COLOMBIA POTENCIA DE LA VIDA

TIC

UNIVERSIDAD SERGIO ARBOLEDA

Análisis de Datos
Nivel Integrador

TALENTO TECH

UNIVERSIDAD SERGIO ARBOLEDA

PREPARACIÓN SEMANA 14

VALIDACIÓN CRUZADA

La Validación Cruzada es una técnica esencial en el campo del aprendizaje automático y la modelización estadística que aborda ciertos problemas asociados con la división tradicional de datos en conjuntos de entrenamiento y prueba. Aquí se explican las razones detrás de la necesidad de la Validación Cruzada y los problemas con la división tradicional:

Necesidad de la Validación Cruzada:

1. **Estimación más precisa del rendimiento del modelo:** La Validación Cruzada proporciona una estimación más precisa del rendimiento de un modelo al evaluarlo en múltiples divisiones de los datos. Esto es crucial para obtener una comprensión más robusta de cómo se comportará el modelo en datos no vistos.
2. **Aprovechamiento máximo del conjunto de datos:** Cuando se tiene un conjunto de datos limitado, la Validación Cruzada permite utilizar al máximo cada punto de datos tanto para entrenamiento como para prueba, mitigando el riesgo de obtener estimaciones sesgadas del rendimiento del modelo.
3. **Identificación de problemas de sobreajuste o subajuste:** La Validación Cruzada ayuda a identificar problemas de sobreajuste (overfitting) o subajuste (underfitting) al evaluar el modelo en diferentes conjuntos de datos de entrenamiento y prueba. Esto es crucial para ajustar la complejidad del modelo.

VALIDACIÓN CRUZADA

Problemas asociados con la división tradicional de datos:

1. **Sensibilidad a la partición inicial:** Dependiendo de cómo se divida inicialmente el conjunto de datos, el rendimiento del modelo puede variar considerablemente. La Validación Cruzada aborda esta sensibilidad al realizar múltiples divisiones y promediar los resultados.
2. **Riesgo de sesgo en la estimación del rendimiento:** Una única partición puede conducir a una estimación sesgada del rendimiento del modelo, especialmente si los datos están desequilibrados o si hay patrones específicos en la distribución de los datos. La Validación Cruzada proporciona una estimación más equitativa.
3. **Utilización ineficiente de los datos:** Al dividir los datos en un conjunto de entrenamiento y otro de prueba, se puede estar desperdiciando información valiosa. La Validación Cruzada aborda esto al utilizar todos los datos para entrenamiento y prueba de manera iterativa.

En resumen, la Validación Cruzada es esencial para obtener estimaciones más confiables del rendimiento del modelo y mitigar los problemas asociados con la división tradicional de datos, permitiendo así una toma de decisiones más informada en el desarrollo de modelos de aprendizaje automático.

TIPOS DE VALIDACIÓN CRUZADA

K-Fold Cross-Validation

En K-Fold Cross-Validation, el conjunto de datos se divide en k particiones (folds) y el modelo se entrena y evalúa k veces. En cada iteración, un fold diferente se utiliza como conjunto de prueba, mientras que los k-1 folds restantes se utilizan como conjunto de entrenamiento.

```python
#K-Fold
from sklearn.model_selection import cross_val_score, KFold
from sklearn.linear_model import LinearRegression
import numpy as np

# Datos de ejemplo
X = np.random.rand(100, 5)
y = 2*X[:, 0] + 3*X[:, 1] + np.random.randn(100)

# Modelo de regresión lineal
modelo = LinearRegression()

# K-Fold Cross-Validation con 5 folds
kf = KFold(n_splits=5, shuffle=True, random_state=42)
resultados = cross_val_score(modelo, X, y, cv=kf)

# Imprimir resultados
print("Resultados K-Fold Cross-Validation:")
print(resultados)
```

```
Resultados K-Fold Cross-Validation:
[-0.08654244  0.04620406  0.63352187  0.5818795   0.60967684]
```

TIPOS DE VALIDACIÓN CRUZADA

Stratified K-Fold Cross-Validation:

Stratified K-Fold es similar a K-Fold, pero asegura que la distribución de las clases sea similar en cada fold, lo cual es crucial para conjuntos de datos desequilibrados.

```python
#Stratified K-Fold
from sklearn.model_selection import StratifiedKFold
from sklearn.svm import SVC
from sklearn.datasets import make_classification

# Datos de ejemplo desequilibrados
X, y = make_classification(n_samples=1000, n_features=20, n_classes=2, weights=[0.95, 0.05], random_state=42)

# Modelo de SVM
modelo_svm = SVC(kernel='linear')

# Stratified K-Fold Cross-Validation con 3 folds
skf = StratifiedKFold(n_splits=3, shuffle=True, random_state=42)
resultados_stratified = cross_val_score(modelo_svm, X, y, cv=skf)

# Imprimir resultados
print("Resultados Stratified K-Fold Cross-Validation:")
print(resultados_stratified)
```

```
Resultados K-Fold Cross-Validation:
[0.38995955 0.48277231 0.58850767 0.46980928 0.61683635]
Resultados Stratified K-Fold Cross-Validation:
[0.9491018  0.95195195 0.95195195]
```

El código de los ejemplos vistos, lo puede encontrar aquí: https://acortar.link/G2FTqd

TIPOS DE VALIDACIÓN CRUZADA

Leave-One-Out Cross-Validation:

