



Curso de especialización inteligencia artificial y big data

Tema 5: Evaluación y selección de modelos

Profesor: Sebastián Rubio Valero

Noviembre 2025



Sistema aprendizaje automático

1 Métodos supervisados

1.1 Hold-out-validation

Este método dividimos nuestros datos en dos conjuntos los de entrenamiento y los de test, por lo general 80%-20%. Una vez obtenido medidos el accuracy (la exactitud), esto lo podemos hacer mediante la matriz de confusión o aplicando alguna de las métricas (R^2 , MAE , *etc.*). Este método es adecuado para volúmenes grandes de datos, más de 1.000.000, donde se reserva una pequeña cantidad para las pruebas, alrededor del 10%. Es útil para hacer una estimación inicial de la capacidad de generalización del modelo de forma sencilla.



Figure 1: Enter Caption

1.2 K-Fold Cross Validation

Es el método más recomendado cuando el volumen de datos es pequeño, por ejemplo inferior a 10.000 muestras. Se obtiene una estimación robusta, al entrenar y evaluar el modelo K veces, y promediar los resultados, se obtiene una estimación del rendimiento del modelo que tiene una menor varianza que la obtenida por el Hol-out. Esto significa que la estimación es menos sensible a la división aleatoria de los datos.

Este método mantiene bajo el sesgo y la varianza, es decir hay un equilibrio entre el sesgo y la varianza. Además detecta el sobreajuste En cada una iteración se realiza

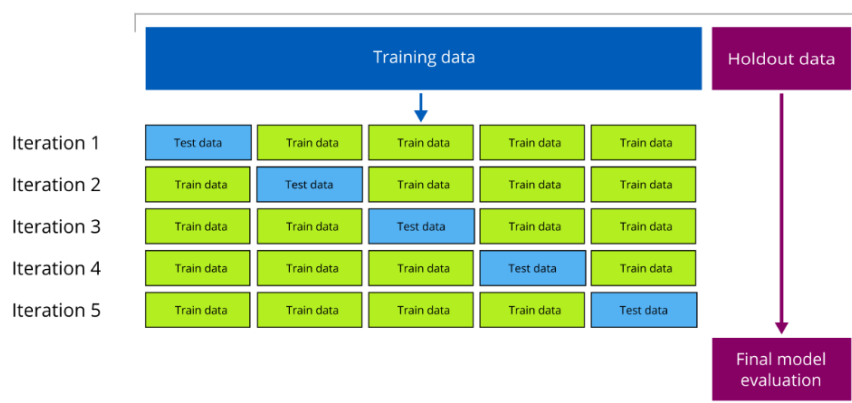


Figure 2: Enter Caption

un entrenamiento y una test, obteniendo así un rendimiento. Estas métricas va a depender del problema que se trate, si es de regresión o de clasificación. Para los modelos de clasificación las métricas son:

- Accuracy(Exactitud): La proporción de predicciones correctas sobre el total de predicciones
- Precision (Precision): De todos los valores positivos que se predijero ¿Cuántos son realmente positivos?, utilizamos la matriz de confusión
- Recal(Exhaustividad/Sensibilidad): De todos los valores que realmente eran positivos, ¿cuántos se predijeron correctamente?, utilizamos la matriz de confusión
- F1 Score: La media armónica de Precision y Recall
- AUC-ROC: El área bajo la curva de características operativas del receptor. Mide la capacidad del modelo de distinguir entre clases

Para los problemas de regresión donde lo que se mide es la magnitud del error cometido, es decir, qué tan cerca están las predicciones del modelo de los valores reales continuo Estas métricas se calculan para cada pliegue de validación, y el resul-

Métrica	Usar Cuando...	Evitar Cuando...
MSE	Penalizar errores grandes.	Hay muchos <i>outliers</i> .
RMSE	Error en unidades originales.	<i>Outliers</i> problemáticos.
MAE	Robustez frente a <i>outliers</i> .	Necesitas diferenciabilidad.
R²	Medir varianza explicada.	Hay sobreajuste.
R² Ajustado	Comparar modelos con más variables.	Modelos simples.
MAPE	Errores en porcentaje.	Valores cercanos a cero.

