
Técnicas Machine Learning 1

Iván G. Torre

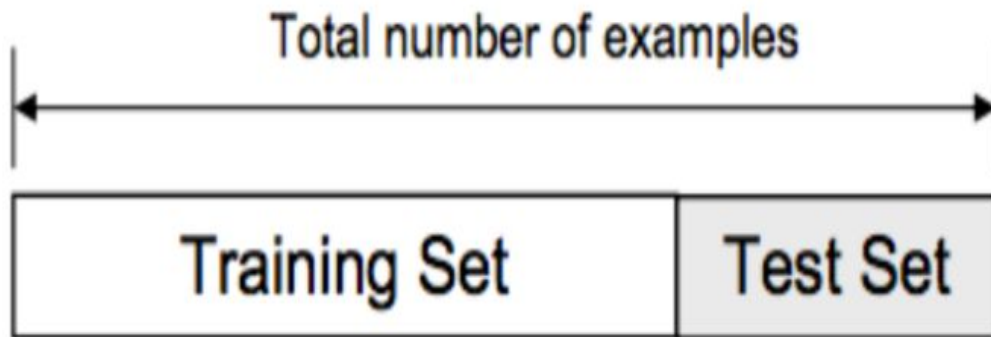
Iván G Torre: ivan.gonzalez.torre@upm.es

