

**Tarea 06 - Reconocimiento de patrones**  
**Giovanni Gamaliel López Padilla**

## Problema 01

Supongamos que  $(X, Y)$  son variables aleatorias discretas con la siguiente distribución conjunta:

	X=1	X=2	X=3	X=4
Y=0	0.1	0.05	0.05	0.15
Y=1	0.12	0.1	0.25	0.18

Queremos predecir  $Y$  en base del valor observado para  $X$ .

- Calcula el clasificador Bayesiano Optimo si equivocarse de categoría tiene costo 1 y no equivocarse tiene costo 0. ¿Cuál es el costo (error) promedio para este clasificador?
- Calcula el clasificador Bayesiano Optimo si clasificar una observación mal cuando el verdadero valor es  $Y = 1$  tiene un costo 3 y en el otro caso tiene costo 2.

Para una  $x$  fija buscamos la asignación  $\hat{Y}(x)$  que minimiza el error mostrado en la ecuación 1.

$$E_{Y|X=x}[L(Y, \hat{Y}(x))] \quad (1)$$

Para este problema se tiene que

$$E_{Y|X=x}[L(Y, \hat{Y}(x))] = L(0, \hat{Y}(x))P(Y = 0|X = x) + L(1, \hat{Y}(x))P(Y = 1|X = x)$$

con  $L(Y, Y) = 0$ .

$$\text{Si } \hat{Y}(x) = 0 \Rightarrow E_{Y|X=x}[L(Y, \hat{Y}(x))] = L(1, 0)P(Y = 1|X = x)$$

$$\text{Si } \hat{Y}(x) = 1 \Rightarrow E_{Y|X=x}[L(Y, \hat{Y}(x))] = L(0, 1)P(Y = 1|X = x)$$

Si

$$\frac{L(1, 0)P(Y = 1|X = x)}{L(0, 1)P(Y = 1|X = x)} > 1$$

entonces  $\hat{Y}(x) = 1$  minimiza el error. De lo contrario  $\hat{Y}(x) = 0$  obtiene el mínimo. Si el cociente es igual a 1 entonces las dos opciones minimizan. En particular elegimos  $\hat{Y}(x) = 0$ .

Por lo tanto el clasificador tiene la siguiente forma:

$$\hat{Y}(x) = \mathbb{I} \left[ \frac{L(1, 0)P(Y = 1|X = x)}{L(0, 1)P(Y = 1|X = x)} > 1 \right] \hat{Y}(x) = \mathbb{I} \left[ \frac{P(Y = 1|X = x)}{P(Y = 1|X = x)} > \frac{L(0, 1)}{L(1, 0)} \right]$$

por el teorema de bayes, se obtiene lo siguiente:

$$\begin{aligned}\hat{Y}(x) &= \mathbb{I} \left[ \frac{P(Y = 1|X = x)}{P(Y = 1|X = x)} > \frac{L(0, 1)}{L(1, 0)} \right] \\ &= \mathbb{I} \left[ \frac{\frac{P(X=x|Y=1)P(Y=1)}{P(X=x)}}{\frac{P(X=x|Y=0)P(Y=0)}{P(X=x)}} > \frac{L(0, 1)}{L(1, 0)} \right] \\ &= \mathbb{I} \left[ \frac{P(X = x|Y = 1)}{P(X = x|Y = 0)} > \frac{L(0, 1)P(Y = 0)}{L(1, 0)P(Y = 1)} \right]\end{aligned}$$

Usando la ley de la probabilidad total, se tiene el siguiente resultado

$$\begin{aligned}P(Y = 0) &= \sum P(Y = 0|X = x)P(X = x) = \sum P(Y = 0, X = x) = 0.35 \\ P(Y = 1) &= \sum P(Y = 1|X = x)P(X = x) = \sum P(Y = 1, X = x) = 0.65\end{aligned}$$

Calculando las probabilidades condicionales se tiene lo siguiente

$$\begin{aligned}P(X = 1|Y = 1) &= \frac{12}{65} & P(X = 1|Y = 0) &= \frac{10}{35} \\ P(X = 2|Y = 1) &= \frac{10}{65} & P(X = 2|Y = 0) &= \frac{5}{35} \\ P(X = 3|Y = 1) &= \frac{25}{65} & P(X = 3|Y = 0) &= \frac{5}{35} \\ P(X = 4|Y = 1) &= \frac{18}{65} & P(X = 4|Y = 0) &= \frac{15}{35}\end{aligned}$$

por lo tanto, los coeficientes son

$$\begin{aligned}\frac{P(X = 1|Y = 1)}{P(X = 1|Y = 0)} &= \frac{420}{650} & \frac{P(X = 1|Y = 1)}{P(X = 1|Y = 0)} &= \frac{875}{325} \\ \frac{P(X = 2|Y = 1)}{P(X = 2|Y = 0)} &= \frac{350}{325} & \frac{P(X = 1|Y = 1)}{P(X = 1|Y = 0)} &= \frac{630}{975}\end{aligned}$$

