El conjunto de datos California Housing

Guillermo Ruiz

En esta actividad se trabajará con el conjunto de datos California Housing. Se reducirá la dimensión y luego se aplicará DBSCAN para agrupar los datos.

Para graficar se debe usar la función `scatter` con el parámetro `cmap='Spectral'` para mostrar los colores.

Con el siguiente código se puede leer los datos.

```
import matplotlib.pyplot as plt
from sklearn.datasets import fetch_california_housing
import umap
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.cluster import DBSCAN
```

```
california_housing = fetch_california_housing(as_frame=True)
df = california_housing.data
y = california_housing.target
df.shape, y.shape
```

```
((20640, 8), (20640,))
```

Que contiene las siguientes características:

```
print(california_housing.feature_names)
```

```
['MedInc', 'HouseAge', 'AveRooms', 'AveBedrms', 'Population', 'AveOccup', 'Latitude', 'Longit
```

Con este conjunto de datos se deberán proyectar los datos en dos dimensiones usando el precio y como color. Se usarán los métodos

- PCA
- UMAP

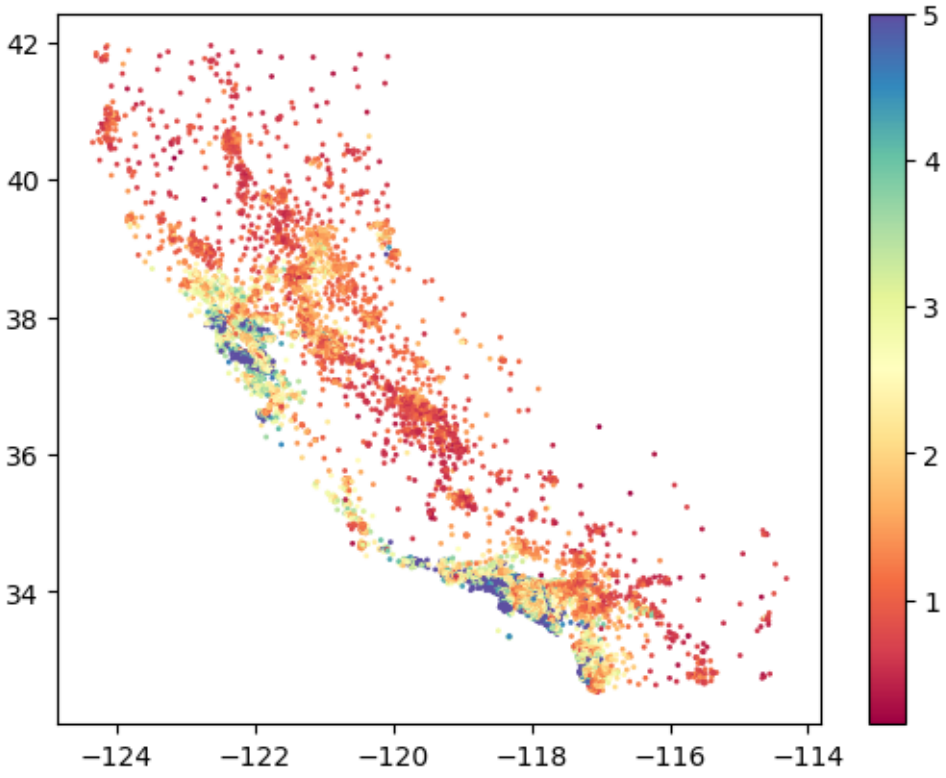
Primero veamos la descripción de los datos.

```
df.describe()
```

	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude
count	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000	20640.00
mean	3.870671	28.639486	5.429000	1.096675	1425.476744	3.070655	35.63186
std	1.899822	12.585558	2.474173	0.473911	1132.462122	10.386050	2.135952
min	0.499900	1.000000	0.846154	0.333333	3.000000	0.692308	32.54000
25%	2.563400	18.000000	4.440716	1.006079	787.000000	2.429741	33.93000
50%	3.534800	29.000000	5.229129	1.048780	1166.000000	2.818116	34.26000
75%	4.743250	37.000000	6.052381	1.099526	1725.000000	3.282261	37.71000
max	15.000100	52.000000	141.909091	34.066667	35682.000000	1243.333333	41.95000

Como tenemos las coordenadas (Longitude y Latitude), podemos poner los puntos en un mapa.
1. Graficar los puntos usando Longitude vs Latitude en los ejes. Usar el precio ycomo color de los puntos. Note que se pueden ver las ciudades de Los Angeles, San Francisco, San Diego y Sacramento.

```
plt.scatter(df['Longitude'],df['Latitude'], s=1, c=y, cmap='Spectral')
plt.colorbar()
plt.show()
```



PCA

2. Usar PCA para bajar las dimensiones a dos y graficar los datos usando `y` como color. Primero aplicar `StandardScaler` para estandarizar los datos (media cero y desviación estandar 1).

Como podemos ver, se ven muy juntos los datos.

UMAP

3. Usar UMAP para visualizar los datos. Graficar la proyección usando `y` como color.

Como podemos ver, se agrupan los datos en clusters.

4. Usar DBSCAN para separarlos (la reducción de dimensión) y graficamos los clusters.
5. Volver a poner los puntos originales en un mapa pero el color será el cluster al que pertenece.