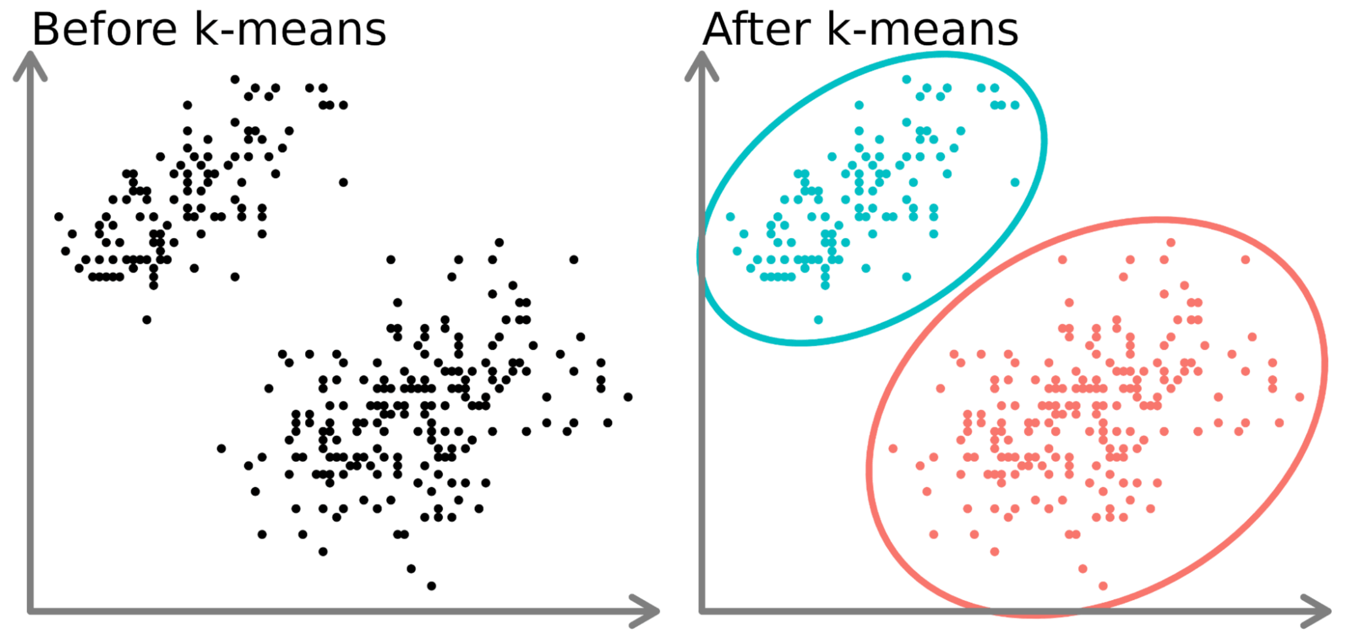


Aprendizaje no Supervisado y Clustering
Partitional Clust.: Kmeans

PhD(e). Jonnatan Arias Garcia – jonnatan.arias@utp.edu.co – jariasg@uniquindio.edu.co

PhD. David Cardenas peña - dcardenasp@utp.edu.co

PhD. Hernán Felipe Garcia - hernanf.garcia@udea.edu.co

Definiciones

- **Aprendizaje no supervisado.** En aprendizaje no supervisado no se cuenta con información sobre la variable de salida.

- Existen diferentes tipos de aprendizaje no supervisado: agrupamiento, estimación de densidad, y reducción de dimensionalidad.

- A continuación se estudia el problema de agrupamiento.

Algoritmo de las K-medias (I)

- Se busca identificar grupos de datos en un espacio multidimensional.

- Un grupo se puede entender como un conjunto de datos cuya distancia entre sí es pequeña comparada con la distancia a los puntos por fuera del grupo.

- Se supone un conjunto de vectores $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ en $\mathbb{R}^D$.

- Se introduce un conjunto de K vectores $\{\boldsymbol{\mu}_k\}_{k=1}^K \in \mathbb{R}^D$.

- Cada vector $\boldsymbol{\mu}_k$ es un prototipo asociado al k-ésimo grupo.

Algoritmo de las K-medias (II)

- Encontrar una asignación de los datos observados $\mathbf{X}$ a los K grupos.

