

Machine Learning – Grid Search & Pipelines

¿Cuál va a ser el mejor modelo
para mi problema de ML?

Escogemos un modelo, por ejemplo Decision Tree. Voy a probar varias combinaciones de hiperparámetros para ver cuál me da mejor rendimiento

Criterion: ‘gini’ o ‘entropy’

max_depth: profundidad del árbol

min_samples_split

Pruebo con gini, max_depth=2, min_samples_split=2

Pruebo con gini, max_depth=3, min_samples_split=2

Pruebo con gini, max_depth=4, min_samples_split=2

Pruebo con gini, max_depth=5, min_samples_split=2

Pruebo con gini, max_depth=6, min_samples_split=2

Pruebo con gini, max_depth=7, min_samples_split=2

Pruebo con gini, max_depth=2, min_samples_split=3

Pruebo con gini, max_depth=2, min_samples_split=4

Pruebo con gini, max_depth=2, min_samples_split=5

Pruebo con gini, max_depth=2, min_samples_split=6

Pruebo con gini, max_depth=2, min_samples_split=7

Pruebo con gini, max_depth=2, min_samples_split=8

Pruebo con entropy, max_depth=2, min_samples_split=2

Pruebo con entropy, max_depth=3, min_samples_split=2

Pruebo con entropy, max_depth=4, min_samples_split=2

Pruebo con entropy, max_depth=5, min_samples_split=2

Pruebo con entropy, max_depth=6, m

Pruebo con entropy, max_depth=7, m

Pruebo con entropy, max_depth=4, m

Pruebo con entropy, max_depth=5, m

Pruebo con entropy, max_depth=6, m

Pruebo con entropy, max_depth=7, m

Pruebo con entropy, max_depth=2, min_samples_split=3

Pruebo con entropy, max_depth=2, min_samples_split=4

Pruebo con entropy, max_depth=2, min_samples_split=5

Pruebo con entropy, max_depth=2, min_samples_split=6

depth=2, min_samples_split=7

depth=2, min_samples_split=8

depth=2, min_samples_split=5

depth=2, min_samples_split=6

depth=2, min_samples_split=7

depth=2, min_samples_split=8



Calculo errores de todos los modelos, y elijo los hiperparámetros cuyo modelo tenga el mínimo error. Por ejemplo, entropy, max_depth=3, min_samples_split=4

Grid Search

Grid Search

No es necesario iterar manualmente cada una de esas combinaciones, tenemos herramientas en sklearn para hacerlo de manera automática.

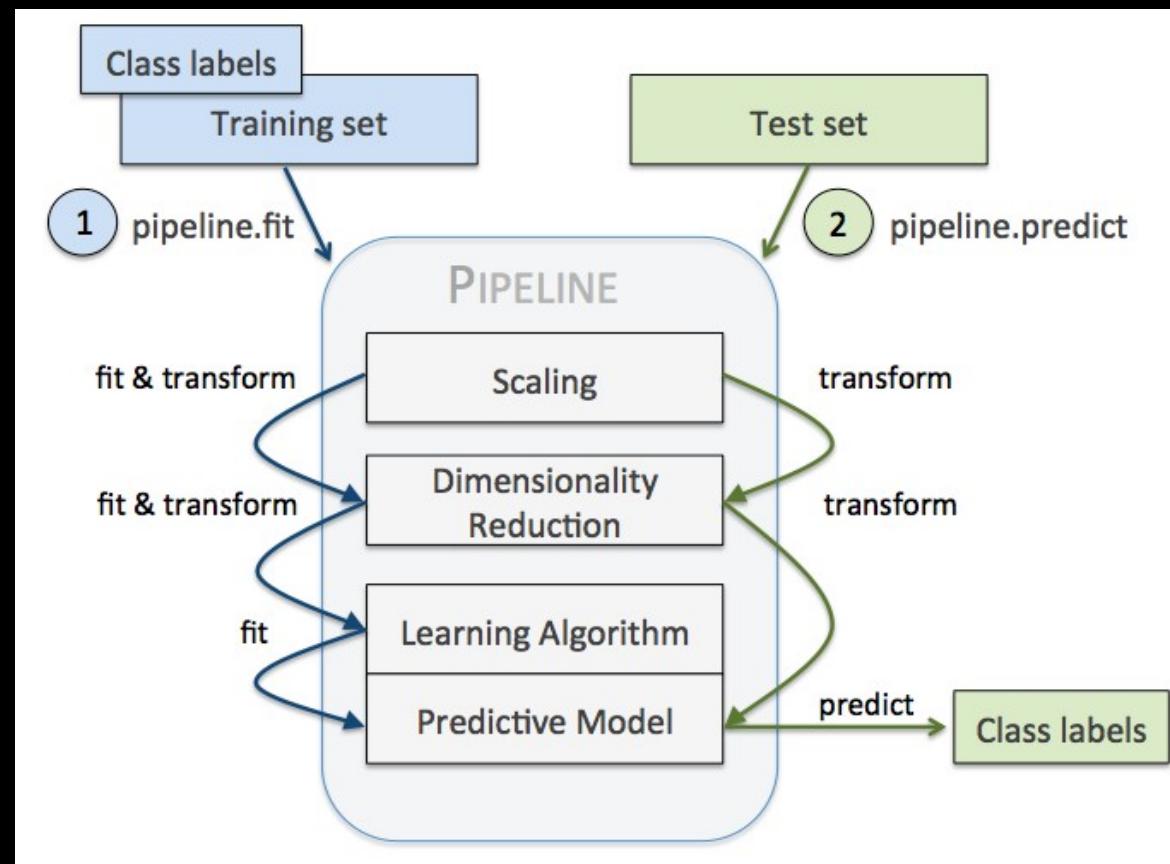
Existen varias técnicas como Grid Search, Random Search, Optimización Bayesiana...

La técnica más utilizada es Grid Search

1. Escojo un modelo
2. Elijo hiperparametros de ese modelo que quiero probar
3. Escojo un rango de valores para cada hiperparametro
4. Entrenamos mediante la función GridSearchCV (veremos en demo), que nos dará la mejor combinación de hiperparametros.
5. CV viene incluido en el argumento “cv”

```
grid = {  
    "max_depth": [3, 5, 10, 15, 20],  
    "max_features": [3, 4, 6, 8],  
    "min_samples_split": range(2, 10)  
}
```

Pipelines



Bibliografía

<https://learning.oreilly.com/library/view/hands-on-machine-learning/9781492032632/ch02.html#idm45022189270600>

<https://towardsdatascience.com/gridsearch-the-ultimate-machine-learning-tool-6cd5fb93d07>