Instituto Tecnológico de Costa Rica

Introducción al reconocimiento de patrones

Tarea 5

Profesor: José Pablo Alvarado

Integrantes:

Sebastián Vargas Zúñiga

Alonso Vega Badilla

I Semestre

Entrega: 30/06/2020

Criterios de evaluación del algoritmo k-medias

- **Índice aleatorio ajustado (ARI):**

Este criterio evalúa la similitud entre dos asignamientos ignorando las permutaciones pero con una posibilidad de normalización, el índice se calcula como el número de pares de elementos en el mismo set en los datos verdaderos y en k-means más el número de pares de elementos en diferentes sets en los datos verdaderos y en k-means dividido entre el número total de pares posibles pero este valor se ajusta aplicándole una especie de normalización donde se le resta la media y se divide entre el valor máximo menos la media. En este caso de desea un valor de 1, de cero para abajo es malo.

- **Puntajes basados en información mutua:**

Este criterio mide que tanto coinciden las predicciones del k-means respecto a las etiquetas que ya tendrían los datos verdaderos, esto se logra agarrando dos conjuntos de etiquetas con n datos para ambas etiquetas y para un conjunto de particiones, a estos dos conjuntos de etiquetas se les saca la probabilidad de que un dato aleatorio se encuentre tanto en una etiqueta U_i(una etiqueta especifica del primer conjunto de etiquetas) y en V_j (una etiqueta especifica del segundo conjunto de etiquetas) y se multiplica por el logaritmo de (esta misma probabilidad dividida entre (la probabilidad individual de que el dato pertenezca a la clase U_i por la probabilidad de que pertenezca a la clase V_j)) y se suma para cada combinación i,j. Con esto obtenemos la información mutua pero este valor se normaliza sacándole el valor esperado y obteniendo la entropía para ambos sets de datos la cual se obtiene mediante la suma para todo i o j de la multiplicación de la probabilidad de que un dato pertenezca a la clase i o j por el logaritmo de esta misma probabilidad, el AMI se obtiene entonces como la información mutua menos el valor esperado de la información mutua dividido entre la media del valor debido a ambas entropías menos el valor esperado de la información mutua.

- **Completeness:**

Es una métrica intuitiva a partir del análisis de entropía, este criterio asume que todos los datos de una clase dada pertenecen a una mismo cluster, su valor se encuentra entre 0 a 1 y se espera un valor alto. Se calcula de la siguiente forma:

$$c = 1 - \frac{H(K|C)}{H(K)}$$

Donde H de K dado C es la entropia condicional debida a las asignaciones de los clusters dadas ciertas clases y H de K es la entropia de los cluster solita.

Homogeneidad: Es una métrica intuitiva a partir del análisis de entropía, este criterio asume que los datos en un cluster solo pertenecen a una clase, su valor se encuentra entre 0 a 1 y se espera un valor alto. Se calcula de la siguiente forma:

$$h = 1 - \frac{H(C|K)}{H(C)}$$

Donde H de C dado K es la entropía condicional debida a las asignaciones de las clases dados los clusters y H de C es la entropía de las clases solita.

- **V-meassure:** a este criterio se le conoce como la media armónica y se calcula a partir de los criterios anteriores de la siguiente forma:

```
v = (1 + beta) * homogeneity * completeness
     / (beta * homogeneity + completeness)
```

Donde beta se utiliza para ajustar el peso que tiene la homogeneidad donde normalmente es 1 pero se puede alterar para lograr dicho efecto.

- **Silhouette Coefficient:** A diferencia de los criterios mostrados previamente este no requiere conocer las etiquetas de los datos verdaderos por lo que es muy util ya que normalmente esto no se conoce entre mayor sea el coeficiente mejor definidos se consideran los clusters y este coeficiente se difine para cada muestra. Para una sola muestra se define mediante la siguiente ecuacion:

$$s = \frac{b - a}{max(a, b)}$$

Donde “a” es la distancia media entre la muestra y los demás puntos en la misma clase y “b” es la distancia media de una muestra y los demás puntos del siguiente cluster más cercano a esta muestra. La medida va de –1 a 1 y si ronda el cero es que hay clusters superpuestos.

Métodos de inicialización

- **K-means++:**

Agarra los dos datos más alejados entre sí y aplica un algoritmo donde busca los k centroides que estén a una distancia máxima entre sí.

- **Random:**

Inicializa los centroides como valores aleatorios del set de datos.

