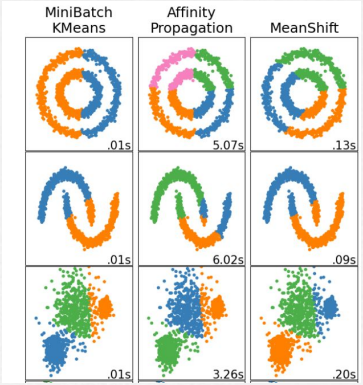
INTRODUCCIÓN A KMEANS

OpenWebinars

QUÉ ES CLUSTERIZACIÓN?

- Un campo que busca agrupar los datos en unos “clusters” compuestos por datos similares
- En vez de predecir un “target”, buscamos asignar este “target” a los datos
- Aunque no tengamos un target específico, construimos los clusters con un objetivo



KMEANS ES EL MODELO MÁS COMÚN

- Kmeans crea unos grupos centrados en la media de los datos que pertenecen al grupo
- Nosotros tenemos que elegir el número de grupos que existen y el algoritmo busca los centroides
- La implementación en sklearn es muy escalable y sigue el patrón típico de un modelo de aprendizaje supervisado

```
model_3 = KMeans(n_clusters=3, random_state=0)
model_3.fit(user_stats[X_variables])
user_stats['predictions_kmeans_3'] = model_3.predict(user_stats[X_variables])

user_stats.predictions_kmeans_3.value_counts()
```

```
0    785
2     51
1     11
Name: predictions_kmeans_3, dtype: int64
```

```
model_10 = KMeans(n_clusters=10, random_state=0)
model_10.fit(user_stats[X_variables])
user_stats['predictions_kmeans_10'] = model_10.predict(user_stats[X_variables])

user_stats.predictions_kmeans_10.value_counts()
```

```
2    335
6    175
8    151
0     82
7     43
9     20
5     17
1     14
3      7
4      3
Name: predictions_kmeans_10, dtype: int64
```

Nos permite sklearn ver donde estan los centroides

```
pd.DataFrame(model_3.cluster_centers_, columns=X_variables)
```

	followers	following	likes	media
0	5.808735e+06	12550.436943	8038.670064	7278.182166
1	8.820971e+07	98593.818182	6443.363636	4792.909091
2	3.911104e+07	1848.549020	2250.098039	14051.294118

```
pd.DataFrame(model_10.cluster_centers_, columns=X_variables)
```

	followers	following	likes	media
0	1.288341e+07	6959.084337	8838.759036	4229.289157
1	5.455245e+07	1087.285714	4605.071429	22738.071429
2	4.065838e+06	6580.238806	9064.191045	6129.005970

EXPLORAMOS EL KMEANS

Expandimos nuestro ejemplo de antes y profundizamos

RESUMEN: EL MODELO DE KMEANS

- Algoritmo clásico y muy robusto que depende de nuestra especificación del número de clusters
- El entrenamiento es sencillo y tenemos la opción de “mini batch kmeans” para entrenar más rápido
- Podemos predecir sobre puntos nuevos para agruparlos según los clusters encontrados en el entrenamiento