FUNDACIÓN YPF

Módulo V - Aprendizaje Supervisado

Clase 16: Validación cruzada

FUNDACIÓN YPF

¿Ponemos a grabar el taller?

FUNDACIÓN YPF

OBJETIVOS de hoy

- REPASO de temáticas anteriores
- Validación Cruzada

FUNDACIÓN YPF

REPASEMOS retomando clases anteriores, ¿Dudas?

FUNDACIÓN YPF

VALIDACIÓN CRUZADA

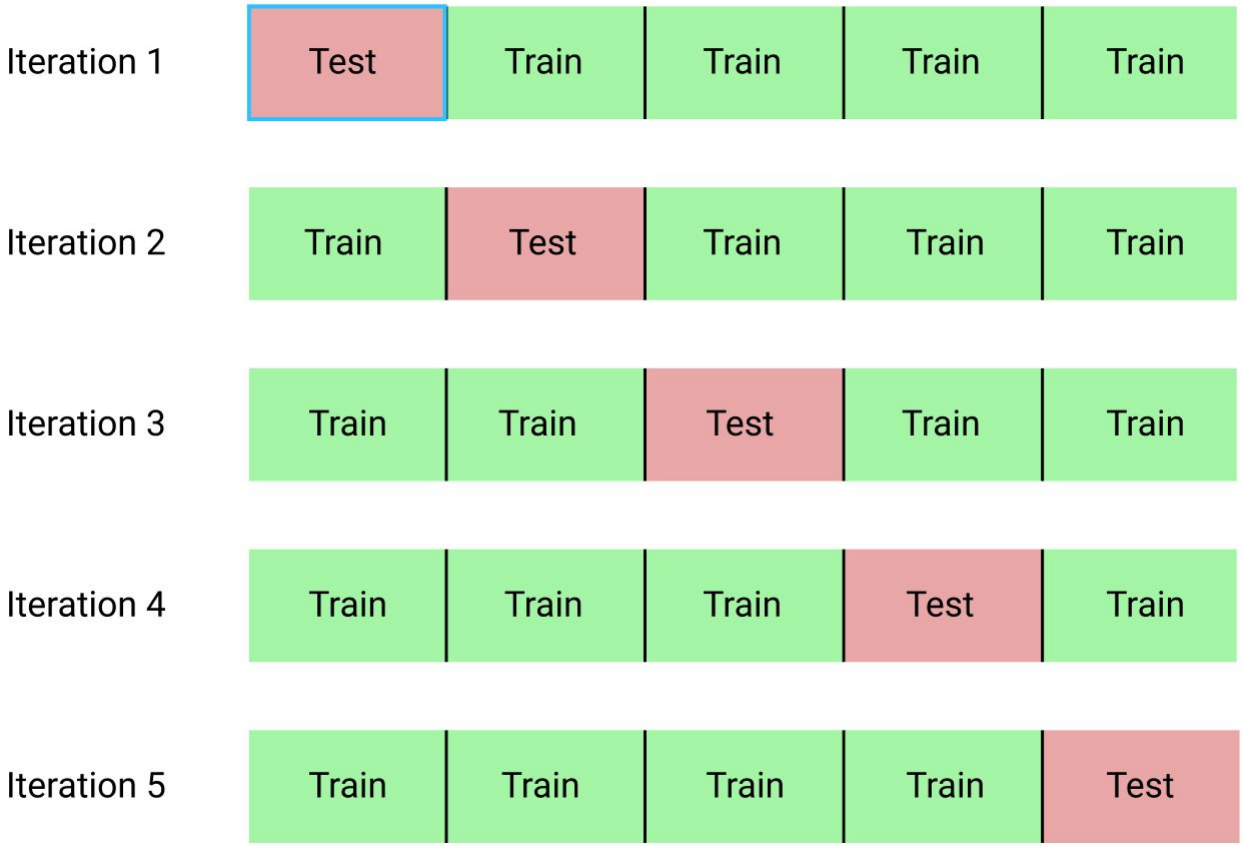
FUNDACIÓN YPF

Validación Cruzada: K-Fold

- La validación cruzada es un procedimiento de **remuestreo** que se utiliza para evaluar modelos de aprendizaje automático en una muestra de datos limitada.
- El procedimiento tiene un único parámetro llamado **k** que se refiere al número de grupos en que se dividirá una muestra de datos dada.