- **PCA based:**

Se usa un algoritmo de PCA para disminuir el número de dimensiones a las más importantes y en esta nueva base se seleccionan los centroides en esta nueva base con algún otro método de inicialización.

- Existe otro método donde uno puede decirle al algoritmo cuales centroides utilizar, esto se utiliza cuando uno sabe más o menos por donde puede andar los centroides con el fin de acelerar el proceso de aprendizaje.

Resultados del código:

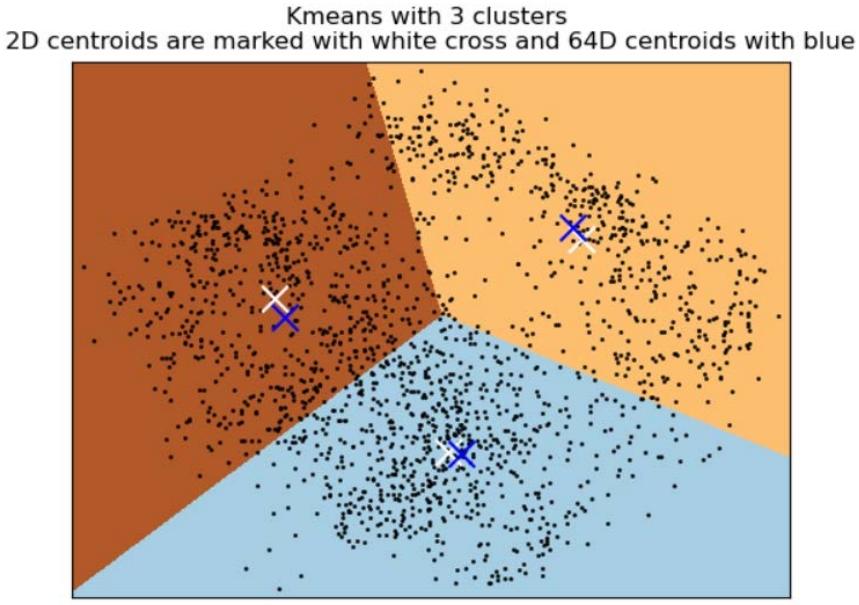

Figura 1. Resultado de algoritmo kmeans con 3 clusters.

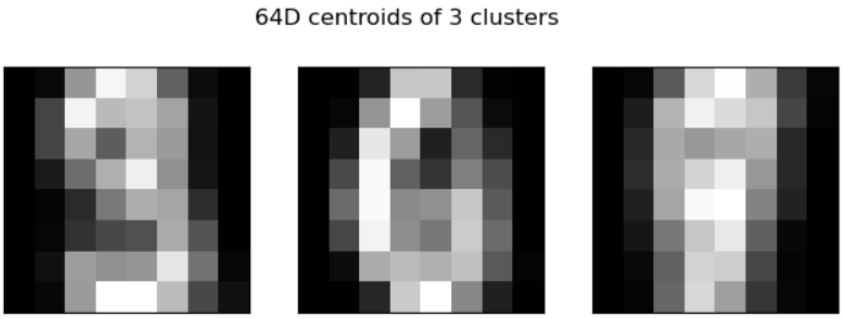

Figura 2. Centroides obtenidos con algoritmo kmeans con 3 clusters.

