Modelos Supervisados

Certificación en Ciencia de Datos

Andrés Martínez

28 Junio 2024

Universidad Externado de Colombia
135 Años

Contenido

- Introducción.
- Definición.
- Importancia.
- Métodos.
- Validación cruzada k-fold.
- Validación cruzada Leave-One-Out.
- Comparación de métodos.
- Implementación.
- Aplicaciones.
- Conclusión.

Introducción

- La validación cruzada es una técnica utilizada para evaluar el desempeño de un modelo.
- Ayuda a asegurar que el modelo no esté sobreajustado y que generalice bien a datos nuevos.
- La validación cruzada implica dividir el conjunto de datos en múltiples subconjuntos.
- Un subconjunto se utiliza para entrenar el modelo y otro para evaluarlo.
- Este proceso se repite varias veces para obtener una medida de desempeño más robusta.
- Ayuda a detectar problemas de sobreajuste.
- Proporciona una estimación más precisa del desempeño del modelo.
- Facilita la selección del mejor modelo y ajuste de hiperparámetros.

Validación cruzada simple

- Se divide el conjunto de datos en dos partes: entrenamiento y prueba.
- El modelo se entrena con la parte de entrenamiento y se evalúa con la parte de prueba.

$$\text{Error} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (1)$$

Figure 1: Hold Out

Validación cruzada k-fold

- Divide los datos en k subconjuntos (folds).
- Se entrena el modelo k veces, cada vez utilizando un fold diferente como conjunto de prueba y los restantes como conjunto de entrenamiento.
- El desempeño del modelo se promedia sobre los k experimentos.

$$\text{Error}_{\text{k-fold}} = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{1}{n_i}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_{ij})^2\right) \quad (2)$$

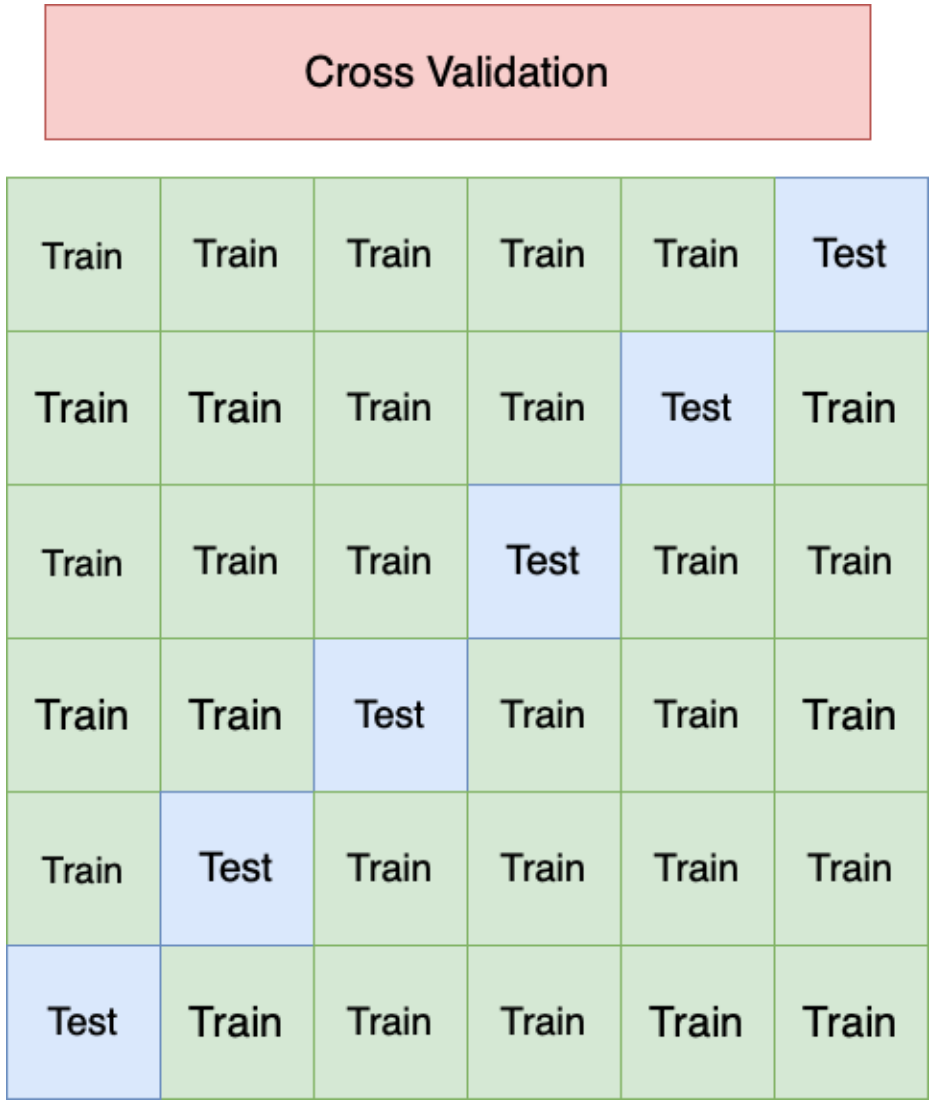

Figure 2: k-fold cross-validation

Validación cruzada Leave-One-Out

- Caso especial de la validación cruzada k-fold donde k es igual al número de observaciones.
- Cada observación es utilizada una vez como conjunto de prueba mientras el resto se utiliza para entrenar.
- Proporciona una evaluación muy precisa pero es computacionalmente costosa.

$$\text{Error}_{\text{LOO}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3}$$

Comparación de métodos

- La validación cruzada k-fold es un buen equilibrio entre sesgo y varianza.
- Leave-One-Out es útil para conjuntos de datos pequeños pero puede ser ineficiente para grandes conjuntos de datos.

Método	Ventajas	Desventajas
k-fold	Buen equilibrio	Computacionalmente caro
Leave-One-Out	Evaluación precisa	Muy caro para grandes datos

Implementación

- La mayoría de las bibliotecas de aprendizaje automático proporcionan funciones para implementar validación cruzada.
- Por ejemplo, en scikit-learn en Python se puede usar `cross_val_score`.
- Evaluación de modelos predictivos.
- Selección de modelos y ajuste de hiperparámetros.
- Estimación de la precisión de modelos en datos no vistos.