Supongamos que el costo de equivocarse es 1, entonces

$$\frac{L(0, 1)}{L(1, 0)} = \frac{1}{1} = 1$$

Con estos resultados obtener que el clasificador esta definido por:  
si  $x = 1$

$$\begin{aligned}\hat{y}(x) &= \mathbb{I} \left[ \frac{P(X = 1|Y = 1)}{P(X = 1|Y = 0)} > \frac{L(0, 1)P(Y = 0)}{L(1, 0)P(Y = 1)} \right] \\ &= \mathbb{I} \left[ \frac{420}{650} > \frac{35}{65} \right] \\ &= 1\end{aligned}$$

si  $x = 2$

$$\begin{aligned}\hat{y}(x) &= \mathbb{I} \left[ \frac{P(X = 2|Y = 1)}{P(X = 2|Y = 0)} > \frac{L(0,1)P(Y = 0)}{L(1,0)P(Y = 1)} \right] \\ &= \mathbb{I} \left[ \frac{70}{65} > \frac{35}{65} \right] \\ &= 1\end{aligned}$$

si  $x = 3$

$$\begin{aligned}\hat{y}(x) &= \mathbb{I} \left[ \frac{P(X = 3|Y = 1)}{P(X = 3|Y = 0)} > \frac{L(0,1)P(Y = 0)}{L(1,0)P(Y = 1)} \right] \\ &= \mathbb{I} \left[ \frac{175}{65} > \frac{35}{65} \right] \\ &= 1\end{aligned}$$

si  $x = 4$

$$\begin{aligned}\hat{y}(x) &= \mathbb{I} \left[ \frac{P(X = 4|Y = 1)}{P(X = 4|Y = 0)} > \frac{L(0,1)P(Y = 0)}{L(1,0)P(Y = 1)} \right] \\ &= \mathbb{I} \left[ \frac{42}{65} > \frac{35}{65} \right] \\ &= 1\end{aligned}$$

Calculando el error se tiene lo siguiente:

Si  $\hat{y}(x) = 1$ , entonces

$$E_{Y|X=x}[L(y, \hat{y}(x))] = L(0,1)P(Y = 0) = P(Y = 0) = 0.35$$

Supongamos que  $L(y, \hat{y}(x))$  esta dada por

$$L(1,0) = 3 \quad L(0,1) = 2$$

es decir, clasificar mal cuando el verdadero valor de  $y$  es 1 tiene un costo de 3, en otro caso tiene un costo de 2.

por lo tanto el clasificador esta dado por

si  $x = 1$

$$\begin{aligned}\hat{y}(x) &= \mathbb{I} \left[ \frac{P(X = 1|Y = 1)}{P(X = 1|Y = 0)} > \frac{L(0,1)P(Y = 0)}{L(1,0)P(Y = 1)} \right] \\ &= \mathbb{I} \left[ \frac{420}{650} > \frac{70}{195} \right] \\ &= 1\end{aligned}$$

si  $x = 2$

$$\begin{aligned}
\hat{y}(x) &= \mathbb{I} \left[ \frac{P(X=2|Y=1)}{P(X=2|Y=0)} > \frac{L(0,1)P(Y=0)}{L(1,0)P(Y=1)} \right] \\
&= \mathbb{I} \left[ \frac{70}{65} > \frac{70}{195} \right] \\
&= 1
\end{aligned}$$

si  $x = 3$

$$\begin{aligned}
\hat{y}(x) &= \mathbb{I} \left[ \frac{P(X=3|Y=1)}{P(X=3|Y=0)} > \frac{L(0,1)P(Y=0)}{L(1,0)P(Y=1)} \right] \\
&= \mathbb{I} \left[ \frac{175}{65} > \frac{70}{195} \right] \\
&= 1
\end{aligned}$$

si  $x = 4$

$$\begin{aligned}
\hat{y}(x) &= \mathbb{I} \left[ \frac{P(X=4|Y=1)}{P(X=4|Y=0)} > \frac{L(0,1)P(Y=0)}{L(1,0)P(Y=1)} \right] \\
&= \mathbb{I} \left[ \frac{42}{65} > \frac{70}{195} \right] \\
&= 1
\end{aligned}$$

## Problema 02

Deriva el clasificador Bayesiano óptimo para el caso de tres clases y una función de costo simétrica cuando:

$$X|Y=1 \sim \mathcal{N}(\mu_1, \Sigma) \quad X|Y=2 \sim \mathcal{N}(\mu_2, \Sigma) \quad X|Y=3 \sim \mathcal{N}(\mu_3, \Sigma)$$

y

$$P(Y=1) = 2P(Y=2) = P(Y=3)$$

Para una  $x$  fija, busquemos la asignación  $\hat{y}(x)$  que minimiza el error

$$E_{Y|X=x}[L(y, \hat{y}(x))]$$

En este caso

$$\begin{aligned}
E_{Y|X=x}[L(y, \hat{y}(x))] &= L(1, \hat{y}(x))P(Y=1|X=x) + \\
&\quad L(2, \hat{y}(x))P(Y=2|X=x) + \\
&\quad L(3, \hat{y}(x))P(Y=3|X=x)
\end{aligned}$$

Si  $\hat{y}(x) = 1$ , entonces

$$\begin{aligned}
E_{Y|X=x}[L(y, \hat{y}(x))] &= L(2, 1)P(Y=2|X=x) + L(3, 1)P(Y=3|X=x) \\
&= P(Y=2|X=x) + P(Y=3|X=x)
\end{aligned}$$

Si  $\hat{y}(x) = 2$ , entonces

$$\begin{aligned} E_{Y|X=x}[L(y, \hat{y}(x))] &= L(1, 2)P(Y = 1|X = x) + L(3, 2)P(Y = 3|X = x) \\ &= P(Y = 1|X = x) + P(Y = 3|X = x) \end{aligned}$$

Si  $\hat{y}(x) = 3$ , entonces

$$\begin{aligned} E_{Y|X=x}[L(y, \hat{y}(x))] &= L(1, 3)P(Y = 1|X = x) + L(2, 3)P(Y = 2|X = x) \\ &= P(Y = 1|X = x) + P(Y = 2|X = x) \end{aligned}$$