- También se busca encontrar el conjunto de vectores $\boldsymbol{\mu}_k$ tal que se minimice la suma de los cuadrados de las distancias entre cada punto y su $\boldsymbol{\mu}_k$ más cercano.

$\mu_k \rightarrow Pto.centro\ o\ medio$

- Se define una *medida de distorsión*

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{n,k} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2,$$

donde $r_{n,k}$ es una variable binaria que indica a cuál de los K grupos se asigna el vector de observación $\mathbf{x}_n$.

- **Objetivo:** encontrar valores de $\{r_{n,k}\}$ y $\{\boldsymbol{\mu}_k\}$ que minimicen J.

Algoritmo de las K-medias (III)

- Lo anterior se puede lograr mediante un proceso iterativo de dos pasos.
  - Se escogen los $\boldsymbol{\mu}_k$ y se minimiza J con respecto a los $\{r_{n.k}\}$ manteniendo los $\boldsymbol{\mu}_k$ fijos.
  - Se minimiza J con respecto a los μ_k manteniendo los $r_{n,k}$ fijos.

- Los dos pasos se repiten hasta lograr la convergencia.

- El primer paso se consigue seleccionando los $r_{n,k}$ como

$$r_{n,k} = \begin{cases} 1, & \text{si } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0, & \text{de otra forma.} \end{cases}$$

- Esto debido a que J es una función lineal de $r_{n,k}$, y los $\mathbf{x}_n$ son independientes.

Algoritmo de las K-medias (IV)

- En el segundo paso se obtiene la derivada de J con respecto a $\boldsymbol{\mu}_k$, y se iguala a cero,

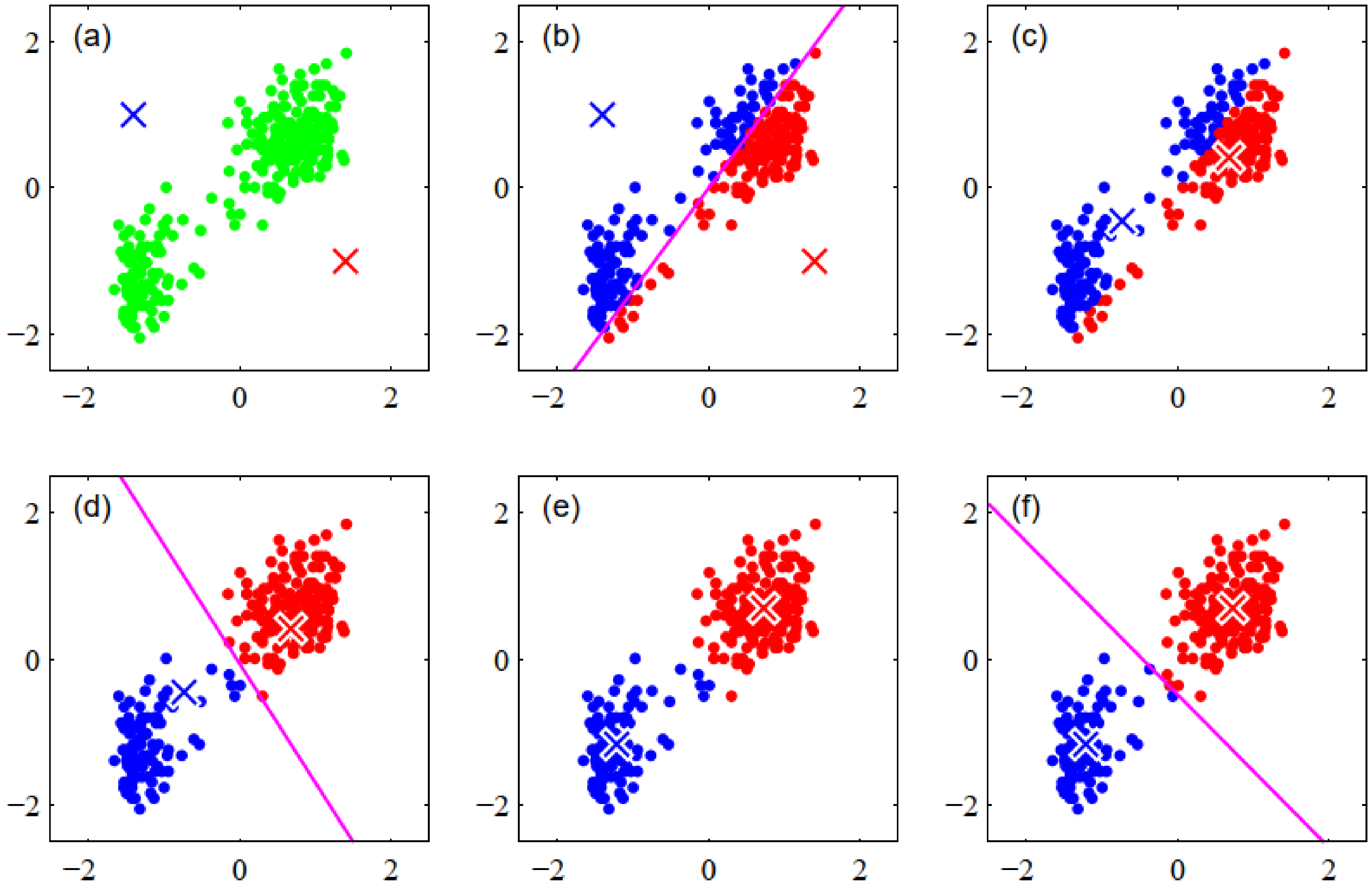
$$2\sum_{n=1}^{N} r_{n,k}(\mathbf{x}_n - \boldsymbol{\mu}_k) = 0.$$

