AWAKELAB

Plan Formativo: Ciencia de Datos	Nivel de Dificultad
Módulo 5: Aprendizaje supervisado	Medio
Tema: Naive bayes	

Intención del aprendizaje o aprendizaje esperado:
• 7. Elabora un modelo predictivo aplicando el algoritmo clasificador Bayesiano para resolver un problema de clasificación utilizando lenguaje Python
Ejercicios planteados
Con el fin de clasificar las distintas actividades que puede realizar una persona, se realizó un experimento a un grupo de 30 voluntarios dentro de un grupo de edad de 19 a 48 años. Cada persona realizó seis actividades (CAMINAR, CAMINAR POR LAS ESCALERAS, BAJAR LAS ESCALERAS, SENTARSE, DE PIE, RECOSTARSE) con un smartphone (Samsung Galaxy S II) conectado en su cintura. Usando el acelerómetro y giroscopio integrados en el smartphone, se tienen la siguiente data: -Aceleración triaxial del acelerómetro (aceleración total) y la aceleración corporal estimada. -Velocidad angular triaxial del giroscopio. -Variables de dominio de tiempo y frecuencia. -Su etiqueta de actividad.

AWAKELAB

-Un identificador del sujeto que realizó el experimento.

El conjunto de datos obtenido se ha dividido aleatoriamente en dos conjuntos (train y test), donde se seleccionó al 70% de los voluntarios para generar los datos de entrenamiento y al 30% los datos de la prueba.

Estos datos se encuentran en los siguientes links:

train: https://raw.githubusercontent.com/natjulian/Contribuci-n-Diplomado-Data-Science-UC-2021/main/Bases%20de%20datos%20Clases/Activity/train.csv

test: https://raw.githubusercontent.com/natjulian/Contribuci-n-Diplomado-Data-Science-UC-2021/main/Bases%20de%20datos%20Clases/Activity/test.csv

Para esta actividad, realice lo siguiente:

1. Cargue los set de datos de entrenamiento y prueba. ¿Cuáles son las dimensiones de estos set de datos?
2. Estudie la cantidad de registros que hay en cada actividad en el set de entrenamiento. Para esto, una opción es utilizar la función de pandas crosstab(index=columna, columns="count"). ¿Cómo se encuentran repartidas las Actividades? ¿Existe algún desbalance en las categorías? (realice un gráfico que acompañe su análisis) Comente.
3. Defina X_train y X_test como corresponda (omitiendo las variables que no corresponden a variables predictoras).
4. Estandarice el set de features de entrenamiento y prueba. Puede utilizar la función `StandardScaler`:

```
from sklearn.preprocessing import StandardScaler
```

AWAKELAB

```
x_stand = StandardScaler().fit_transform(x)
```

5. Defina y_train e y_test como corresponda (recuerde que debe estar codificada numéricamente). Para recodificar un vector o columna y de varias categorías a formato numérico puede utilizar:

```
from sklearn import preprocessing

encoder=preprocessing.LabelEncoder().fit(y)

y_new=encodertrain.transform(y)
```

6. Utilice la función GridSearchCV para realizar 10 validaciones cruzadas con la lista de parámetros definida anteriormente. Determine la mejor combinación de parámetros y obtenga el modelo con dichos parámetros.
7. ¿Cómo es la capacidad predictiva de este modelo en el set de prueba? Comente.

Caso

APRENDIZAJE DE MÁQUINA SUPERVISADO

Preguntas guía

Recursos Bibliográficos:

[1] Teorema de Bayes

https://economipedia.com/definiciones/teorema-de-bayes.html

[2] Naive Bayes

https://scikit-learn.org/stable/modules/naive_bayes.html

[3] Clasificador bayesiano

https://medium.com/datos-y-ciencia/algoritmos-naive-bayes-fudamentos-e-implementaci%C3%B3n-4bcb24b307f