Usando el teorema de bayes se obtiene que

$$\begin{aligned} P(Y = 2|X = x) + P(Y = 3|X = x) &= \frac{P(X = x|Y = 2)P(Y = 2)}{\sum_j P(X = x|Y = j)P(Y = j)} + \frac{P(X = x|Y = 3)P(Y = 3)}{\sum_j P(X = x|Y = j)P(Y = j)} \\ P(Y = 1|X = x) + P(Y = 3|X = x) &= \frac{P(X = x|Y = 1)P(Y = 1)}{\sum_j P(X = x|Y = j)P(Y = j)} + \frac{P(X = x|Y = 3)P(Y = 3)}{\sum_j P(X = x|Y = j)P(Y = j)} \\ P(Y = 1|X = x) + P(Y = 2|X = x) &= \frac{P(X = x|Y = 1)P(Y = 1)}{\sum_j P(X = x|Y = j)P(Y = j)} + \frac{P(X = x|Y = 2)P(Y = 2)}{\sum_j P(X = x|Y = j)P(Y = j)} \end{aligned}$$

si se cumpla la condición

$$\begin{aligned} P(X = x|Y = 1)P(Y = 1) &\leq P(X = x|Y = 2)P(Y = 2) \\ P(X = x|Y = 1)P(Y = 1) &\leq P(X = x|Y = 3)P(Y = 3) \end{aligned}$$

entonces se cumple que

$$\begin{aligned} E_{Y|X=x}[L(y, 1)] &\leq E_{Y|X=x}[L(y, 2)] \\ E_{Y|X=x}[L(y, 1)] &\leq E_{Y|X=x}[L(y, 3)] \end{aligned}$$

por lo que elegimos  $\hat{y}(x) = 1$ .

Si se cumpla la condición

$$\begin{aligned} P(X = x|Y = 2)P(Y = 2) &\leq P(X = x|Y = 1)P(Y = 1) \\ P(X = x|Y = 2)P(Y = 2) &\leq P(X = x|Y = 3)P(Y = 3) \end{aligned}$$

entonces se cumple que

$$\begin{aligned} E_{Y|X=x}[L(y, 2)] &\leq E_{Y|X=x}[L(y, 1)] \\ E_{Y|X=x}[L(y, 2)] &\leq E_{Y|X=x}[L(y, 3)] \end{aligned}$$

por lo que elegimos  $\hat{y}(x) = 2$ .

Si se cumpla la condición

$$P(X = x|Y = 3)P(Y = 3) \leq P(X = x|Y = 1)P(Y = 1)$$

$$P(X = x|Y = 3)P(Y = 3) \leq P(X = x|Y = 2)P(Y = 2)$$

entonces se cumple que

$$\begin{aligned} E_{Y|X=x}[L(y, 3)] &\leq E_{Y|X=x}[L(y, 1)] \\ E_{Y|X=x}[L(y, 3)] &\leq E_{Y|X=x}[L(y, 2)] \end{aligned}$$

por lo que elegimos  $\hat{y}(x) = 3$ .

Por lo tanto, para una  $x$  fija debemos calcular tres coeficientes y compararlos con 1 para determinar la clasificación correspondiente. Una alternativa es calcular el logaritmo de los coeficientes y compararlo con 0. Los coeficientes son los siguientes:

$$\begin{aligned} \frac{P(X = x|Y = 1)P(Y = 1)}{P(X = x|Y = 2)P(Y = 2)} &= \frac{P(X = x|Y = 1)2P(Y = 2)}{P(X = x|Y = 2)P(Y = 2)} = \frac{2P(X = x|Y = 1)}{P(X = x|Y = 2)} \\ \frac{P(X = x|Y = 1)P(Y = 1)}{P(X = x|Y = 3)P(Y = 3)} &= \frac{P(X = x|Y = 1)P(Y = 3)}{P(X = x|Y = 3)P(Y = 3)} = \frac{P(X = x|Y = 1)}{P(X = x|Y = 3)} \\ \frac{P(X = x|Y = 2)P(Y = 2)}{P(X = x|Y = 3)P(Y = 3)} &= \frac{P(X = x|Y = 2)P(Y = 2)}{2P(X = x|Y = 3)P(Y = 2)} = \frac{P(X = x|Y = 2)}{2P(X = x|Y = 3)} \end{aligned}$$

Dado que  $X|Y = j \sim \mathcal{N}(\mu_j = \Sigma)$ , entonces calculando los logaritmos se tiene que

$$\begin{aligned} \log \left( \frac{P(X = x|Y = 1)}{P(X = x|Y = 2)} \right) &= \log(2) + (\mu_1 - \mu_2)^T \Sigma^{-1} x + \frac{1}{2}(\mu_1^T \Sigma \mu_1 - \mu_2^T \Sigma \mu_2) = \ell_{1,2}^T x + b_{1,2} \\ \log \left( \frac{P(X = x|Y = 1)}{P(X = x|Y = 3)} \right) &= (\mu_1 - \mu_3)^T \Sigma^{-1} x + \frac{1}{2}(\mu_1^T \Sigma \mu_1 - \mu_3^T \Sigma \mu_3) = \ell_{1,3}^T x + b_{1,3} \\ \log \left( \frac{P(X = x|Y = 2)}{P(X = x|Y = 3)} \right) &= (\mu_2 - \mu_3)^T \Sigma^{-1} x + \frac{1}{2}(\mu_2^T \Sigma \mu_2 - \mu_3^T \Sigma \mu_3) - \log 2 = \ell_{2,3}^T x + b_{2,3} \end{aligned}$$

por lo tanto

$$\hat{y}(x) = \begin{cases} 1 & \text{para } \ell_{1,2}^T x + b_{1,2} \geq 0 \text{ y } \ell_{1,3}^T x + b_{1,3} \geq 0 \\ 2 & \text{para } \ell_{1,3}^T x + b_{1,3} \leq 0 \text{ y } \ell_{2,3}^T x + b_{2,3} \geq 0 \\ 3 & \text{para } \ell_{2,3}^T x + b_{2,3} \leq 0 \text{ y } \ell_{1,2}^T x + b_{1,2} \leq 0 \end{cases}$$