Table 1: Guía rápida para selección de métricas.

tado final es el promedio de los K valores obtenidos, proporcionando una estimación de error más estable.

Código python

```
from sklearn.datasets import load_iris
from sklearn.model_selection import cross_val_score, KFold
from sklearn.linear_model import LogisticRegression

X, y = load_iris(return_X_y=True)
model = LogisticRegression(max_iter=1000)

cv = KFold(n_splits=5, shuffle=True, random_state=42)
scores = cross_val_score(model, X, y, cv=cv)

print("Accuracy por fold:", scores)
print("Media de accuracy:", scores.mean())
```

1.3 Leave-One-Out Cross Validation (LOOCV)

Es una simple validación cruzada. Cada conjunto de aprendizaje se crea tomando todas las muestras excepto una, siendo el conjunto de prueba la muestra que se deja fuera. Así, par n muestras, tenemos n conjuntos de entrenamiento diferentes y n conjuntos de prueba diferentes.

En este método debes tener en cuenta que es muy costoso ya que se crea un modelo por cada instancia, por lo tanto debes utilizarlo cuando tengas pocas características y pocas instancias

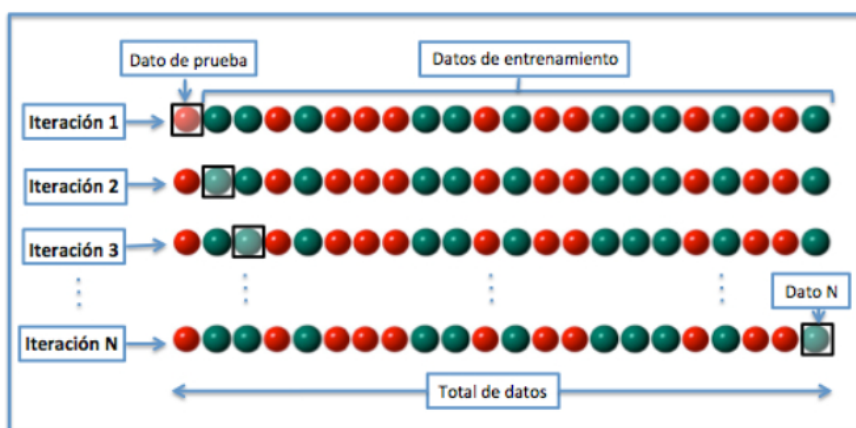


Figure 3: Enter Caption

1.4 K-fold estratificado

Algunos problemas de clasificación pueden presentar un gran desequilibrio en la distribución de la clase objetivo: por ejemplo, puede haber varias veces más muestras negativas que positivas. En estos casos se recomienda utilizar este método, para asegurar que las frecuencias relativas de las clases se conserven aproximadamente en cada parte de entrenamiento y validación. Es decir, cada conjunto contiene aproximadamente el mismo porcentaje de muestras de cada clase objetivo que el conjunto completo