Jorge Calero: jorge.calero.sanz@alumnos.upm.es

José Olarrea: jose.olarrea@upm.es

División entre Train y Test en Clasificación

- Train: Model learn
- Test: Prediction



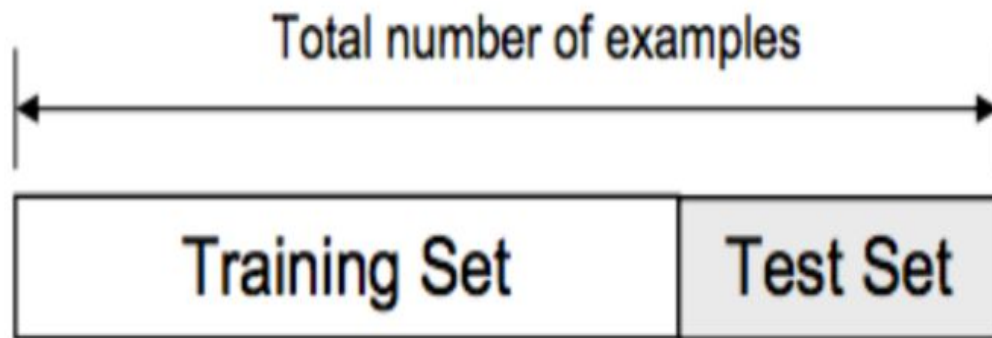
Train y Test: Problema

- Datos ordenados

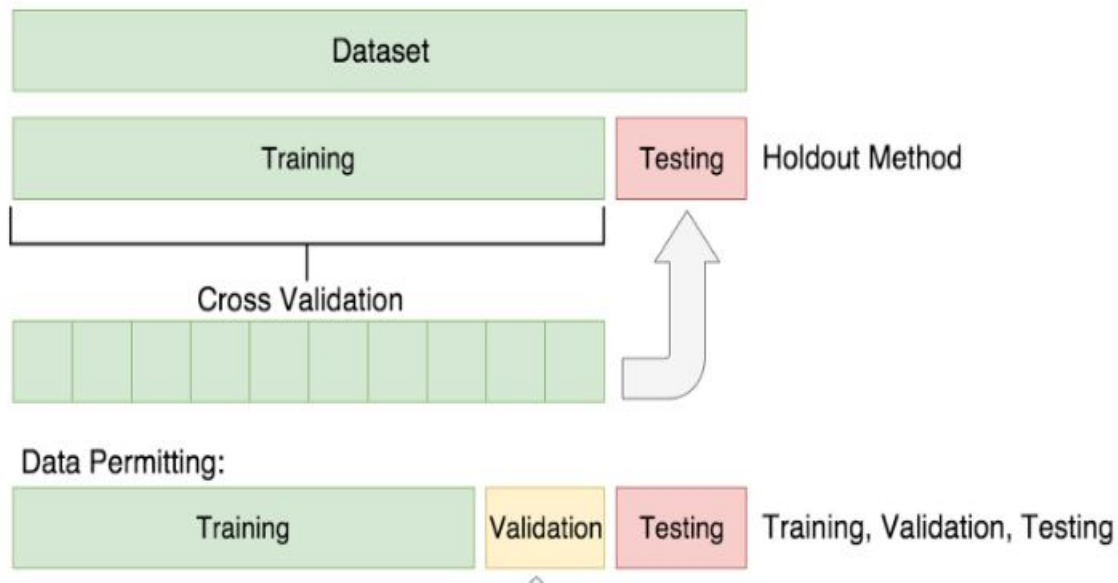


Train y Test: Punto de división Train-Test

- Arbitrario
- Overfitting



Train y Test: Validación Cruzada



Train y Test: Validación Cruzada

- K-Fold:
 - Se divide en k subconjuntos
 - Uno es validation, (k-1) train

5-fold CV

DATASET

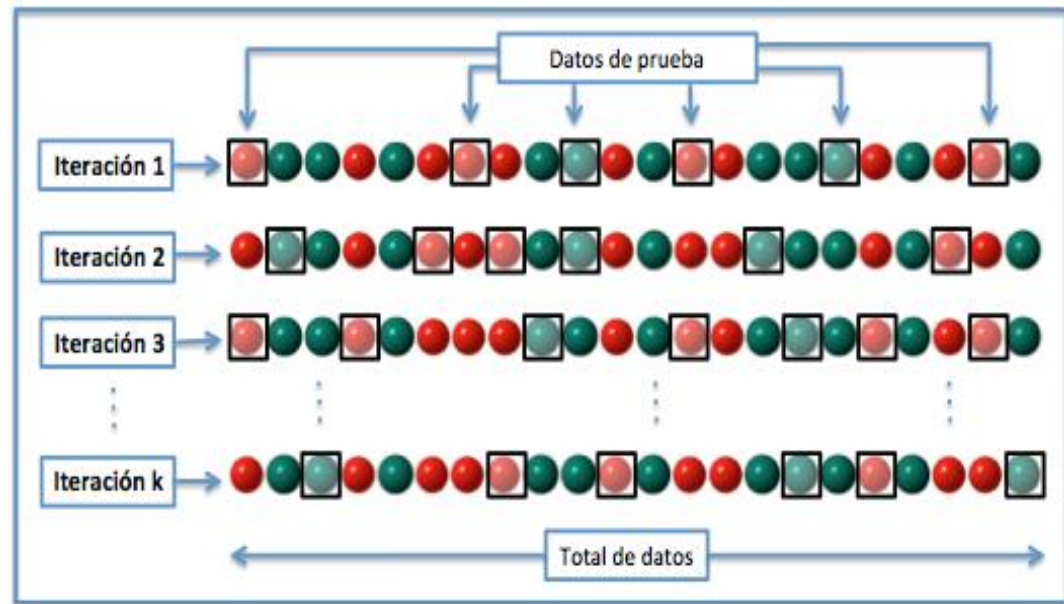


The diagram illustrates the 5-fold cross-validation process. A bracket labeled 'DATASET' spans the top of a table. The table has 5 rows, each representing an estimation. Each row contains 6 cells. The first cell of each row is labeled 'Estimation 1' through 'Estimation 5'. The remaining 5 cells in each row represent the split of the dataset: one is 'Test' (green background) and the other four are 'Train' (blue background). The 'Test' set rotates through the 5 folds.

Estimation 1	Test	Train	Train	Train	Train
Estimation 2	Train	Test	Train	Train	Train
Estimation 3	Train	Train	Test	Train	Train
Estimation 4	Train	Train	Train	Test	Train
Estimation 5	Train	Train	Train	Train	Test

