Clase N°4. Hiperparámetros

Los hiperparámetros son variables utilizadas para controlar el algoritmo de aprendizaje con el fin de mejorar la performance del modelo y suelen ser ajustados manualmente por el usuario. En Python se puede usar el comando **model.get_params()** para obtener una lista de todos los parámetros con sus respectivos valores en cada modelo.

Àrbol de regresión (Ajuste de hiperparàmetros)

DecisionTreeRegressor (ccp_alpha=0.0, **criterion='mse'**,
max_depth=None, max_features=None, max_leaf_nodes=None,
min_impurity_decrease=0.0, min_impurity_split=None,
min_samples_leaf=1, min_samples_split=2,
min_weight_fraction_leaf=0.0, presort='deprecated',
random_state=0, splitter='best')

DecisionTreeRegressor

- class_weight=None. Importancia relativa de los valores de clasificación.
- criterion=*'mse'/' squared_error'/ 'absolute_error'.*
- max_depth=3. Distancia max entre a raíz y las hojas.

Árbol de clasificación (Ajustar hiperparàmetros)

DecisionTreeClassifier (ccp_alpha=0.0,class_weight=None, **criterion='entropy'**, max_depth=3,max_features=None, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=Non, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, presort='deprecated',random_state=None, splitter='best')

DecisionTreeClassifier

- class_weight=None. Importancia relativa de los valores de clasificación.
- criterion= *'gini'/ 'entropy'*. Mide la calidad de la división en el árbol. Valores próximos a "1" indican impureza o desorden, mientras que los valores cercanos a "0" muestran pureza u orden.
- max_depth=3. Distancia máx. entre la raíz y las hojas.
- max_features=None. Número max de variables a considerar.
- max_leaf_nodes=20. Número max de hojas.
- min_impurity_decrease=0.0
- min_impurity_split=None. (Deprecado)
- min_samples_leaf=1. Podar si quedan menos que este número de ejemplos.
- min_samples_split=2. Continuar si quedan al menos esta cantidad de ejemplos.
- min_weight_fraction_leaf=0.0. Porcentaje mínimo de ejemplo para continuar.

Para visualizar un parámetro en particular,

- **get_depth ().** Muestra la profundidad del árbol
- **get_n_leaves ().** Muestra el nro. de hojas del árbol.

- **get_metadata_routing ().** Muestra la meta data de la ruta del objeto.

Método train-test Split

La técnica del Train-Test Split consiste en **descomponer de manera aleatoria una serie de datos**. Una parte servirá para el entrenamiento del modelo de Machine Learning, mientras que la otra permitirá probarlo para la validación. Por lo general, se reserva entre un 70 % y 80 % de los datos de la serie para el entrenamiento y la proporción restante de 30% o 20% para la evaluación y validación. En caso que el modelo se encuentre sobreajustado o subajustado se pueden ajustar los parámetros utilizando el método de validación cruzada.

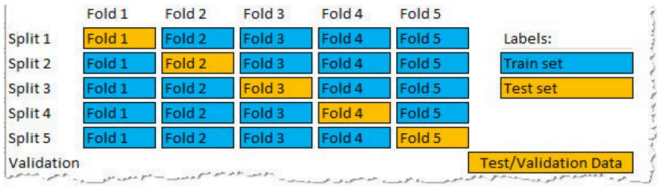
Método Cross-Validation (k-fold)

Un método común de validación cruzada es el **método K-Fold**, el cual consiste en dividir el conjunto de datos de entrenamiento en K pliegues o subconjuntos más pequeños. Luego, entrenamos el modelo K veces, utilizando un pliegue diferente como conjunto de validación en cada iteración y los otros K-1 pliegues como conjunto de entrenamiento.

Por ej. Si dividimos un conjunto en 5 subgrupos, resulta que k=5, entonces el modelo toma k-1=4 subconjuntos para entrenar y 1 subconjunto para evaluar. Luego evalúa las métricas de cada partición y selecciona al conjunto que maximiza la exactitud de la predicción.

Normalmente, divide los datos en 3 conjuntos.

- **Entrenamiento** : se utiliza para entrenar el modelo y optimizar los hiperparámetros del modelo.
- **Prueba** : se usa para verificar que el modelo optimizado funciona con datos desconocidos para probar que el modelo generaliza bien
- **Validación** : durante la optimización, cierta información sobre el conjunto de prueba se filtra en el modelo por su elección de los parámetros, por lo que realiza una verificación final en datos completamente desconocidos



Ejemplo de una división de datos de validación cruzada de 5 veces.

En el enfoque de validación cruzada más común, utiliza parte del conjunto de capacitación para las pruebas. Lo hace varias veces para que cada punto de datos aparezca una vez en el conjunto de prueba.

ACTIVIDAD

Considerando la base de datos analizada en la clase sobre indicadores socioeconómicos de la República Argentina, siendo la variable respuesta “poverty” y los factores: “school dropout” y “birth mortal”, evaluar en cuál de todos los supuestos se logra mejorar las métricas del modelo de árbol de regresión:

1. Incorporando una tercer variable observable “deficient_infra”.
2. Limitando la profundidad del árbol a un nivel de 4.
3. Aceptando un nivel mínimo de impurezas de: 0.03.