Análisis Avanzado de Datos

EVALUACIÓN

02 AL 05 DE ABRIL

CRITERIOS DE EVALUACIÓN:

Comprensión y preparación de datos

- Calidad del análisis exploratorio y la preparación de datos para el modelado.

Selección y Justificación del Spline

- Justificación de la elección de tipo de spline
- Justificación y metodología para la selección del número y la ubicación de nodos

Métodos de validación cruzada

- Uso efectivo de técnicas de validación cruzada para ajuste de hiperparámetros y evaluación del modelo.

Selección y Justificación del Kernel:

- Correcta selección y entrenamiento de las características usadas para el kernel.
- Ha implementado la metodología adecuada para la selección del kernel.

Evaluación del Modelo:

- Se ha implementado metodologías de evaluación de modelos
- Ha utilizado métricas adecuada para evaluar el rendimiento

Los criterios para autoevaluación, coevaluación y heteroevaluación, se define a partir de la rúbrica anexa en la plataforma Moodle del curso

PARTE 1: 70 PUNTOS

DATASET A UTILIZAR:

House Prices: Advanced Regression Techniques:

El conjunto de datos contiene 79 variables explicativas que abarcan casi todos los aspectos de las viviendas residenciales, como el tipo de material del techo, el número de habitaciones, el año de construcción, la presencia de ciertas instalaciones (piscina, garaje, etc.), entre otros.

Recuerda: **No debemos aceptar ciegamente todos los atributos y sus interpretaciones, sino cuestionar y entender el contexto detrás de ellos**.

Análisis Avanzado de Datos

Link de descargas: está en la página del curso y se llama dataparte1.rar, en él encuentra la descripción de cada columna

Nota: Debe hacer todo su proceso de modelado con los datos llamados train.csv. en la evaluación por parte del profesor se usaron unos datos de test

OBJETIVO:

El objetivo principal es predecir el precio de venta (**SalePrice**) de cada vivienda y tener el mejor modelo, tanto para spline como usando kernel

DESARROLLO:

1. Exploración de datos (10 puntos):

- Visualizar las relaciones entre las características y el precio de vivienda e identificar posibles relaciones no lineales
- Identificar y tratar datos faltantes si los hay
- Si es necesario usar variables categóricas
- Evaluar la normalización de características según el método a utilizar
- Dividir el conjunto de datos, según el método de validación a utilizar

2. Modelado con spline (30 puntos):

- Definir qué características va a utilizar para predecir el precio de venta, y aplicar regresión spline.
- Explorar el tipo de spline (por ejemplo, spline cúbico o natural)
- Ajustar el modelo y evaluar su rendimiento en el conjunto de prueba.
- Ajustar el número y la ubicación de los nodos según sea necesario.
- Usar métricas como RMSE, MAE y R^2 para evaluación.

3. Modelado con Regresión con Kernel (30 punto):

- Elegir un kernel a utilizar y el número de característica a utilizar
- Ajustar el modelo en el conjunto de entrenamiento, posiblemente usando validación cruzada para encontrar el mejor parámetro de ancho de banda ventana para el kernel.
- Evaluar el rendimiento del modelo.
- Usar métricas apropiadas como RMSE, MAE y R^2 para la evaluación.
- Ajustar el modelo según sea necesario, quizás probando diferentes kernels o ajustando parámetros.

ENTREGABLES:

Un notebook con código bien comentado y organizado que documente el análisis y modelado, incluyendo:

1. **Análisis Exploratorio de Datos y pre-procesamiento**: Respectivas gráficas y descripciones para el análisis realizado. Explicar las consideraciones realizadas y los criterios para el pre-procesamiento y selección de datos

2. **Spline**: Tipo de spline seleccionado y razones para esta elección. Ecuaciones o representación del spline utilizado, Métricas de rendimiento en el conjunto de entrenamiento y validación según sea el caso. Recuerde entregar todo el proceso realizado.

3. **kernel**: Tipo de kernel elegido y justificación. Decisión sobre parámetros clave, como el ancho de banda/ Ventana, metodología usada para encontrar este parámetro y código dela misma. Métricas de rendimiento en el conjunto de entrenamiento y validación según sea el caso. Recuerde entregar todo el proceso realizado.

PARTE 2: 30 PUNTOS

DATASET A UTILIZAR

Los sistemas de bicicletas compartidas ofrecen una solución de transporte urbano flexible y automatizada, facilitando el estudio de patrones de movilidad mediante el registro detallado de cada alquiler. Estos datos son clave para entender y prever la demanda, mejorando así la gestión del sistema, la planificación urbana y fomentando el transporte sostenible.
Link de descargas: Para este análisis, utilizaremos el conjunto de datos disponible en https://archive.ics.uci.edu/dataset/275/bike+sharing+dataset, que también se encuentra en e-aulas como dataparte2.rar

OBJETIVO:

Predecir el número total de bicicletas alquiladas (cnt) en un día, utilizando un modelo de regresión con kernel basado en características como condiciones climáticas, temporada, hora del día, temperatura, etc.

ENTREGABLES:

Un notebook que incluya:

- **Exploración y Visualización de Datos:** Código y gráficos que proporcionen un entendimiento inicial de los datos.
- **Preparación de Datos:** Código que procese y prepare los datos para el modelado.
- **Modelado de Regresión con Kernel**: Implementación del modelo, incluyendo la elección de características y la optimización de parámetros.
- **Evaluación del Modelo:** Análisis del rendimiento del modelo mediante métricas establecidas.