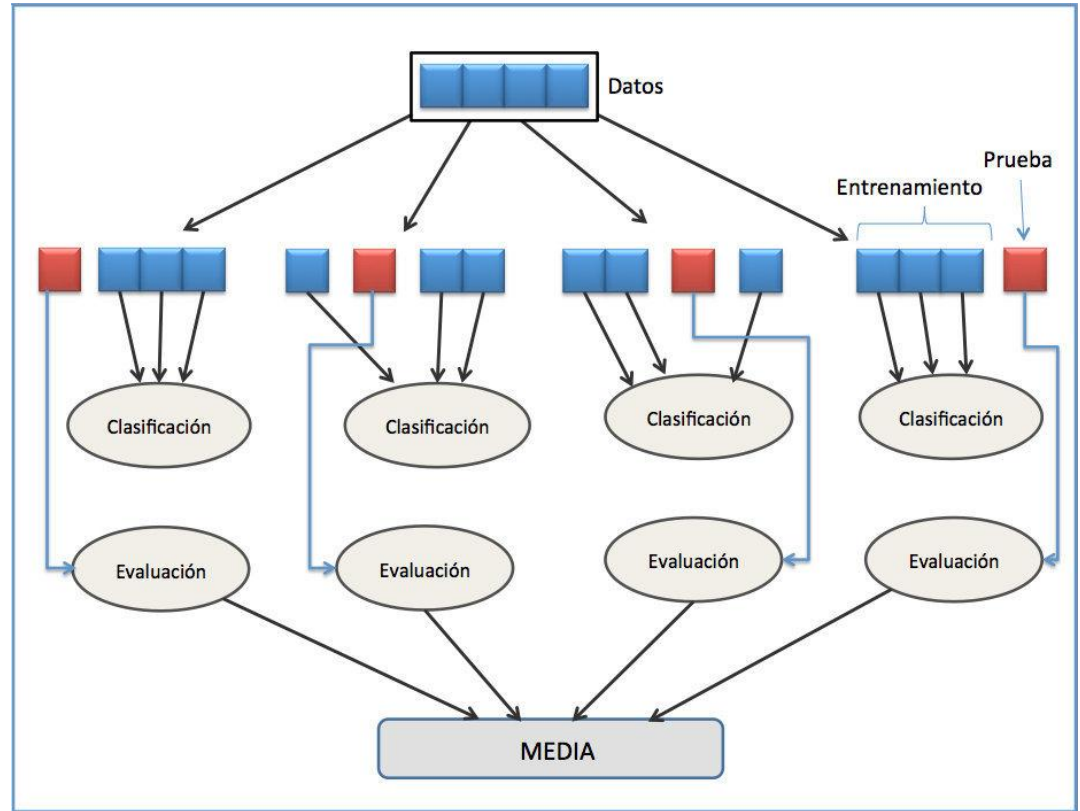
Train y Test: Validación Cruzada

- CV aleatoria:
 - Se elige aleatoriamente los datos de validación, el resto es train
- No depende del número de iteraciones
- Puede haber repetición



Train y Test.