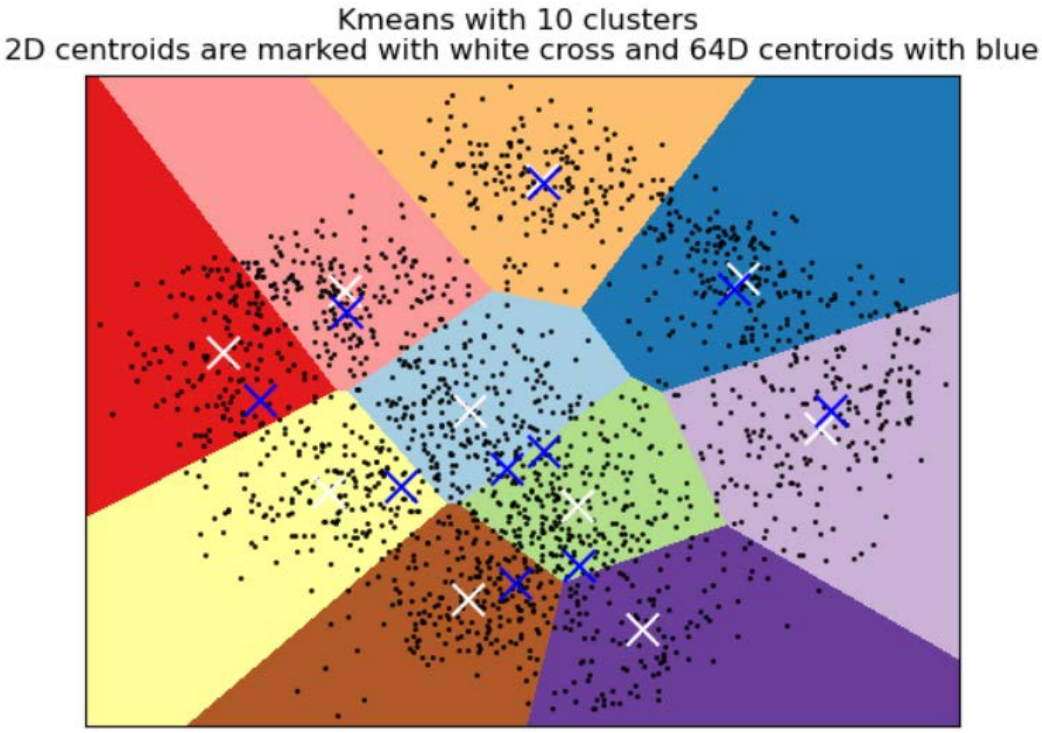

Figura 3. Resultado de algoritmo kmeans con 10 clusters.

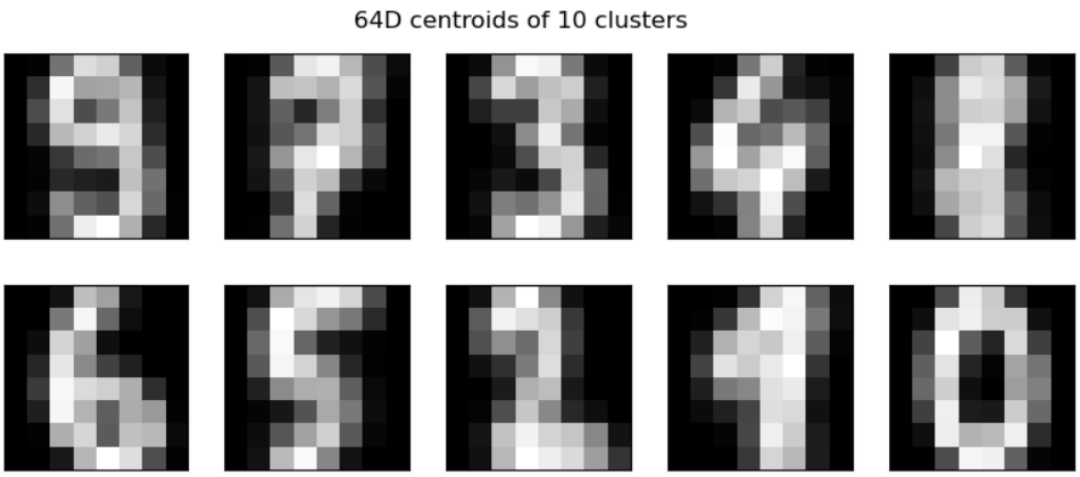

Figura 4. Centroides obtenidos con algoritmo kmeans con 10 clusters.

