Validación cruzada en ML

Qué es la validación cruzada en machine learning?

La **validación cruzada** es una técnica fundamental en *machine learning* utilizada para evaluar el rendimiento de un modelo de aprendizaje supervisado y garantizar que este generalice bien a datos no vistos. Consiste en dividir los datos disponibles en conjuntos separados para entrenamiento y prueba, repitiendo este proceso de manera sistemática para obtener una evaluación robusta y evitar el sobreajuste.

¿Cómo funciona?

El objetivo principal es estimar cómo funcionará un modelo entrenado en un conjunto de datos sobre datos que no ha visto antes. La validación cruzada típicamente implica los siguientes pasos:

1. **División de datos**:
    - Los datos se dividen en varios subconjuntos llamados *folds*.
    - Por ejemplo, en una validación cruzada de 5 particiones (*5-fold cross-validation*), los datos se dividen en 5 partes aproximadamente iguales.
2. **Entrenamiento y prueba**:
    - En cada iteración, uno de los subconjuntos se utiliza como conjunto de prueba (*test set*), y los demás se utilizan para entrenar el modelo (*training set*).
    - Esto se repite hasta que cada subconjunto haya sido usado una vez como conjunto de prueba.
3. **Promedio de resultados**:
    - Al final, los resultados obtenidos en cada iteración (por ejemplo, métricas como precisión, *F1-score*, etc.) se promedian para obtener una estimación más confiable del desempeño del modelo.

Tipos comunes de validación cruzada

1. **k-Fold Cross-Validation**:
    - Es la más común.
    - Los datos se dividen en *k* partes (o *folds*), y el modelo se entrena y evalúa *k* veces.
2. **Leave-One-Out Cross-Validation (LOOCV)**:

- Cada observación en los datos se usa una vez como conjunto de prueba y el resto como conjunto de entrenamiento.
- Es útil para conjuntos de datos pequeños, pero puede ser computacionalmente costosa.

3. **Stratified k-Fold Cross-Validation**:
   - Similar a *k-Fold*, pero asegura que la proporción de clases en los conjuntos de entrenamiento y prueba sea similar a la proporción en los datos originales.
   - Es especialmente útil para datos desbalanceados.
4. **Time Series Split**:
   - Se utiliza para series temporales, respetando el orden temporal de los datos.
   - El conjunto de prueba contiene observaciones más recientes que las del conjunto de entrenamiento.

Ventajas de la validación cruzada

- **Mejor estimación del rendimiento del modelo**: Al evaluar sobre diferentes particiones, se obtiene una evaluación más confiable.
- **Reducción del riesgo de sobreajuste**: Ayuda a entender si el modelo se está ajustando demasiado a los datos de entrenamiento.
- **Uso eficiente de los datos**: Todos los datos se utilizan para entrenamiento y prueba, maximizando la información disponible.

Desventajas

- **Costos computacionales**: Puede ser costosa en tiempo y recursos, especialmente si *k* es grande o el modelo es complejo.
- **Configuración cuidadosa**: Si los datos no se dividen adecuadamente (por ejemplo, respetando el orden en series temporales), puede llevar a resultados engañosos.

En resumen, la validación cruzada es una herramienta esencial para evaluar y comparar modelos, ayudando a seleccionar el que mejor se desempeña en datos no vistos.