- CV uno fuera:
 - Muy costoso computacionalmente

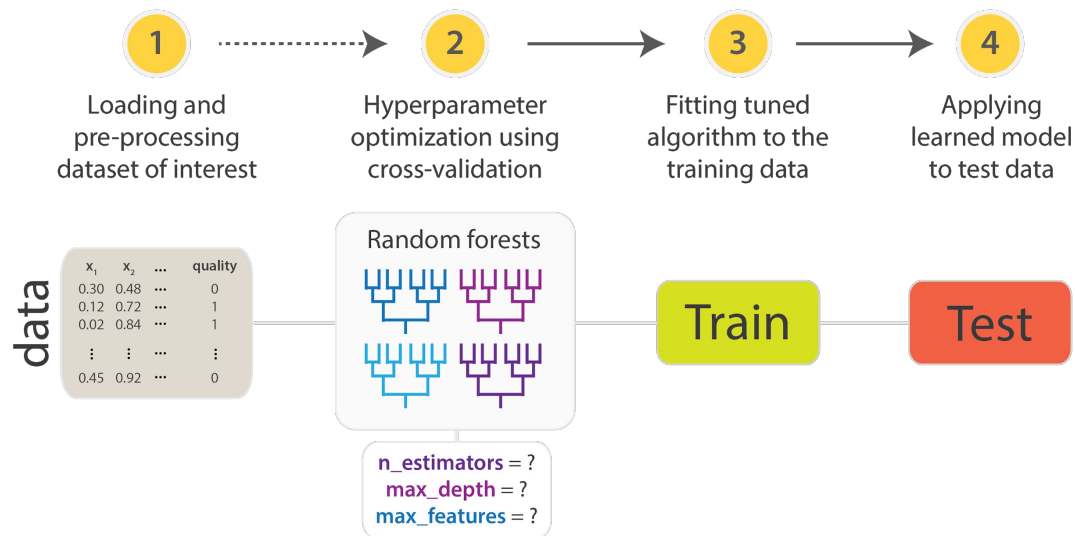


Cross Validation: Ejercicio



Hyperparameter optimization

- Datos -> OK
- Elección Modelo -> OK
- Tuneado del modelo ->



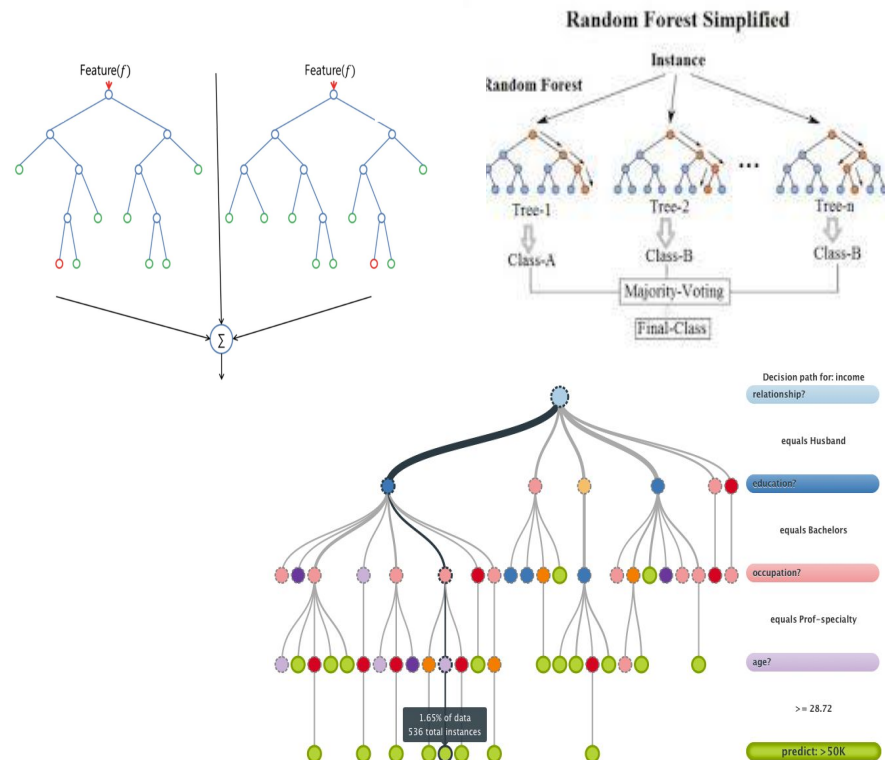
Hyperparameter optimization: Random Forest

- **n_estimators** : integer, optional (default=10)
 - The number of trees in the forest.
- **max_depth** : integer or None, optional (default=None)

The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.
- **n_jobs** : int or None, optional (default=None)
 - The number of jobs to run in parallel for both fit and predict.

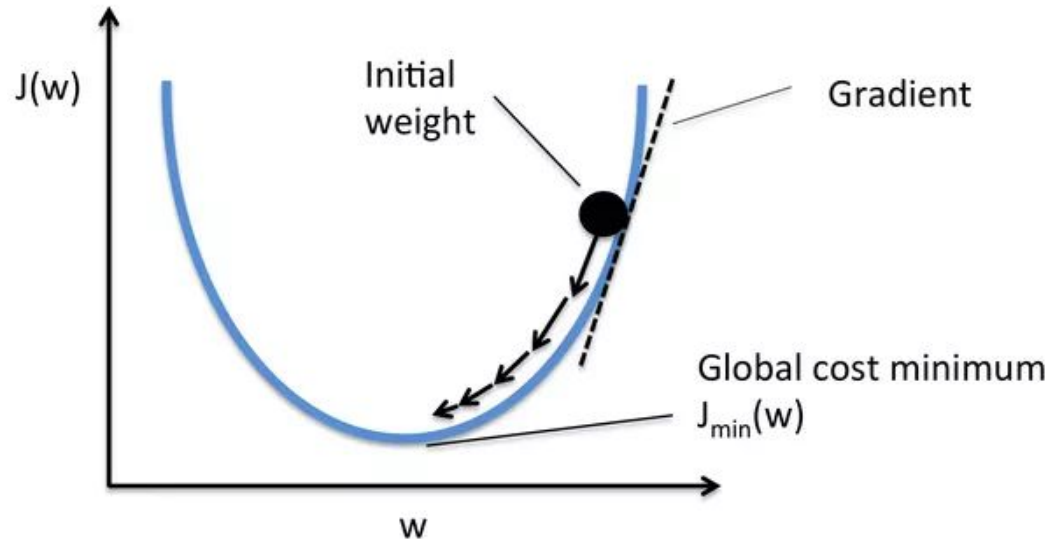
None means 1 unless in a [joblib.parallel_backend](#) context.