## Problema 03

Genere un conjunto de datos de entrenamiento de 200 puntos en  $\mathcal{R}^3$  muestreando 100 puntos con coordenadas independientes de una normal  $\mathcal{N}(4, 1)$  y 100 puntos de una normal  $\mathcal{N}(8, 1)$ . Ejecute el algoritmo de agrupamiento de k-medias, para  $k = 2, 3, \dots, 15$ , usando el conjunto de datos de entrenamiento. Para cada  $k$  use diez puntos iniciales aleatorios y solo guarde la solución que tenga el menor valor de la función objetivo de k-medias.

## Punto 01

Muestra en una gráfica el valor de la función objetivo de k-medias resultante sobre el conjunto de datos de entrenamiento como una función de k. Comenta lo que ves. Qué valor de k seleccionaría basándose solo en esta gráfica?

En la figura 1 se muestran los resultados de la función objetivo objetivos con el número de clusters  $k = 2, 3, \dots, 15$ . En esta gráfica se observa el comportamiento descendiente de la función objetivo conforme aumentan el número de clusters a tomar en cuenta. Este comportamiento aparenta llegar a una convergencia a un valor fijo conforme el número de clusters aumenta.

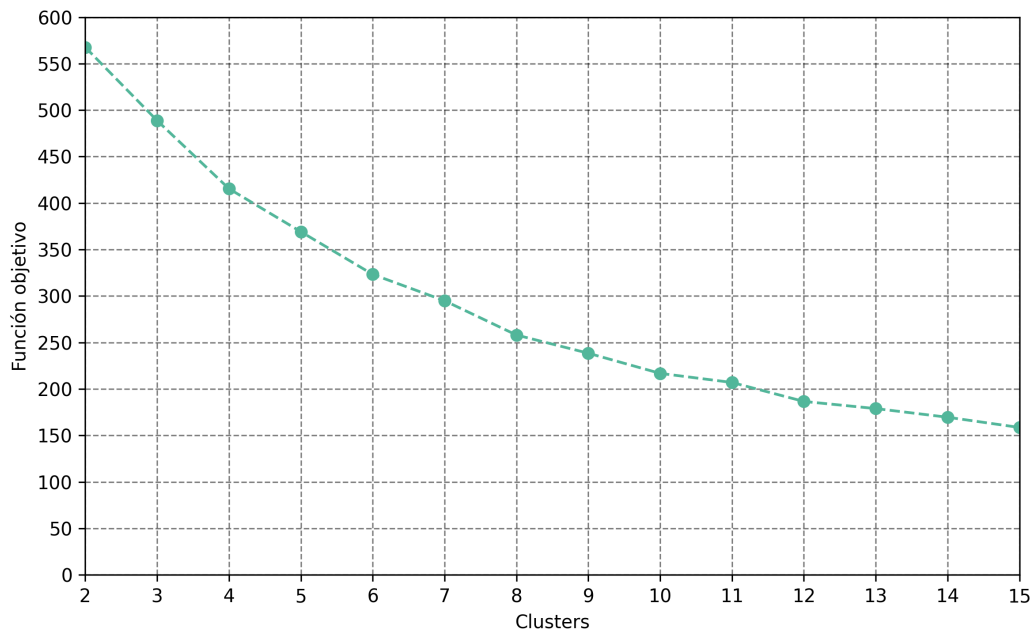


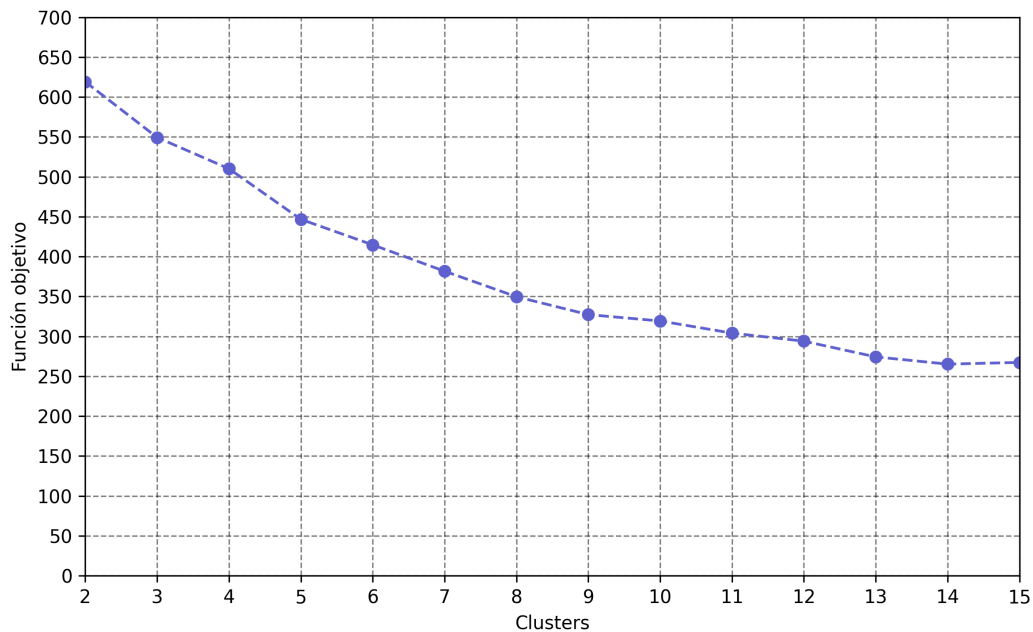
Figura 1: Valor de la función objetivo para los datos de entrenamiento.