FUNDACIÓN YPF

Validación Cruzada: K-Fold



FUNDACIÓN YPF

Validación Cruzada: K-Fold

Pasos para realizar la validación cruzada:

1) Desordenar los datos

2) Separar en K folds del mismo tamaño

3) Para cada fold que separamos:

- Elegir un fold como Test set, y los K-1 folds restantes como Train set.
- Entrenar y evaluar el modelo.
- Guardar el resultado de la evaluación y descartar el modelo.

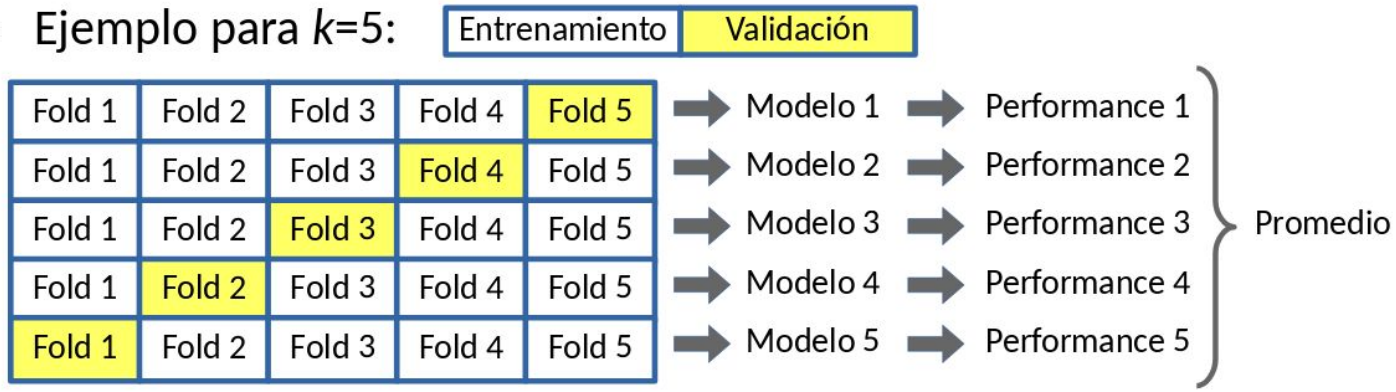
4) Obtener una medida de performance del modelo como el promedio de las K evaluaciones obtenidas en (3). También es una buena práctica incluir una medida de la varianza de las métricas obtenidas.

FUNDACIÓN YPF

Validación Cruzada: K-Fold

- Permite tener una estimación más realista de la performance del modelo.
- Generalmente resulta en una estimación menos sesgada de la habilidad del modelo para predecir.
- El valor de K no es trivial. Una mala elección puede resultar una mala evaluación del desempeño.

- Ejemplo para *k*=5:



FUNDACIÓN YPF

Descanso

Nos vemos en 10 minutos

FUNDACIÓN YPF

¿Cómo encuentro la óptima combinación de atributos + algoritmos + hiper parámetros ?

FUNDACIÓN YPF

Optimización

Un **MODELO** es un *pipeline* que se conforma de:

- Distintos atributos (selección y transformación de atributos)
- Distintos algoritmos (árboles, KNN, Linear Regression,etc)
- Distintos **hiper parámetros** de cada algoritmo.