-1 means using all processors. See [Glossary](#) for more details.
- **class_weight** : dict, list of dicts, “balanced”, “balanced_subsample” or None, optional (default=None)



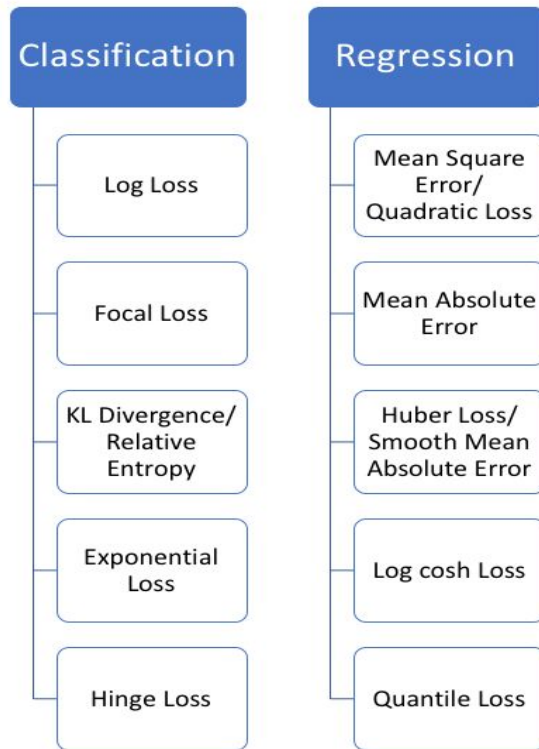
Hyperparameter optimization: loss function

- Necesitamos una métrica de error que minimizar



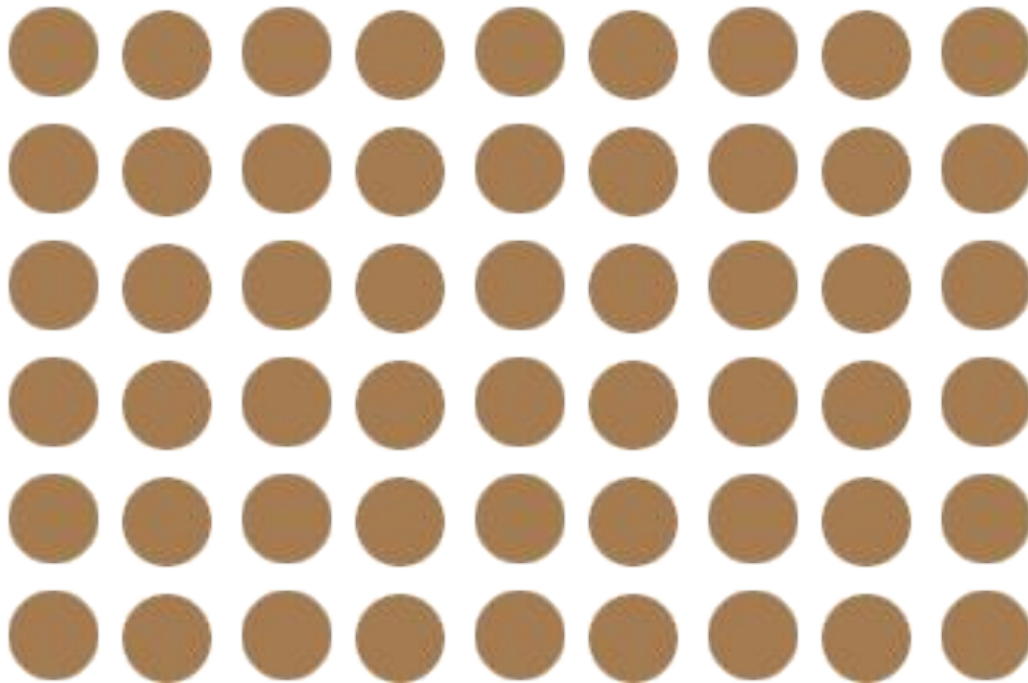
Hyperparameter optimization: loss function