- Despejando se obtiene,

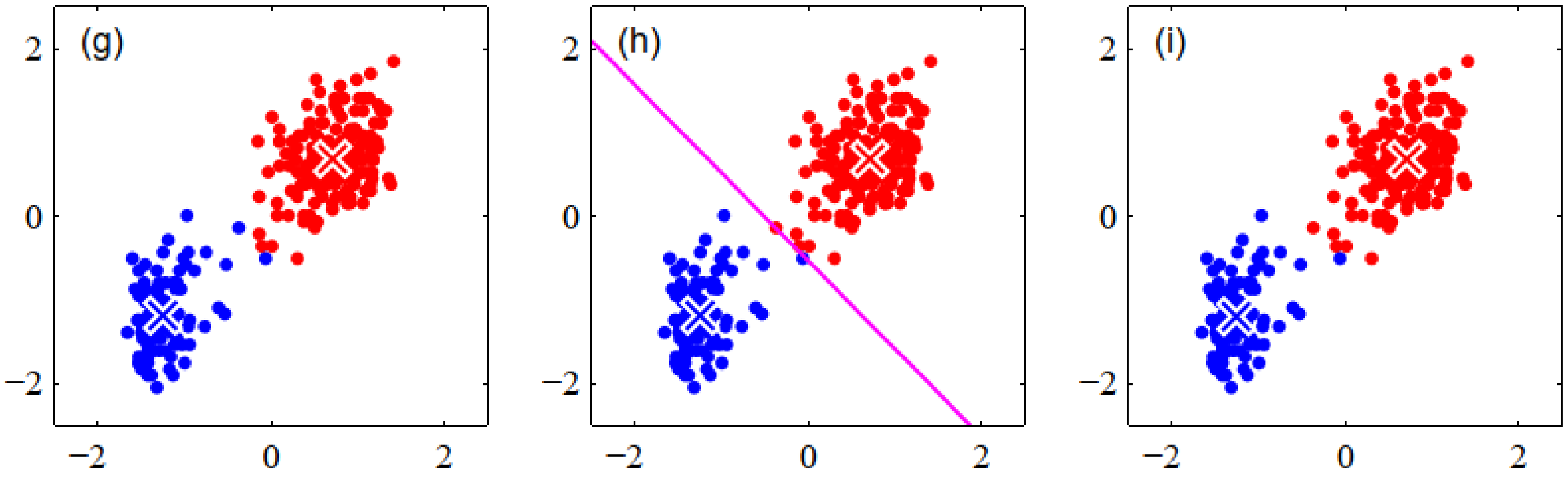
$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} r_{n,k}\mathbf{x}_n}{\sum_{n=1}^{N} r_{n,k}}.$$