El valor de k que seleccionaría de la gráfica sería 3. Esto debido a que el score obtenido tiene una diferencia con su anterior k mayor a comparación de las demás diferencias obtenidas.

## Punto 02

Genere un conjunto de datos de validación del mismo tamaño que el conjunto de entrenamiento y de la misma manera. Para cada k, asigne cada punto o dato de validación a la media de clúster aprendida más cercana en la parte (a) y muestre en una gráfica el valor de la función objetivo resultante de k-medias usando los datos de validación como una función de k. Comenta lo que ves. Qué valor de k se seleccionaría si usara como criterio de selección el mínimo valor de la función objetivo para los datos de validación?

En la figura 2 se muestra los resultados de la función objetivo para los datos de validación. Si se usará el criterio del valor mínimo de la función de objetivo el k seleccionado sería el k=14. En la tabla 1 se muestran los cinco k con menor valor en la función objetivo.



**Figura 2:** Valor de la función objetivo para los datos de validación.

$k$	Función objetivo
14	265.194122
15	267.335407
13	274.270514
12	293.978512
11	304.018999

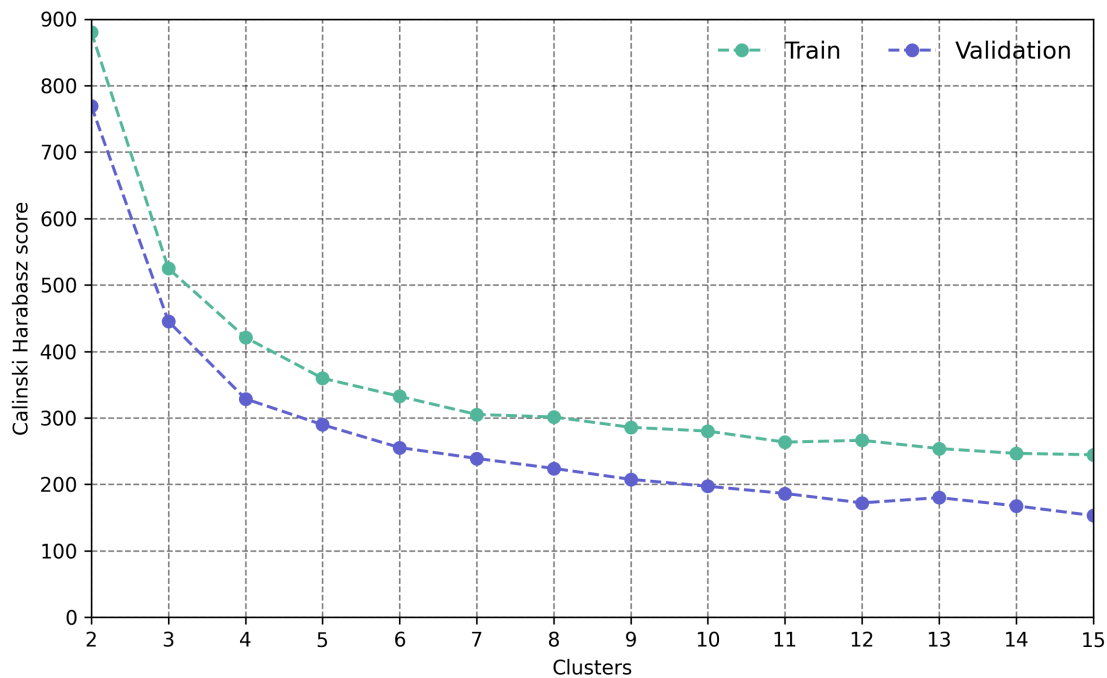
**Tabla 1:** Valores de  $k$  con menor costo objetivo.

### Punto 03

Para cada  $k$ , calcule y muestre en una gráfica el índice (score) Calinski Harabasz (CH) calculado para los datos de entrenamiento. Muestra CH como una función de  $k$ , y comente al respecto. Que valor de  $k$  maximiza este criterio?

En la figura 3 se muestran los scores obtenidos con el índice de Calinski Harabasz para los datos de entrenamiento y validación. Si se quisiera obtener un máximo del índice se tomaría a  $k=2$ . Esto debido a que en los datos de entrenamiento y validación en donde se presenta este criterio.