- Dependiendo del modelo, existen muchas.
- Problema muy desarrollado para cálculo numérico entre otros



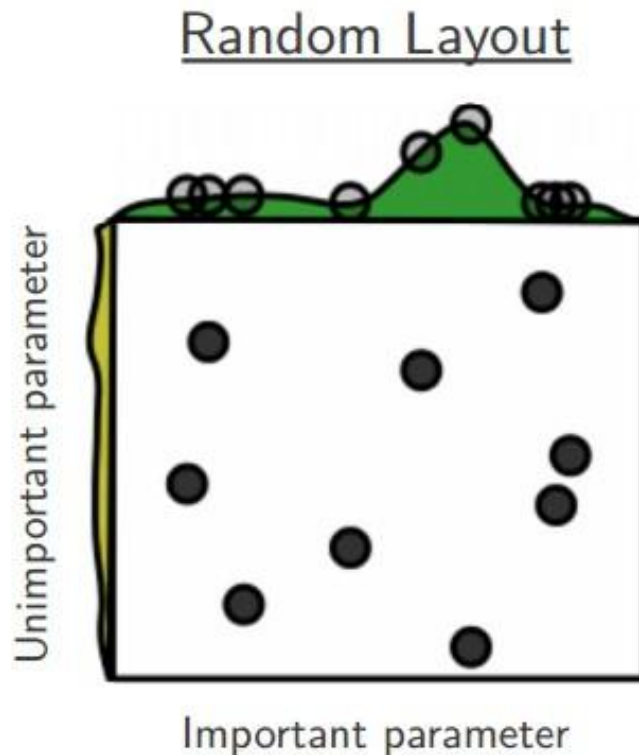
Hyperparameter optimization: Grid Search

- Búsqueda exhaustiva
- Paralelizable



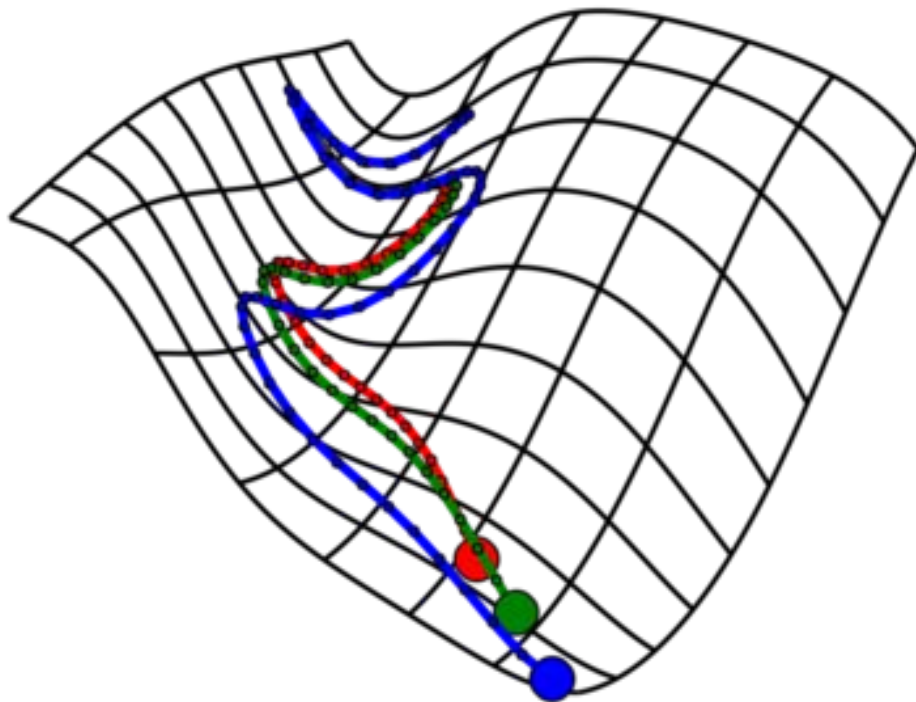
Hyperparameter optimization: Búsqueda aleatoria

- Búsqueda no exhaustiva
- Paralelizable



Hyperparameter optimization: Búsqueda gradiente

- Búsqueda no exhaustiva
- No Paralelizable
- Ampliamente estudiado



GridSearch en Clasificación: Ejercicio

