Lab 7

Integrantes:

- Manuel Archila 161250
- Juan Avila 20090
- Diego Franco 20240

Ejericicio 1

```
import numpy as np
from sklearn.svm import SVR
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.preprocessing import StandardScaler
from sklearn.gaussian_process import GaussianProcessRegressor
from sklearn.gaussian_process.kernels import RBF

def simulate_classifier_performance(C, gamma):
    X, y = load_dataset()
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

    scaler = StandardScaler()
    X_train_scaled = scaler.fit_transform(X_train)
    X_test_scaled = scaler.transform(X_test)

    clf = SVR(C=C, gamma=gamma, kernel='rbf')
    clf.fit(X_train_scaled, y_train)

    y_pred = clf.predict(X_test_scaled)

    mse = mean_squared_error(y_test, y_pred)
    r2 = r2_score(y_test, y_pred)

    return mse, r2

def load_dataset():
    np.random.seed(0)
    num_samples = 1000
    num_features = 10

    X = np.random.rand(num_samples, num_features)
    y = np.random.rand(num_samples)

    return X, y
```

```
np.random.seed(0)
X = np.random.rand(100, 2) * 10
mse_scores, r2_scores = zip(*[simulate_classifier_performance(C, gamma) for C, gamma in X])

X_train_mse, X_test_mse, mse_train, mse_test = train_test_split(X, mse_scores, test_size=0.2, ran

X_train_r2, X_test_r2, r2_train, r2_test = train_test_split(X, r2_scores, test_size=0.2, random_s

kernel = 1.0 * RBF(length_scale=1.0)
gp_mse = GaussianProcessRegressor(kernel=kernel, n_restarts_optimizer=10)
gp_mse.fit(X_train_mse, mse_train)

param_grid_mse = {
    'C': np.logspace(-3, 3, 7),
    'gamma': np.logspace(-3, 3, 7)
}
svm = SVR(kernel='rbf')  # Puedes probar otros kernels aquí también
grid_search_mse = GridSearchCV(svm, param_grid_mse, cv=5, scoring='neg_mean_squared_error')
grid_search_mse.fit(X, mse_scores)

best_params_mse = grid_search_mse.best_params_
best_mse = -grid_search_mse.best_score_



print("Mejores hiperparámetros para MSE:", best_params_mse)
print("Error cuadrático medio (MSE) con mejores hiperparámetros:", best_mse)
```

```
Mejores hiperparámetros para MSE: {'C': 0.001, 'gamma': 0.001}
Error cuadrático medio (MSE) con mejores hiperparámetros: 0.00016081174532542151
```

a. ¿Por qué este es un ejemplo de meta-modelado? - Porque se está creando un modelo que describe el comportamiento de otro modelo, en este caso el SVM. Se utiliza un modelo de regresión gausiana como el modelo sustitutivo para predecir el rendimiento de un suport vector machine SVM utilizando los datos reales que fueron obtenidos con la función simulate_classifier_performance. Luego de esto se encontraron los mejores hiperparámetros para el SVM utilizando el modelo sustitutivo.

b. ¿Cuál es el modelo sustitutivo? - El modelo de sustitutivo es el modelo de regresión gausiana, el cual se utiliza para predecir el rendimiento de un modelo SVM. Este se entrena con el conjunto de datos generado por el el clasificador SVM mezclando diferentes valores de C y gamma

c. Explique adecuadamente la construcción de su modelo y qué aplicaciones puede tener en la vida real

- Inicialmente se generan datos simulados para representar el rendimiento del SVM entrenando varios modelos con distintos hiperparámetros para luego calcular la métrica de MSE.

- Luego con los datos simulados se separaron los datos en 80% para entrenamiento y 20% para pruebas. Esto con el fin de poder entrenar el modelo sustitutivo y luego evaluarlo con los datos de prueba.
- Para calcular la metrica se creó un modelo de regresión gausiana para predecir el rendimiento de la SVM.
- Luego se utiliza gridSearch para encontrar los mejores hiperparametros para la SVM utilizando los modelos sustitutivos entrenados. Esto implica probar diferentes combinaciones de valores de C y gamma en una escala logarítmica y evaluar su rendimiento utilizando cross-validation.
- Tiene usos como la optimización de hiperparámetros y en la aceleración de procesos costosos. En casos de optimización de hiperparámetros como fue en este modelo se usa mucho en aprendizaje automático, ya que encontrar los mejores hiperparámetros puede ser un proceso muy costoso, por lo que utilizar modelos sustitutivos para estimar el rendimiento del modelo actual de esta forma puede ahorrar bastante tiempo y recursos.

Ejercicio 2

a. Explique el concepto de modelo sustituto y su papel en la aproximación de sistemas complejos. Mencione al menos un ejemplo de situaciones en las que el modelado sustituto sea particularmente beneficioso.

Un modelo sustituto es un modelo que tiene la peculiaridad de ser relativamente más sencillo, el cual se utiliza para simular un sistema más complejo y así poder entender, analizar o predecir el comportamiento del sistema complejo.