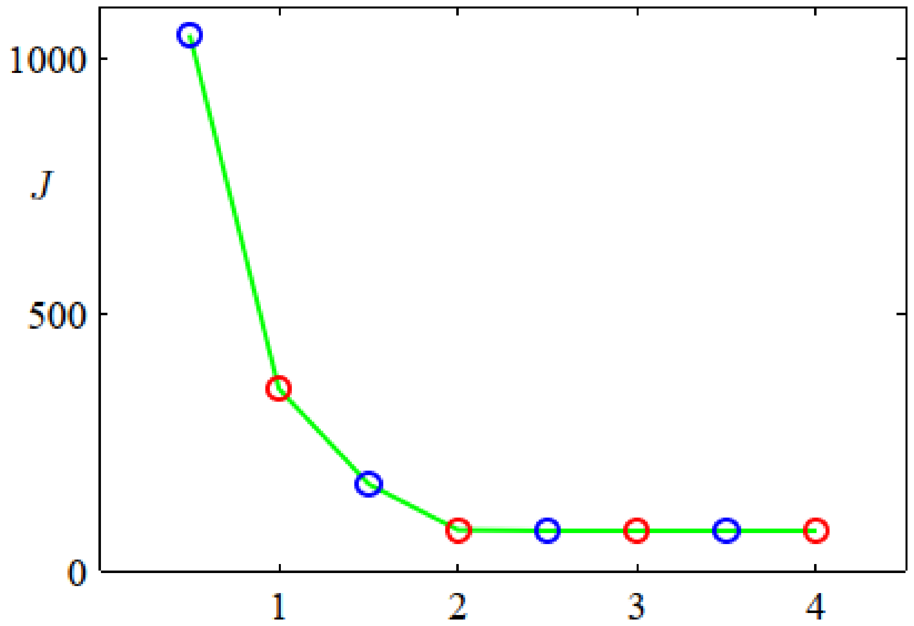
- Nótese que el denominador es igual al número de puntos asignados al grupo k.
- Igualmente $\boldsymbol{\mu}_k$ es la media de los datos $\mathbf{x}_n$ asignados al grupo k.
- Las dos fases de asignación de datos y cálculo de las medias se repiten hasta que no existan cambios en la asignación de grupos.

Algoritmo de las K-medias: ejemplo (I)



Algoritmo de las K-medias: ejemplo (II)





Algoritmo de las K-medias: otros espacios

- La distancia Euclidiana puede reemplazarse por una medida de disimilaridad $\mathcal{V}(\mathbf{x}, \mathbf{x}')$ que dependa de la aplicación y datos específicos.

- En este caso, la función de costo está dada como

$$\widetilde{J} = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{n,k} \mathcal{V}(\mathbf{x}_n, \boldsymbol{\mu}_k).$$

Aplicación: segmentación de imágenes (I)

- El objetivo de la segmentación es dividir una imagen en regiones que tengan una apariencia visual razonablemente homogénea.

