Universidad Externado

Proyecto Capstone

Modelo de Predicción

Introducción Ciencia de Datos

Junio 2024

Departamento de Matemáticas

Capstone: Modelos supervisados para la predicción de KPI en una organización

June 9, 2024

Definición

Los modelos supervisados son una categoría de algoritmos de aprendizaje automático que se entrenan utilizando datos etiquetados. Esto significa que cada ejemplo del conjunto de datos de entrenamiento tiene una etiqueta o valor objetivo asociado, que el modelo intenta predecir. El objetivo principal de un modelo supervisado es aprender una función que mapea entradas a salidas basándose en ejemplos de entrada-salida del conjunto de datos de entrenamiento.

Estructura de los Modelos Supervisados

La estructura de un modelo supervisado generalmente incluye los siguientes pasos:

1. **Recolección de Datos:** Obtención de un conjunto de datos etiquetado que incluye ejemplos de entrada y sus correspondientes salidas.
2. **Preprocesamiento de Datos:** Limpieza y transformación de los datos para prepararlos para el análisis. Esto puede incluir la normalización, manejo de valores faltantes y codificación de variables categóricas.
3. **División de Datos:** Separación del conjunto de datos en subconjuntos de entrenamiento y prueba para evaluar el rendimiento del modelo.
4. **Selección del Modelo:** Elección de un algoritmo de aprendizaje supervisado adecuado para el problema.
5. **Entrenamiento del Modelo:** Utilización del conjunto de datos de entrenamiento para ajustar el modelo a los datos.
6. **Evaluación del Modelo:** Medición del rendimiento del modelo utilizando el conjunto de datos de prueba y métricas de evaluación específicas.
7. **Ajuste del Modelo:** Refinamiento del modelo mediante la optimización de hiperparámetros y mejoras en la arquitectura.
8. **Predicción:** Aplicación del modelo entrenado a nuevos datos para hacer predicciones.

Tipos de Modelos Supervisados

Los modelos supervisados se pueden clasificar en dos categorías principales: regresión y clasificación.

Regresión

La regresión se utiliza cuando la variable objetivo es continua. El objetivo es predecir un valor numérico. Ejemplos de modelos de regresión incluyen:

- **Regresión Lineal:** Modelo que asume una relación lineal entre las variables de entrada y la variable objetivo.
- **Regresión Polinómica:** Extensión de la regresión lineal que puede modelar relaciones no lineales.
- **Regresión de Ridge y Lasso:** Variantes de la regresión lineal que incluyen términos de regularización para evitar el sobreajuste.

Clasificación

La clasificación se utiliza cuando la variable objetivo es categórica. El objetivo es asignar etiquetas a las entradas. Ejemplos de modelos de clasificación incluyen:

- **Regresión Logística:** Modelo utilizado para predecir la probabilidad de una clase binaria.
- **Máquinas de Soporte Vectorial (SVM):** Modelos que encuentran el hiperplano que mejor separa las clases en el espacio de características.
- **Árboles de Decisión:** Modelos que dividen el espacio de características en regiones basadas en reglas de decisión.
- **Bosques Aleatorios:** Conjunto de árboles de decisión entrenados de manera independiente y cuyos resultados se combinan para mejorar la precisión.
- **Redes Neuronales:** Modelos inspirados en la estructura del cerebro humano que pueden aprender representaciones complejas y no lineales de los datos.

Objetivo de la Actividad

El objetivo de esta actividad es que los estudiantes desarrollen y apliquen modelos supervisados a un conjunto de datos específico, abordando un problema real en sus respectivas empresas. Los estudiantes investigarán, analizarán y compararán diferentes algoritmos supervisados, implementarán modelos, evaluarán su desempeño y presentarán sus hallazgos.

Estructura de la Actividad

Duración: 5 semanas

Materiales Necesarios

- Acceso a conjuntos de datos relevantes.
- Material de lectura sobre modelos supervisados (e.g., regresión lineal, árboles de decisión, máquinas de soporte vectorial).
- Herramientas para la recopilación y análisis de datos (e.g., Python, R).
- Herramientas para la visualización de datos (e.g., matplotlib, seaborn, ggplot2).

Semana 1: Investigación y Selección de Algoritmos

Acción: Investigar diferentes algoritmos de aprendizaje supervisado que se ajusten a la medición de la variable objetivo.

Resultado: Informe comparativo de los algoritmos investigados, destacando sus ventajas y desventajas.

Semana 2: Preparación de los Datos y Definición del Problema

Acción: Recopilar y preparar el conjunto de datos. Definir el problema específico a resolver mediante modelos supervisados.
Resultado: Documento con la descripción del conjunto de datos, el problema definido y los objetivos del modelo.

Semana 3 y 4: Implementación y Evaluación de Modelos

Acción: Implementar varios modelos supervisados utilizando Python o Orange, entrenar los modelos y evaluar su desempeño utilizando métricas adecuadas (e.g., precisión, recall, F1-score, AUC).
Resultado: Lista de modelos implementados y evaluaciones de desempeño.

Semana 5: Interpretación y Presentación de Resultados

Acción: Interpretar los resultados obtenidos, compararlos y seleccionar el modelo más adecuado. Crear visualizaciones que respalden los hallazgos y prepararse para presentar los resultados.
Resultado: Informe final con los resultados del modelo seleccionado, las visualizaciones y conclusiones.

Entrega Final

Al finalizar las 5 semanas, cada estudiante deberá presentar un informe completo que incluya todos los resultados y documentos desarrollados durante la actividad. Este informe debe contener:

1. Introducción y descripción del problema.
2. Informe de algoritmos supervisados investigados.
3. Descripción del conjunto de datos y definición del problema.
4. Resultados de la implementación y evaluación de modelos.
5. Interpretación de resultados y modelo seleccionado.
6. Conclusiones y recomendaciones.