Como se mencionó anteriormente el papel de estos modelos sustitutos es poder adentrarse en la exploracion de sistemas o modelos más complejos sin la necesidad de tener que analizarlos directamente, ya que estos pueden ser muy complejos y no se puede tener una idea clara de su comportamiento. Además que por el hecho de ser complejos su analisis directo puede representar un costo computacional muy alto.

Ejemplo

Uno de los usos más comunes que se le da a este tipo de modelos es la predicción del clima, ya que el sistema climatico es muy complejo y no se puede analizar directamente, por lo que se utilizan modelos sustitutos para poder predecir su comportamiento

b. En el contexto del modelado sustituto (meta-modelado), ¿qué se entiende por sesgo de selección del modelo y cómo puede afectar la precisión de las predicciones del modelo sustituto? Detalle al menos una estrategia para mitigar este sesgo.

El sesgo de selección del modelo es el error que se comete al elegir un modelo sustituto, ya que este puede no ser el más adecuado para el sistema que se quiere modelar. Esto puede afectar la precisión de las predicciones del modelo sustituto ya que si el modelo sustituto no es el adecuado, las predicciones que se hagan con este no serán precisas. Este sesgo puede surgir cuando se selecciona un modelo basado en criterios subjetivos o cuando se elige un modelo que se ajusta demasiado (overfitting) a los datos de entrenamiento.

Estrategía

Validación cruzada: Utilizar técnicas de validación cruzada, como la validación cruzada k-fold o la validación cruzada leave-one-out, para evaluar el rendimiento del modelo sustituto en datos no vistos. Esto ayuda a estimar cómo se comportará el modelo en situaciones del mundo real y reduce la probabilidad de sobreajuste.

c. Analice el equilibrio entre la precisión del modelo y la eficiencia computacional al elegir la complejidad de un modelo sustituto. ¿En qué circunstancias optaría por un modelo más complejo y cuándo sería preferible un modelo más simple?

- Optar por un modelo más complejo:
    - Si se dispone de una gran cantidad de datos de alta calidad, un modelo más complejo podría aprovechar más el hecho de tener buena calidad de información.
    - Si la precisión es de vital importancia y los recursos computacionales no son un problema, un modelo más complejo podría ser la mejor opción porque se podria hacer una investigacion mucho más detallada.
    - Si se sabe que el sistema base es altamente complejo y que las relaciones son difíciles de capturar con un modelo simple, entonces un modelo más complejo podría ser necesario para obtener resultados precisos.
- Optar por un modelo más simple:
    - Cuando se sabe que no se tiene una amplia libertad sobre los recursos computacionales, un modelo más simple podría ser la mejor opción.
    - En algunos casos, la simplicidad del modelo puede ser deseable para facilitar la interpretación y comunicación de los resultados. Los modelos más simples suelen ser más fáciles de comprender y explicar.
    - Los modelos más simples tienden a ser menos propensos al sobreajuste, lo que significa que pueden generalizar mejor a nuevos datos. Esto es especialmente importante si se tienen datos limitados o ruidosos.

d. ¿Cuáles son las limitaciones del modelado sustituto y qué tipos de problemas pueden no ser adecuados para la aproximación del modelo

sustituto? Proporcione ejemplos para ilustrar estas limitaciones.

- Las limitaciones del pueden ser:
  - Los modelos sustitutos son aproximaciones de los modelos base por lo que pueden nos ser perfectas. Esto hace que no se pueda capturar completamente la complejidad del modelo base. Ejemplo: En el caso de tener una SVM con alto comportamiento no lineal puede ser que el modelo de regresión gausiana no pueda capturar completamente estos comportamientos.

  - Los modelos sustitutos pueden ser no muy buenos para predecir valores fuera del rango de los datos de entrenamiento. Ejemplo: En el caso de tener un modelo sustituto que predice el rendimiento de un modelo SVM, si se le pide que prediga el rendimiento de un SVM con un valor de C o gamma que no se encuentre en el rango de los datos de entrenamiento, el modelo sustituto no podrá predecirlo de forma correcta.

e. Imagine que tiene un problema complejo y de grandes dimensiones con recursos computacionales limitados. ¿Cómo abordaría la reducción de dimensionalidad y la selección de características antes de construir un modelo sustituto? ¿Qué métodos o técnicas podría emplear?

- Se puede utilizar un analisis de componentes principales o PCA para poder reducir la dimensionalidad de los datos y así poder trabajar con un modelo más simple. Esto se puede hacer ya que el PCA es una técnica de reducción de dimensionalidad que se utiliza para reducir un conjunto de variables correlacionadas a un número más pequeño de variables no correlacionadas, llamadas componentes principales.d

- Por otro lado se pueden utilizar autoencoders que son redes neuronales que permiten reducir la dimensionalidad, entrenando a la red para reconstruir las entradas después de comprimirlas en una representación de dimensiones más pequeñas que la original.