En Leave-One-Out Cross-Validation, se utiliza un solo dato como conjunto de prueba en cada iteración. Es útil cuando se tiene un conjunto de datos pequeño.

```python
#Leave-One-Out Cross-Validation

from sklearn.model_selection import LeaveOneOut
from sklearn.neighbors import KNeighborsClassifier
from sklearn.datasets import load_iris

# Datos de ejemplo (iris)
iris = load_iris()
X, y = iris.data, iris.target

# Modelo de clasificación k-NN
modelo_knn = KNeighborsClassifier()

# Leave-One-Out Cross-Validation
loo = LeaveOneOut()
resultados_loo = cross_val_score(modelo_knn, X, y, cv=loo)

# Imprimir resultados
print("Resultados Leave-One-Out Cross-Validation:")
print(resultados_loo)
```

```
Resultados K-Fold Cross-Validation:
[ 0.49058583  0.48907417 -0.0314514   0.5469021   0.55665921]
Resultados Stratified K-Fold Cross-Validation:
[0.9491018  0.95195195 0.95195195]
Resultados Leave-One-Out Cross-Validation:
[1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 0. 1.
 0. 1. 1. 1. 1. 1. 1. 1. 1. 1. 0. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 0. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 0.
 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
 1. 1. 1. 1. 1. 1.]
```

TIPOS DE VALIDACIÓN CRUZADA

EJEMPLO PRACTICO En este ejemplo, se utiliza un conjunto de datos aleatorio generado por make_classification de Scikit-Learn. El modelo de clasificación utilizado es un Bosque Aleatorio (RandomForestClassifier). Se aplica K-Fold Cross-Validation con 5 folds para evaluar el rendimiento del modelo. Los resultados impresos son las puntuaciones de la validación cruzada para cada fold.

```python
from sklearn.model_selection import cross_val_score, KFold
from sklearn.ensemble import RandomForestClassifier
from sklearn.datasets import make_classification

# Generar un conjunto de datos aleatorio
X, y = make_classification(n_samples=1000, n_features=20, n_classes=2, random_state=42)

# Modelo de clasificación - Bosques Aleatorios
modelo_rf = RandomForestClassifier(n_estimators=100, random_state=42)

# K-Fold Cross-Validation con 5 folds
kf = KFold(n_splits=5, shuffle=True, random_state=42)

# Resultados de la validación cruzada
resultados = cross_val_score(modelo_rf, X, y, cv=kf)

# Imprimir resultados
print("Resultados de K-Fold Cross-Validation:")
print(resultados)
```

```
Resultados de K-Fold Cross-Validation:
[0.895 0.86  0.935 0.88  0.925]
```

ÉTICA EN EL ANÁLISIS DE DATOS

La ética en el análisis de datos es un aspecto crucial que debe ser considerado en todas las etapas del proceso. Algunos puntos clave relacionados con la ética en el análisis de datos incluyen:

1. **Transparencia y Responsabilidad:** Es esencial que los profesionales de datos sean transparentes en sus métodos y resultados. Deben ser capaces de explicar cómo se recopilaron, procesaron y utilizaron los datos. La responsabilidad implica asumir las consecuencias de las decisiones basadas en los resultados del análisis.
2. **Privacidad y Protección de Datos:** Respetar la privacidad de los individuos es fundamental. El manejo ético de datos implica garantizar que la información personal se maneje con cuidado y que se tomen medidas para protegerla contra accesos no autorizados.
3. **Sesgo y Equidad:** Los modelos de datos pueden contener sesgos inherentes, y es fundamental identificar y abordar estos sesgos. La equidad en el análisis de datos implica garantizar que los modelos no discriminen injustamente a ciertos grupos.
4. **Consentimiento Informado:** Obtener el consentimiento informado de las personas cuyos datos se están utilizando es un principio ético básico. Las personas deben estar informadas sobre cómo se utilizarán sus datos y tener la opción de dar su consentimiento.

ÉTICA EN EL ANÁLISIS DE DATOS

5. **Interpretación Responsable:** La interpretación de los resultados del análisis de datos debe hacerse de manera responsable y contextualizada. Evitar la sobre interpretación y reconocer las limitaciones del análisis es parte de la ética en la comunicación de resultados.
6. **Seguridad de los Modelos en Producción:** Garantizar que los modelos implementados en producción sean seguros y no representen riesgos inesperados. Esto es particularmente importante en aplicaciones críticas como la atención médica y la toma de decisiones automatizada.
7. **Cumplimiento Legal y Normativo:** Los profesionales de datos deben cumplir con todas las leyes y regulaciones aplicables relacionadas con la protección de datos y la privacidad.
8. **Revisión por Pares y Evaluación Ética:** La revisión por pares y la evaluación ética de proyectos de análisis de datos pueden ayudar a garantizar la integridad y la ética en la práctica del análisis de datos.

En resumen, la ética en el análisis de datos implica un enfoque consciente y reflexivo para garantizar que los beneficios de la analítica de datos se logren de manera ética y responsable, evitando consecuencias no deseadas o discriminación.

REGRESIÓN AVANZADA

Breve Revisión de la Regresión Lineal Ordinaria: La regresión lineal ordinaria busca encontrar la línea (o hiperplano) que mejor se ajuste a los datos minimizando la suma de los cuadrados de las diferencias entre las observaciones reales y las predicciones del modelo.

En este ejemplo, la línea roja representa la regresión lineal ordinaria que busca minimizar la distancia vertical entre los puntos de datos y la línea. La regresión avanzada, como Ridge y LASSO, se introduce para abordar limitaciones y mejorar la capacidad del modelo para adaptarse a situaciones más complejas.

¿Preguntas?