AWAKELAB

Plan Formativo: Ciencia de Datos	Nivel de Dificultad
Módulo: Aprendizaje no Supervisado	Bajo / medio
Tema: PCA	

Intención del aprendizaje o aprendizaje esperado:

- Elabora un modelo predictivo utilizando técnicas de reducción dimensional para resolver un problema de aprendizaje de máquina

Ejercicios planteados

La base de datos *House_price.csv* contiene información de casas. La idea es obtener un modelo que permita predecir el valor de venta de una casa. Las variables son las siguientes:

- LotFrontage: Pies lineales de calle conectados a la propiedad
- LotArea: Tamaño del lote en pies cuadrados
- MasVnrArea: Área de revestimiento de mampostería en pies cuadrados
- BsmtFinSF1: Pies cuadrados del sótano terminados tipo 1
- BsmtFinSF2: Pies cuadrados con del sótano acabados tipo 2
- BsmtUnfSF: Pies cuadrados sin terminar de área del sótano
- TotalBsmtSF: Total de pies cuadrados de área del sótano
- 1stFlrSF: Pies cuadrados del primer piso

AWAKELAB

- 2ndFlrSF: Pies cuadrados del segundo piso
- LowQualFinSF: Pies cuadrados con acabado de baja calidad (todos los pisos)
- GrLivArea: Pies cuadrados de área habitable sobre el nivel (suelo)
- GarageArea: Tamaño del garaje en pies cuadrados
- WoodDeckSF: Área de la plataforma de madera en pies cuadrados
- OpenPorchSF: Área del porche abierto en pies cuadrados
- EnclosedPorch: Área del porche cerrado en pies cuadrados
- 3SsnPorch: área de porche de tres estaciones en pies cuadrados
- ScreenPorch: Área del porche de la pantalla en pies cuadrados
- PoolArea: Área de la piscina en pies cuadrados
- MiscVal: Valor de la función miscelánea
- **SalePrice: el precio de venta de la propiedad en dólares. Esta es la variable objetivo que se quiere modelar.**

a) Cargue la base de datos, ¿hay alguna columna que no sea útil para el análisis?. Revise si existen casos faltantes en la base de datos, según la descripción de las variables, ¿qué pudiera significar un NA? Tome decisiones de qué hacer si existen casos faltantes. Si desea imputar los valores nulos de alguna variable por un valor puede utilizar data["columna"].fillna('valor', inplace = True).
b) b) La variable a modelar corresponde al precio de venta de la casa en dólares. En base a los modelos que hemos aprendido en clases, ¿qué modelo(s) podría(n) ser de utilidad para predecir el precio de venta de casas mediante las variables predictoras?
c) Vamos a implementar un modelo de regresión lineal para modelar el precio de venta de casas. Considerando que en la base de datos

AWAKELAB

tenemos varias variables que dependen de otras, ¿cuál pudiera ser el problema que surgiría en este contexto? ¿por qué esto es un problema? Discuta.

d) Realice un análisis de la correlación entre las variables. Comente, ¿existen variables muy correlacionadas entre sí? ¿Por qué cree que ocurre?

e) Suponga que este es un problema de gran volumen, donde tenemos muchas columnas y registros. Cuando existen variables muy correlacionadas el costo computacional para obtener el modelo puede ser muy elevado. Obtenga el tiempo de procesamiento para obtener una regresión lineal. Para esto, podemos obtener la regresión lineal utilizando la función LinearRegression de sklearn.linear_model y calcular el tiempo de procesamiento de la siguiente forma:

```
from datetime import datetime
start=datetime.now()
[aquí va el código del modelo]
print(datetime.now()-start) #esto entrega el tiempo que demoró en correr el modelo.
```

f) Obtenga las componentes principales utilizando las variables estandarizadas, si desea explicar un 80% de la variabilidad, ¿cuántas componentes debería elegir? Argumente.

g) Corra el modelo de regresión lineal con las componentes principales como variables predictoras. Vuelva a calcular el tiempo de procesamiento. ¿En cuál caso suele demorarse menos? ¿Qué ganamos al realizar componentes principales? Concluya.

Caso

Preguntas guía

- Reducción de dimensionalidad

AWAKELAB

- Regresion lineal

Recursos Bibliográficos:

Referencias

[1] Técnicas de reducción de dimesionalidad
https://topbigdata.es/6-algoritmos-de-reduccion-de-la-dimensionalidad-con-python/#:~:text=La%20reducci%C3%B3n%20de%20la%20dimensionalidad%20es%20una%20t%C3%A9cnica%20de%20preparaci%C3%B3n,de%20entrenar%20un%20modelo%20predictivo.

[2] Reducción de dimensionalidad
https://pharos.sh/reduccion-dimensional-en-python-con-scikit-learn/

[3] PCA
https://www.cienciadedatos.net/documentos/35_principal_component_analysis

[4] LDA
https://economipedia.com/definiciones/analisis-discriminante.html