**Figura 3:** Índice de Calinski Harabasz obtenido a partir de la predicción de los datos de entrenamiento y validación.

## Problema 04

En este problema compararán varios esquemas de inicialización para el algoritmo de agrupamiento k-medias

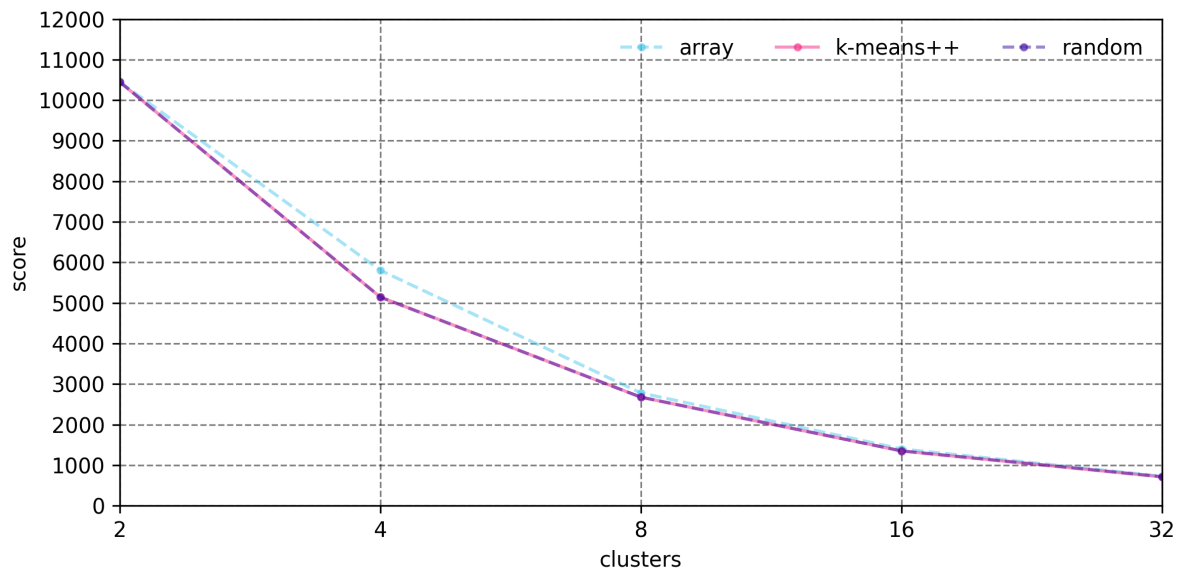
- El esquema de inicialización aleatorio o ‘random’.
- El esquema de inicialización ‘k-means++’.
- Proponga su propio esquema de inicialización.

Para  $k \in \{2, 4, 8, 16, 32\}$  y para los 3 esquemas de inicialización, ejecute el algoritmo k-medias para la imagen a color [Colorful-Flowers.jpg](#) y obtenga la representación comprimida. Considere cada pixel como un punto en 3 dimensiones.

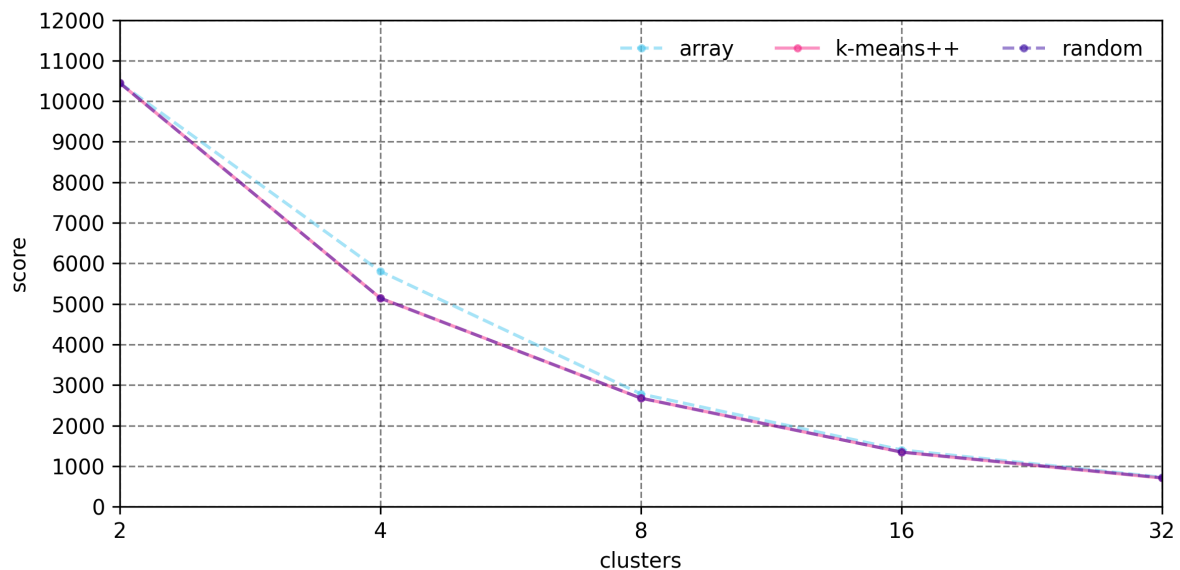
### Punto 01

Ejecute cada inicialización 5 veces con diferentes semillas y muestre en una gráfica el mínimo y la media de la función objetivo de k-medias como una función de k.

En la figura 4 se muestra el promedio y mínimo de la función objetivo de inicializador del método de k-means. En esta se visualiza que la tendencia de las dos es semejante. La diferencia promedio entre la media y mínimo es 1.0729. Comparado con el orden de magnitud de la función de costo este valor es muy pequeño. Por lo que tomar el mínimo un promedio de los resultados no varía mucho en las predicciones de cada modelo.



(a) Promedios de la función objetivo.