FUNDACIÓN YPF

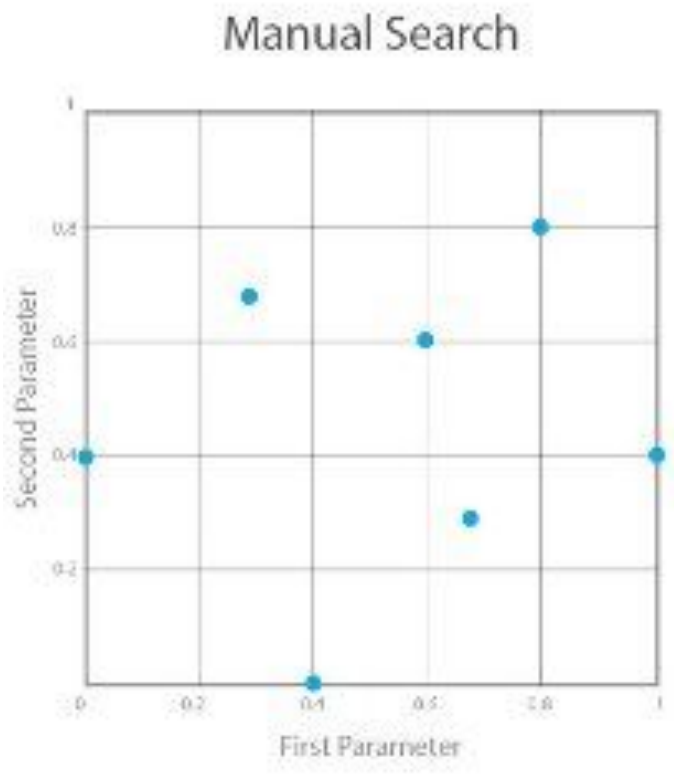
Optimización

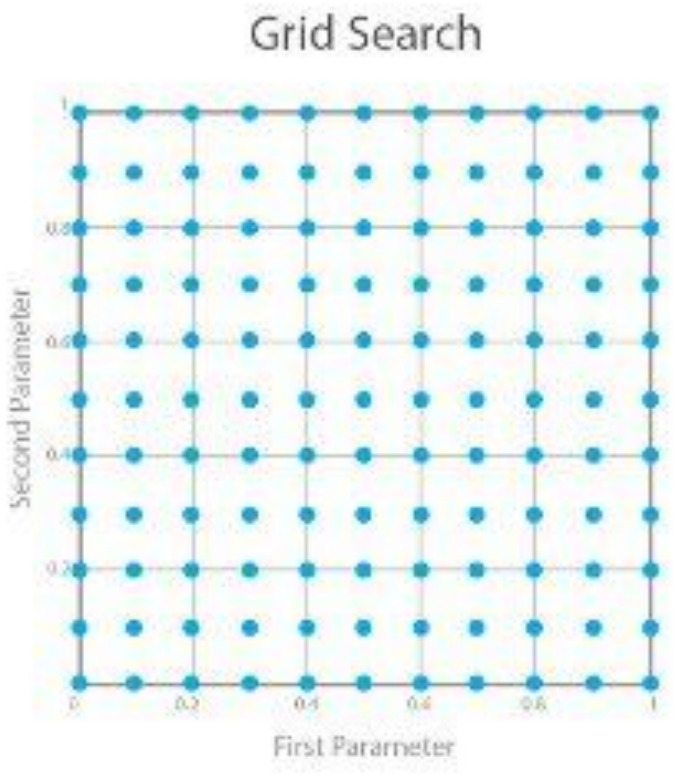
1) Definir el espacio de hiper parámetros
2) Explorar el espacio de búsqueda, definiendo un modelo para cada combinación posible de hiper parámetros. Para cada uno, entrenarlo y evaluar su desempeño.
3) Elegir la combinación con mejor desempeño, y entrenar el **único modelo** definitivo usando todos los datos.

FUNDACIÓN YPF

Optimización: Gridsearch

Técnica para buscar la mejor combinación de atributos + algoritmos + hiperparámetros de manera tal que el modelo sea el óptimo.





FUNDACIÓN YPF

Validación Cruzada y Optimización: GridSearchCV

```
from sklearn.model_selection import GridSearchCV
```

```
parameters_optimize = {
                      'max_features': ['auto', 'sqrt', 'log2', None],
                      'max_depth': [2,3, 4],
                      'criterion': ['gini', 'entropy'],
                      'bootstrap': [True, False],
                      'n_estimators': [2, 5, 10, 15, 20]
                      }
```

```
random_forest_hyp = RandomForestClassifier()

random_forest_search = GridSearchCV(random_forest_hyp,
                                    cv = 20,
                                    param_grid = parameters_optimize,
                                    n_jobs = 3)
random_forest_search.fit(X_train_imp, Y_train);
```

```
print('The best parameteres after GridSearchCV', random_forest_search.best_params_)
```

```
The best parameteres after GridSearchCV {'bootstrap': True, 'criterion': 'gini', 'max_depth': 4, 'max_features': 'aut
o', 'n_estimators': 20}
```

FUNDACIÓN YPF

Repasamos en Kahoot

FUNDACIÓN YPF

¿Dudas?

FUNDACIÓN YPF

Descanso

Nos vemos en 10 minutos

FUNDACIÓN YPF

Sección práctica:

Métricas y overfitting con la Notebook 18

FUNDACIÓN YPF

Trabajamos en salas de zoom

Métricas y overfitting

Trabajaremos con la Notebook 18

En los grupos establecidos, aprendemos cómo evaluar los modelos y evitar el overfitting optimizando los hiperparametros

50 minutos de actividad

FUNDACIÓN YPF

REPASAMOS

Revisamos los conceptos y código trabajados en la notebook 18

FUNDACIÓN YPF

¿Dudas?

FUNDACIÓN YPF

¡Muchas gracias!