CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

Curso
Machine Learning con Python

Trabajo Práctico Módulo 1

Objetivos

Aplicar los conocimientos adquiridos en el Módulo 1

Consignas

Ejercicio 1 (40 puntos)

El objetivo de este ejercicio es obtener un modelo de regresión lineal múltiple que describa el comportamiento de los datos en el archivo advertising.csv (data tomada de kaggle.com).

En este archivo de datos se registra las ventas (Sales) en relación con las siguientes variables

- Monto invertido en publicidad en TV (TV)
- Monto invertido en publicidad en Radio (Radio)
- Monto invertido en publicidad en Prensa (Newspaper)

Desarrolle un Jupyter Notebook que

1) (5 ptos) Cargue los datos del archivo advertising.csv suministrado en un dataframe de Pandas
2) (5 ptos) Realice un gráfico de dispersión de los valores de Sales vs los índices de los datos

3) (5 ptos) Utilice el comando train_test_split para separar los datos en conjuntos de entrenamiento y prueba. La fracción de datos de prueba debe estar entre 20% y 30%.
4) (10 ptos) Determine un modelo de regresión lineal múltiple utilizando los datos de entrenamiento. Este modelo debe estimar Sales a partir de las variables de entrada
5) (15 ptos) Evalúe el desempeño del modelo de regresión lineal obtenido en el apartado anterior. Para esto
    - Calcule las métricas Error Absoluto Medio, Error Cuadrático Medio, Puntuación R2, Puntuación de Varianza Explicada.
    - Realice un gráfico de dispersión de datos pronosticados (eje vertical) vs. datos de prueba (eje horizontal)
    - Realice un gráfico de línea de los errores porcentuales ($error_{\%}$ vs índices) del modelo sobre los datos de prueba

$$error_{\%} = \left| \frac{Y_{test} - \hat{Y}_{test}}{Y_{test}} \right| \times 100$$

    - Comente brevemente los resultados

    El modelo obtenido debe alcanzar una puntuación R2 por encima de 0.9 y los errores porcentuales deben estar por debajo de 30%

Ejercicio 2 (60 puntos)

El objetivo de este ejercicio es obtener diferentes modelos de clasificación utilizando los datos de vinos del “Wine recognition dataset” de Scikit-Learn, que puede obtener con la instrucción

from sklearn.datasets import load_wine

Este conjunto de datos contiene 178 registros, cada uno de los cuales consta de valores de trece (13) atributos numéricos y la clase de vino correspondiente (0, 1 y 2).

Desarrolle un Jupyter Notebook que

1) (5 ptos) Cargue los datos en un dataframe de Pandas

Centro de e-Learning SCEU UTN - BA. Medrano 951 2do piso
(1179) // Tel. +54 11 7078- 8073 / Fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

2) (10 ptos) Genere 2 datasets, uno con los datos originales y otro con los datos de entrada normalizados (restando la media y dividiendo por la desviación estándar). Utilice el comando train_test_split para separar ambos datasets en conjuntos de entrenamiento y prueba. La fracción de datos de prueba debe estar entre 20% y 30%. Si asociamos a los datos originales y normalizados índices del 0 al 177, los conjuntos de entrenamiento deben ser construidos de forma que los índices de los datos originales y normalizados presentes en ambos conjuntos de entrenamiento sean los mismos. De forma similar deben construirse los conjuntos de prueba
3) (10 ptos) Determine dos modelos de clasificación basados en regresión logística, uno con cada conjunto de entrenamiento. Evalúe el desempeño de cada modelo sobre el conjunto de prueba correspondiente. Para esto, determine la precisión (accuracy) y la matriz de confusión para cada modelo
4) (10 ptos) Determine dos modelos de clasificación basados en KNN, uno con cada conjunto de entrenamiento. Utilice el mismo K y la misma opción de weights ('uniform' o 'distance') para ambos modelos. Evalúe el desempeño de cada modelo sobre el conjunto de prueba correspondiente. Para esto, determine la precisión (accuracy) y la matriz de confusión para cada modelo
5) (10 ptos) Determine dos modelos de clasificación basados en SVM, uno con cada conjunto de entrenamiento. Utilice kernel de tipo lineal para ambos modelos. Evalúe el desempeño de cada modelo sobre el conjunto de prueba correspondiente. Para esto, determine la precisión (accuracy) y la matriz de confusión para cada modelo

   Todos los modelos (regresión logística, KNN y SVM) obtenidos con los datos de entrenamiento normalizados deben tener una precisión (accuracy) por encima del 95%
6) (5 ptos) Comente sobre el desempeño de los diferentes modelos

NOTA: Puede obtener más información sobre el dataset en
https://scikit-learn.org/stable/datasets/toy_dataset.html#wine-dataset

Formato de presentación:

- El trabajo puede realizarse individualmente o en grupos de 2 participantes
- Los participantes deben entregar un archivo comprimido que contenga los Jupyter notebooks desarrollados. El nombre del archivo comprimido debe tener el formato TPI_Apellidos. Por ejemplo, TPI_Canelon o TPI_Barreto_Canelon. El nombre de los Jupyter notebooks debe tener el formato TPI_Apellidos_Ejer#.ipynb, donde # debe reemplazarse por el número de ejercicio. Por ejemplo, TPI_Canelon_Ejer1.ipynb o TPI_Barreto_Canelon_Ejer1.ipynb
- El archivo comprimido debe generarse con la aplicación 7-zip, que puede descargarse gratuitamente desde https://www.7-zip.org/.

Fecha límite de entrega:

Nominal: 25/09/2024 - 23:59 hrs

Recuperatorio: 02/10/2024 - 23:59 hrs

Las fechas de entrega son inapelables, ya que están configuradas automáticamente en el Campus. La plataforma no permitirá que los participantes entreguen fuera de la fecha/hora indicada. Si no pueden hacerlo en la primera, podrán hacerlo en el recuperatorio.

Criterios de evaluación

- La calificación total del trabajo estará en función del número de consignas realizadas correctamente. Si alguna consigna no funciona de manera correcta o genera un error en el Jupyter notebook, se restarán puntos del total correspondiente a esa consigna

Centro de e-Learning SCEU UTN - BA. Medrano 951 2do piso
(1179) // Tel. +54 11 7078- 8073 / Fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

Bibliografía utilizada y sugerida

scikit-learn. Documentación oficial.
Disponible desde: URL: https://scikit-learn.org/stable/

NumPy. Documentación oficial.
Disponible desde: URL: https://numpy.org/

Pandas. Documentación oficial.
Disponible desde: URL: https://pandas.pydata.org/

Matplotlib. Documentación oficial.
Disponible desde: URL: https://matplotlib.org/