UNIVERSIDAD PERUANA DE CIENCIAS APLICADAS
CIENCIAS DE LA COMPUTACIÓN

MACHINE LEARNING
Laboratorio k-means
(Primer Semestre del 2019)

Objetivos de aprendizaje:

- Generar cluster usando k-means.

---

1. Actividad en Weka

1.1. Food (20 minutos)

Un entrenador físico está analizando la dieta que ingiere uno de los deportista que entrena. Al entrenador le gustaría poder clasificar los alimentos en grupos homogéneos, los más similares posibles, para de esta manera, analizar la composición de los alimentos. Se le pide a Ud. que le ayude al realizar esta tarea usando el conjunto de datos `food.arff`. Para este ejercicio use el algoritmo `SimpleKMeans`. Los atributos de este conjunto de datos se describen a continuación

- name
- energy
- protein
- fat
- calcium
- iron

Preguntas de discusión

- Describa al modelo. ¿Es un buen modelo? Justifique su respuesta.
- ¿Existen diferencias cuando normaliza los datos?

1.2. Seed (20 minutos)

El conjunto de datos `seeds_dataset.arff` contiene mediciones de propiedades geométricas de granos pertenecientes a tres variedades diferentes de trigo. Este conjunto de datos contiene siete atributos numéricos los cuales se describen a continuación.

- Atributo1. area A
- Atributo2. perimeter P
- Atributo3. compactness C $= 4 * pi * A/P^2$
- Atributo4. length of kernel
- Atributo5. width of kernel
- Atributo6. asymmetry coefficient

- Atributo7. length of kernel groove

Se le pide que analice este conjunto de datos agrupando los objetos más similares entre sí. Para este ejercicio use el algoritmo `SimpleKMeans`.

Preguntas de discusión

- Describa al modelo. ¿Es un buen modelo? Justifique su respuesta.
- ¿Existen diferencias cuando normaliza los datos?

2. Actividad en RapidMiner

2.1. Food (20 minutos)

Un entrenador físico está analizando la dieta que ingiere uno de los deportista que entrena. Al entrenador le gustaría poder clasificar los alimentos en grupos homogéneos, los más similares posibles, para de esta manera, analizar la composición de los alimentos. Se le pide a Ud. que le ayude al realizar esta tarea usando el conjunto de datos `food.csv`. Use el componente `k-Means` para resolver el problema.

Preguntas de discusión

- Describa al modelo. ¿Es un buen modelo? Justifique su respuesta.
- ¿Existen diferencias cuando normaliza los datos?
- Los datos son consistentes con Weka.

2.2. Seed (20 minutos)

El conjunto de datos `seeds_dataset.csv` contiene mediciones de propiedades geométricas de granos pertenecientes a tres variedades diferentes de trigo. Se le pide que analice este conjunto de datos agrupando los objetos más similares entre sí. Use el componente `k-Means` para resolver el problema.

Preguntas de discusión

- Describa al modelo. ¿Es un buen modelo? Justifique su respuesta.
- ¿Es necesario trabajar los valores nulo? Si requiere reemplazar valores nulo puede usar el componente `Replace Missing Values`
- ¿Existen diferencias cuando normaliza los datos?
- Los datos son consistentes con Weka. ¿El algoritmo `SimpleKMeans` permite atributos nulos?

3. Actividad en Python

3.1. Food (10 minutos)

Obtenga los centroides del modelo en Python. Para esto puede usar el script que sigue a continuación.

```
from sklearn.cluster import KMeans

kmeans = KMeans(n_clusters=2)
modelo = kmeans.fit(X)
print(modelo)

centroides = modelo.cluster_centers_
print(centroides)
```

3.2. Seed (10 minutos)

De forma similar obtenga los centroides del modelo en Python.

Preguntas de discusión

- ¿Es posible trabajar los valores nulo?
- ¿Existen diferencias cuando normaliza los datos?
- Los datos son consistentes con Weka y RapidMiner.

Monterrico, 6 de junio de 2019.