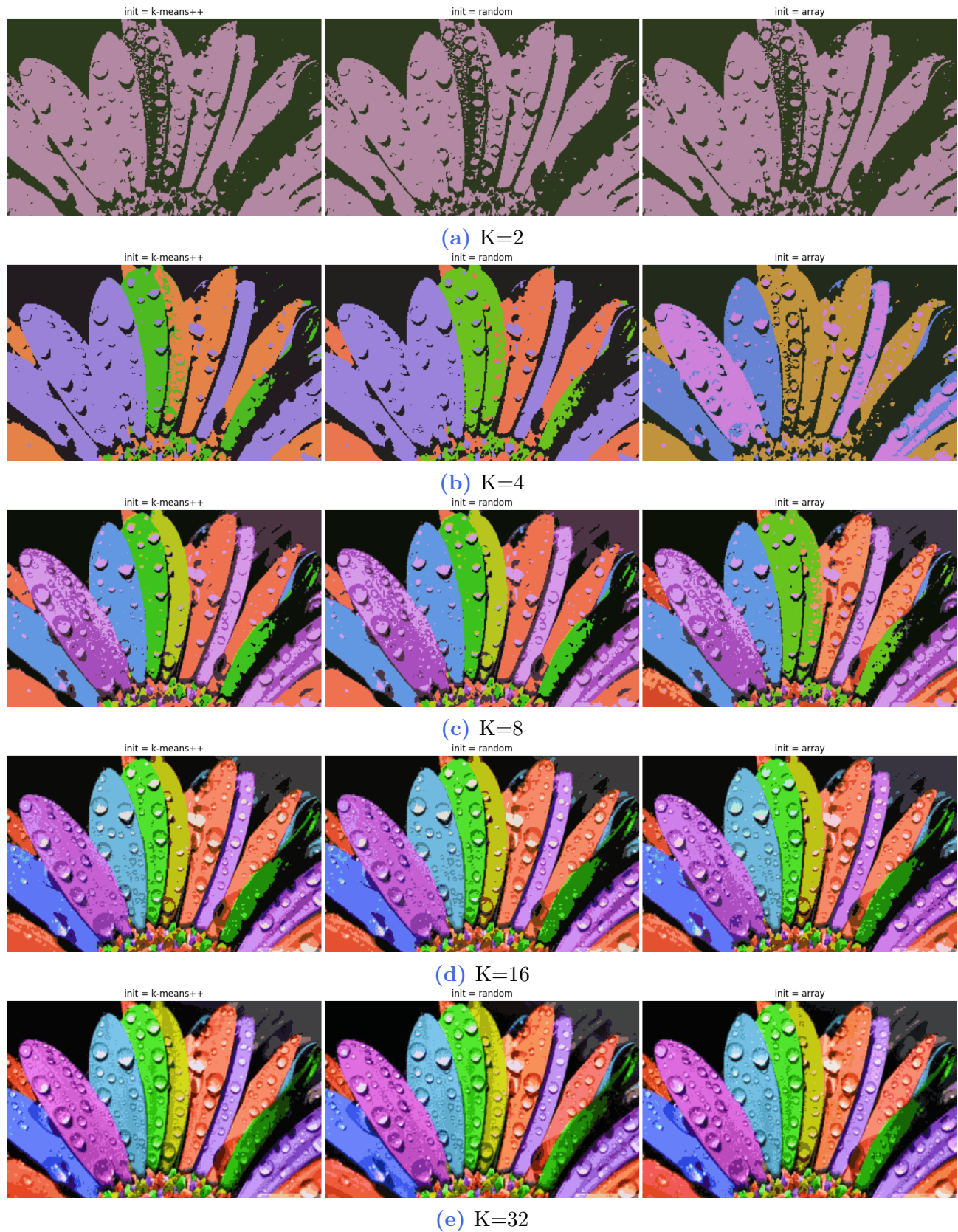
(b) Mínimo de la función de objetivo.

**Figura 4:** Promedios y mínimo de la función objetivo de cada inicializador dado en el método de k-means.

## Punto 02

Para cada procedimiento y para cada  $k$ , visualice la imagen comprimida usando la mejor corrida aleatoria. Cual es el menor valor de  $k$  para el cual está satisfecho con el resultado obtenido?

Se selecciono el modelo que obtuviera la menor función objetivo dado el modelo y una  $k$ . En la figura se muestran las compresiones realizadas por cada modelo.



**Figura 5:** Compresiones de imágenes para cada modelo y número de clusters dado.

## Problema 05

Es importante que las compañías de tarjetas de crédito puedan reconocer transacciones fraudulentas con tarjetas de crédito para que a los clientes no se les cobren artículos que no han comprado.

Descarga la base de datos 'creditcard.csv' de la pagina Kaggle, <https://www.kaggle.com/mlg-ulb/creditcardfraud> Descripción de la base de datos: El conjunto de datos contiene transacciones realizadas con tarjetas de crédito en septiembre de 2013 por titulares de tarjetas europeos. Este conjunto de datos presenta transacciones que ocurrieron en dos días, donde tenemos 492 fraudes en 284,807 transacciones. El conjunto de datos está muy desequilibrado, la clase positiva (fraudes) representa 0, 172 % de todas las transacciones. La base de datos solo contiene variables de entrada numéricas que son el resultado de una transformación PCA. Desafortunadamente, debido a problemas de confidencialidad, no se pueden proporcionar las características originales, ni más información general sobre los datos. Las características  $V_1, V_2, \dots, V_{28}$  son los principales componentes obtenidos con PCA, las únicas características que no han sido transformadas con PCA son 'Tiempo' y 'Cantidad' ('Time' and 'Amount'). La función 'Tiempo' ('Time') contiene los segundos transcurridos entre cada transacción y la primera transacción en el conjunto de datos. La función 'Cantidad' ('Amount') es la cantidad de la transacción. Característica 'Class' es la variable respuesta y toma valor 1 en caso de fraude y 0 en caso contrario.

