Elements of Machine Learning 2024

Ejercicio

22.marzo.2024

1. Implementar el algoritmo de K-means para que funcione con cualquier matriz de datos X de tamaño $n \times d$ (n es el número de datos u observaciones, d es la dimensión de los datos).

   - Como parámetros, el algoritmo debe recibir la matriz de datos X, y el número de clústers requerido k.
     (También conviene agregar algunos criterios de paro para que el algoritmo termine cuando alcanza la convergencia. Por ejemplo, un número máximo de iteraciones, y una toleracia `tol` para contrastar la diferencia entre el vector de centroides de la iteracion anterior y el vector de centroides actual

     $$error = ||old\_centroids - centroids||_2$$

     esto puede calcularse usando la norma euclideana o similar.)
   - Como salida, el algoritmo debe devolver la lista o vector de etiquetas, de tamaño n.
   - También como salida, el algoritmo debe devolver la lista o vector de centroides, de tamaño $k \times d$.

   Usar la distancia euclideana de los vectores de tamaño d para calcular las distancias.

   Aplique su algoritmo con los datos del conjunto Iris a continuación para verificar la funcionalidad de su implementación.

2. Aplicar el algoritmo de K-means al conjunto de datos `k-means_data.csv`, para diferentes valores de k.

   En cada caso, calcular el error MSE (*mean squared error*), y hacer una gráfica de "codo" para ilustrar el número ideal de clústers en el conjunto de datos.