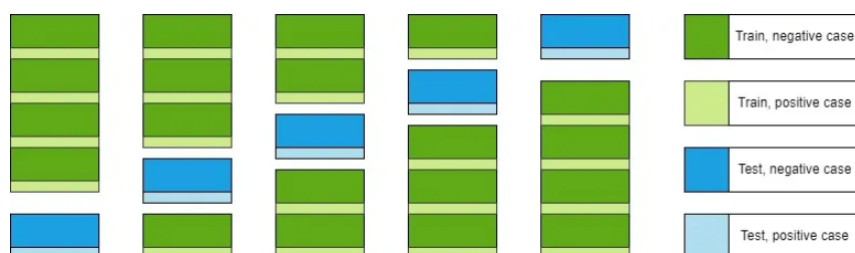


Figure 4: Enter Caption

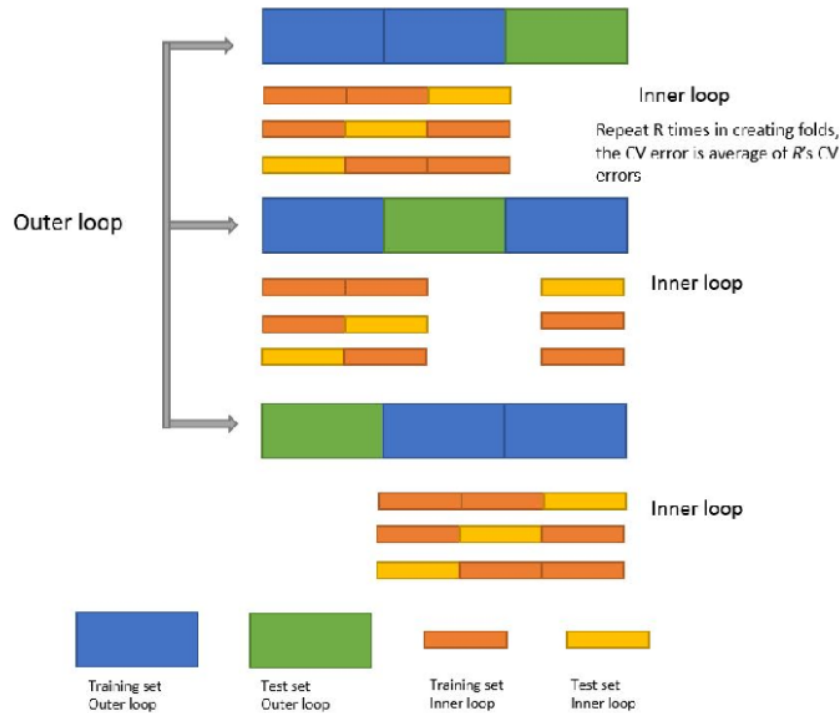
Código python

```
from sklearn.model_selection import StratifiedKFold
cv = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)
scores = cross_val_score(model, X, y, cv=cv)
```

1.5 Nested Cross Validation

Es un método que se utiliza para entrenar modelos en los que también hay que optimizar los hiperparámetros. Estima el error de generalización del modelo subyacente y sus hiperparámetros. La elección de los parámetros que maximizan la CV no anidada sesga el modelo.

Este método consta de dos bucles: uno interno y otro externo. En el bucle interno se usa para ajustar los hiperparámetros, probando diferentes combinaciones en un subconjunto de los datos. El bucle externo evalúa el modelo final entrenado, utilizando el rendimiento del bucle interno para obtener una estimación más precisa del error de generalización y evitar así el sobreajuste.



Showing the illustration of nested cross-validation, when $K, V = 3$.

Figure 5: Enter Caption

1.6 Time Series Cross Validation

1.7 Barajar los datos

Barajar los datos en la Validación Cruzada K-Fold es muy recomendable para mejorar la validez de la evaluación del modelo. Al establecer `barajar=Verdadero`, la barajación rompe cualquier orden inherente al conjunto de datos que pudiera introducir sesgos durante el proceso de validación. Esto garantiza que cada pliegue sea representativo de todo el conjunto de datos, lo que es crucial para evaluar lo bien que generaliza el modelo a los nuevos datos. Sin embargo, es importante evitar el barajado en los casos en que la secuencia de puntos de datos sea significativa, como ocurre con los datos de series temporales, para preservar la integridad del proceso de aprendizaje.

2 Métodos no supervisados

2.1 Train-Test Split con Métricas Internas

2.2 Stability-Based Validation

2.3 Consensus Clustering

2.4 Gap Statistic