Ejecuta 5 veces k-medias con diferentes semillas (random seeds) usando el esquema de inicialización 'k-means++'.

### Punto 01

Muestra en una gráfica el mínimo de la función objetivo como una función de  $k$ , para  $k \in \{2, 3, 4, 5, 6\}$

En la figura 6 se visualiza la función objetivo obtenida para el dataset [creditcar.csv](#). En esta se presenta una disminución en su valor conforme el número de clusters aumenta.

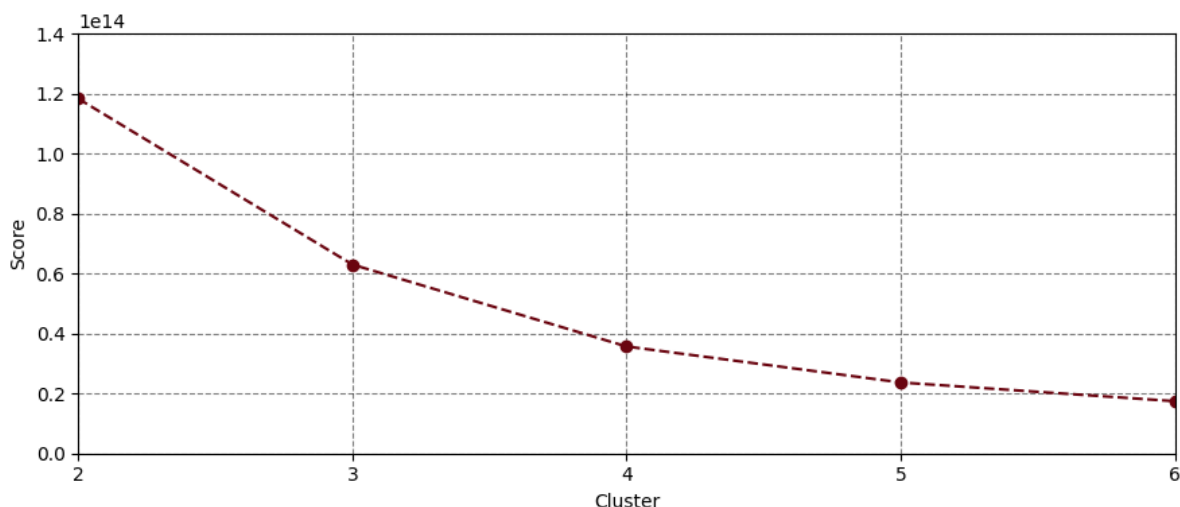
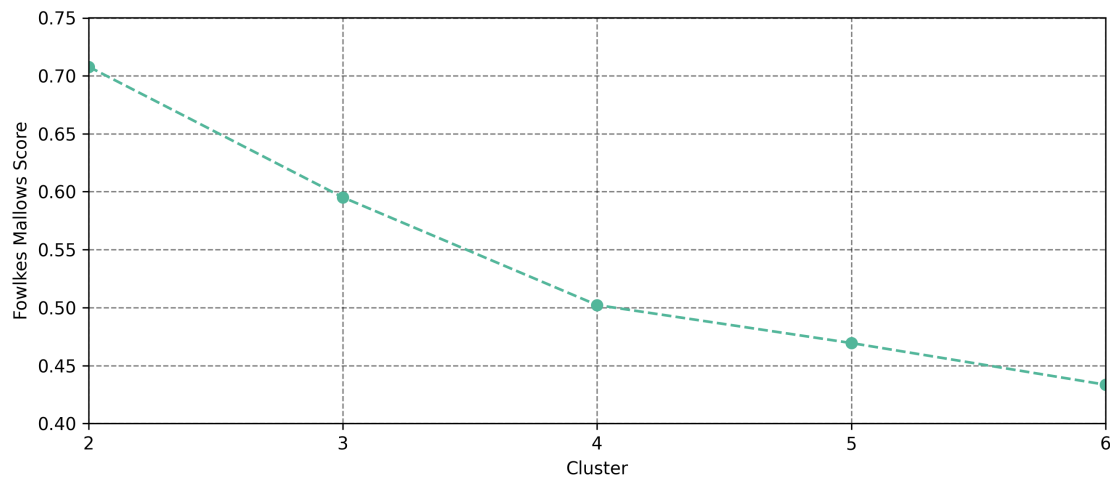


Figura 6: Mínimo de la función objetivo para cada  $k$  calculada.

## Punto 02

Calcula y muestra el índice Fowlkes Mallows score (FM) usando el conjunto de entrenamiento. Muestra en una gráfica FM como una función de  $k$ , y comenta al respecto. Qué valor de  $k$  maximiza este criterio?

En la figura 7 se visualiza el índice de Fowlkes Mallows para cada  $k$  calculada con el conjunto de entrenamiento. El índice de Fowlkes Mallows sigue una forma semejante a la figura 6. La diferencia se encuentra en el orden de magnitud de los valores de cada score obtenido. Si se quisiera maximizar el índice de Fowlkes Mallows se usaría  $k=2$ . Ya que para este valor se obtiene el valor máximo del índice.



**Figura 7:** Índice de Fowlkes Mallows obtenido a partir de los resultados usando el conjunto de entrenamiento del dataset [creditcard.csv](#).