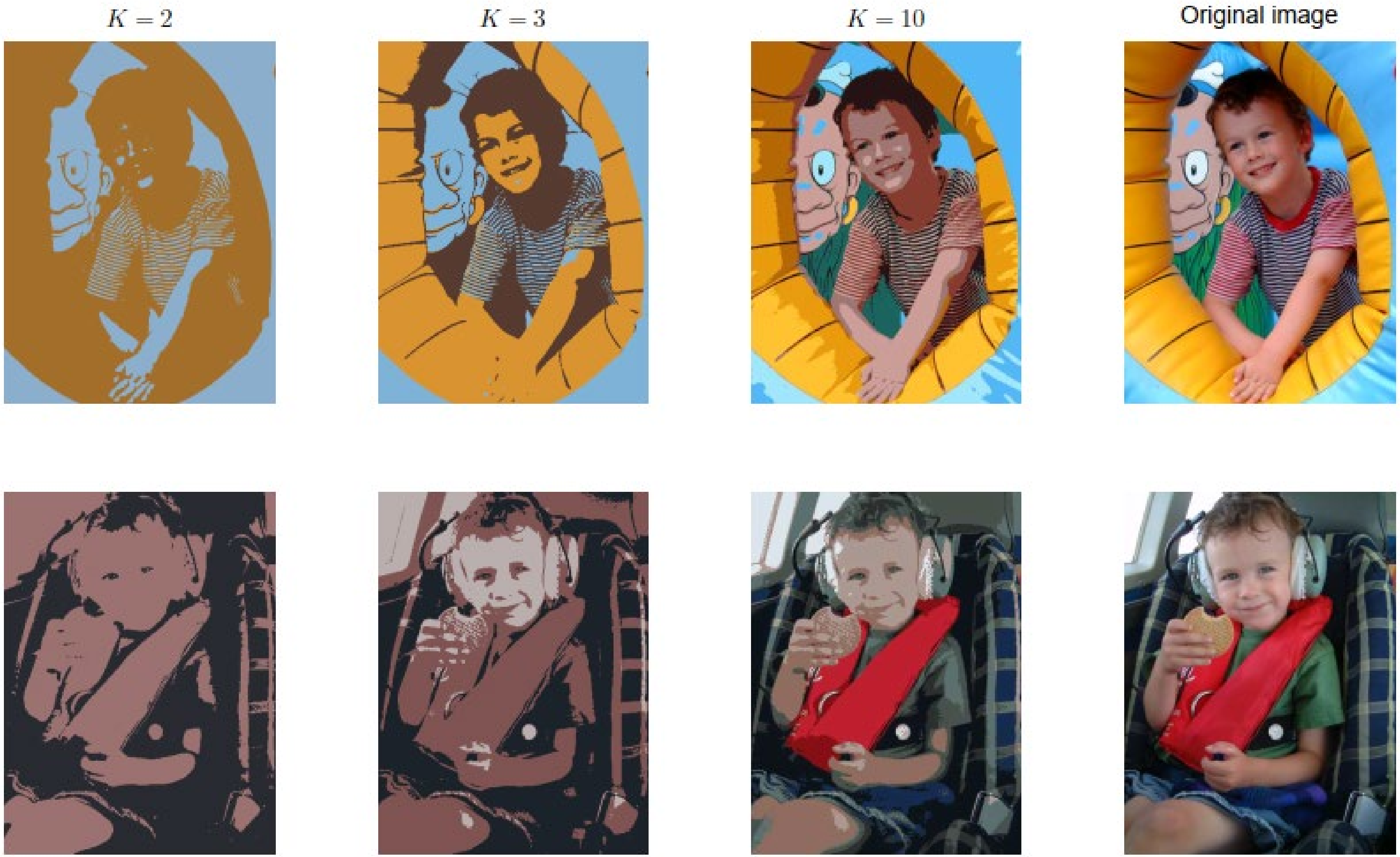
- Esas regiones suelen corresponder a objetos o partes de objetos.

- En esta aplicación, cada pixel se representa por un vector de intensidad [R,G,B], donde cada variable toma valores entre 0 y 1.

- Se usa agrupamiento por k-medias para diferentes valores de k.

- La imagen se redibuja cambiando el valor [R,G,B] de cada punto por el valor [R,G,B] dado por el centro μ al que ese punto ha sido asignado.

Aplicación: segmentación de imágenes (II)



Aplicación: compresión de imágenes (I)

- Para los N datos, se almacena únicamente la identidad del grupo al que pertenece cada dato.

- También se almacenan los valores de los centros μ_K.

- Si se transmitiera la imagen en codificación [R,G,B] con 8 bits de precisión, para transmitir la imagen completa se necesitarían

$$24 \times N \text{ bits} .$$

- Si se corre primero K-medias sobre la imagen, la información a transmitir consistiría en la identidad del grupo al que pertenece cada pixel ($\log_2 K$ bits), más la codificación [R,G,B] de los K centros

$$24 \times K + N \log_2 K \text{ bits} .$$

Aplicación: compresión de imágenes (II)

- En el ejemplo anterior, las imágenes tienen dimensiones de 240x180 píxeles.
- Esto da un valor de $N = 43200$ muestras.
- Transmitir la imagen completa implicaría transmitir

$$24 \times N \text{ bits } = 1.036.800 \text{ bits.}$$

- Transmitir haciendo K-medias primero implicaría transmitir ($K = 2$),

$$24 \times K + N \log_2 K \text{ bits } = 43248 \text{ bits.}$$