluación, se observa lo siguiente:



- El modelo de Random Forest presenta valores notablemente menores que los obtenidos por la Regresión Lineal. Esto indica que Random Forest logra predicciones más precisas en promedio.
- El MAE también es menor en el modelo Random Forest en comparación con la Regresión Lineal. Esto significa que, en promedio, las predicciones de Random Forest se encuentran más cercanas a los valores reales.
- El modelo Random Forest alcanza un R^2 de 0.8163, lo que sugiere que explica más del 81% de la variabilidad de los datos. En contraste, la Regresión Lineal obtiene un R^2 de 0.5758, lo cual indica una capacidad predictiva menor.

En conjunto, los resultados muestran que **Random Forest supera ampliamente a la Regresión Lineal** en todas las métricas evaluadas. Esto se debe a que los modelos de ensamble como Random Forest pueden capturar relaciones no lineales y complejas entre las variables, mientras que la Regresión Lineal se limita a una relación estrictamente lineal.

Por lo tanto, se concluye que, para este conjunto de datos, Random Forest es el modelo más adecuado, al ofrecer un balance superior entre precisión y capacidad de generalización.

3 Métodos No Supervisados

3.1 Datos y preprocesamiento

Para esta sección se utilizó el dataset *Iris*. Dicho dataset contiene 150 muestras de flores distribuidas en tres clases: *Iris setosa*, *Iris versicolor* e *Iris virginica*. Cada muestra está caracterizada por cuatro variables numéricas: longitud y ancho del sépalo, así como longitud y ancho del pétalo.

El objetivo del análisis no supervisado fue explorar la estructura interna de los datos y evaluar en qué medida los algoritmos de agrupamiento eran capaces de identificar la separación natural entre las clases sin usar las etiquetas originales.

3.2 Algoritmos y parámetros

Con el fin de mejorar la representación de las relaciones no lineales entre los datos, se aplicó **t-SNE**. Esta técnica de reducción de dimensionalidad no lineal permitió visualizar agrupamientos más definidos, mostrando cómo el dataset formaba clusters un poco más compactos aunque aún 2 de las especies presentaban cierto grado de solapamiento.

Además, mediante el *método del codo* se determinó que el número óptimo de clusters a utilizar era 3, ajustando así los parámetros del modelo de clustering para reflejar la estructura subyacente del dataset. GMM es una generalizacion de KMeans así que también necesitamos proveerle una cantidad de clusters.

3.3 Comparación

Métrica	PCA + KMeans	t-SNE + GMM
Silhouette Score	0.5512	0.5012
Davies-Bouldin Index	0.6660	0.7483
Adjusted Rand Index (ARI)	0.7163	0.9039
NMI	0.7419	0.8997

Table 2: Comparación de métricas entre PCA+KMeans y t-SNE+GMM

Los resultados permiten comparar el desempeño de dos enfoques no supervisados: **PCA+KMeans** y **t-SNE+GMM**. A partir de los valores de las métricas de evaluación, se observa lo siguiente:

- El modelo **PCA+KMeans** obtuvo un mayor *Silhouette Score* y un menor *Davies-Bouldin Index*, lo que indica que las agrupaciones fueron más compactas y mejor separadas globalmente en comparación con t-SNE+GMM.
- En contraste, **t-SNE+GMM** alcanzó valores muy superiores en métricas basadas en etiquetas reales: un *Adjusted Rand Index (ARI)* de 0.9039 y una *Normalized Mutual Information (NMI)* de 0.8997. Esto sugiere que la estructura de clustering generada por t-SNE+GMM se asemeja mucho más a las clases originales del dataset *Iris*.

En conjunto, los resultados muestran que **PCA+KMeans ofrece agrupaciones más compactas y separadas**, pero **t-SNE+GMM se aproxima mejor a la verdadera partición de las clases del dataset**. Esto refleja un compromiso entre métricas geométricas (mejor para PCA+KMeans) y métricas de concordancia con etiquetas reales (mejor para t-SNE+GMM).

Este mejor desempeño puede explicarse por varias razones:

- **Representación no lineal de los datos**: a diferencia de métodos lineales como PCA, *t-SNE* preserva relaciones de vecindad locales en el espacio de baja dimensión, capturando estructuras complejas no lineales presentes en el dataset de Iris.
- **Separabilidad de los grupos**: al proyectar los datos en 2 dimensiones mediante *t-SNE*, las clases se distribuyen de forma más clara y diferenciada, lo que facilita el trabajo posterior del modelo de clustering.
- A diferencia de K-Means, que asume clusters esféricos y equidistantes, GMM permite modelar distribuciones elípticas y asignaciones probabilísticas, adaptándose mejor a las características reales de los datos que no siguen agrupaciones en forma esférica.

En conjunto, estas características explican por qué la combinación *t-SNE* + GMM logra capturar de forma más realista la estructura subyacente del dataset *Iris*, alcanzando métricas superiores de desempeño en comparación con otras combinaciones probadas.

Repositorio

El proyecto está disponible aquí!. Para mayor información de gráficos.