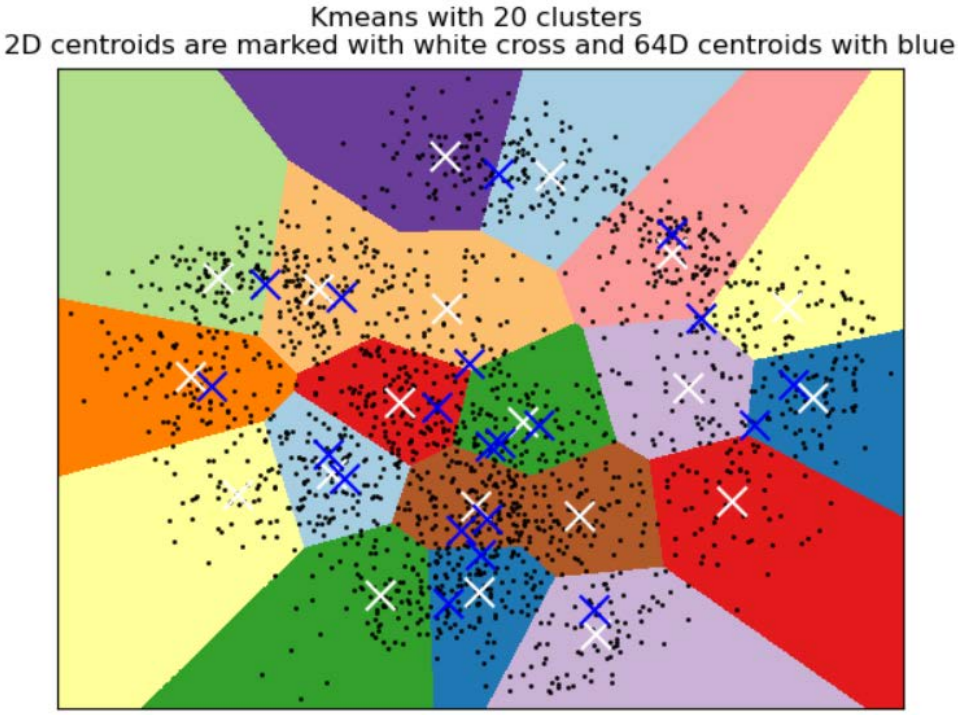

Figura 5. Resultado de algoritmo kmeans con 20 clusters.

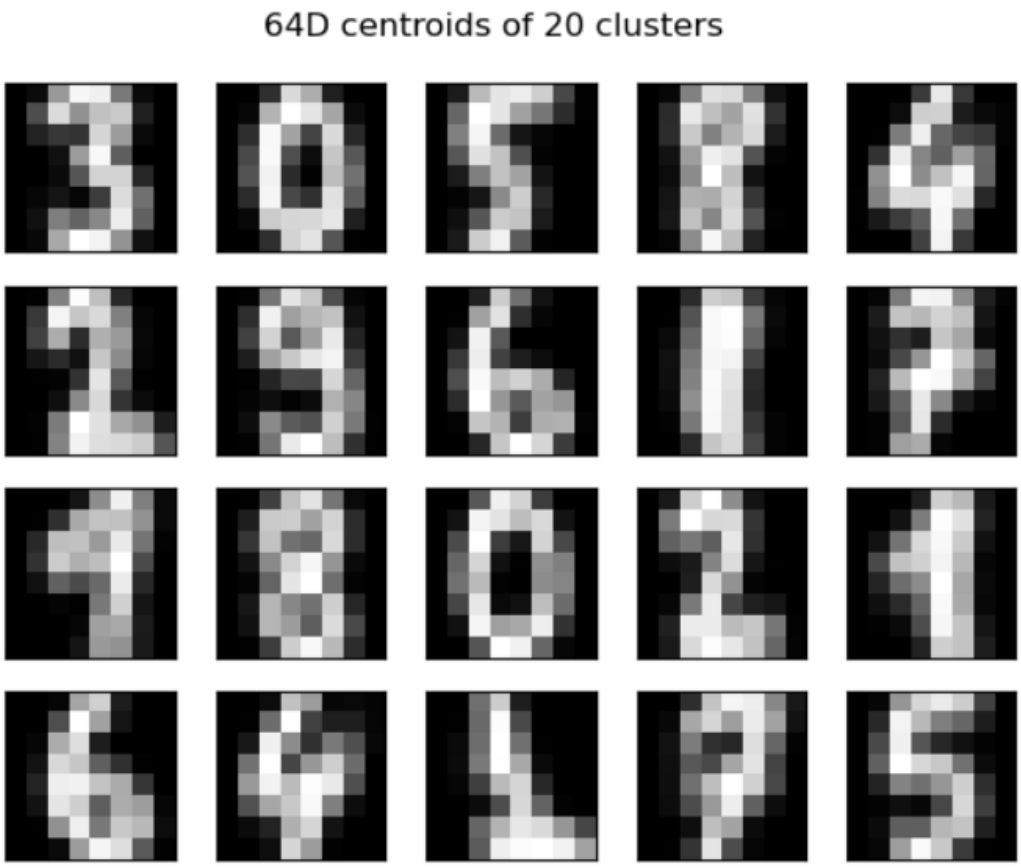

Figura 6. Centroides obtenidos con algoritmo kmeans con 20 clusters.