Computer Science Department
Universidad de Ingeniería y Tecnología

Proyecto 2: Clustering
Prof. Ariana Villegas

---

1. Sección teórica. (30 pts)

1.1. Fundamentos de K-means y GMM. (10 pts)

Considere algoritmos de agrupamiento para datos $D = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$ con $x^{(i)} \in \mathbb{R}^d$.

1. La función objetivo de k-means se define como $J(\mu_1, \ldots, \mu_k, z) = \sum_{i=1}^{N} \sum_{j=1}^{k} z_{ij} \|x^{(i)} - \mu_j\|^2$, donde $z_{ij} \in \{0, 1\}$ indica las asignaciones de clúster. Explique por qué esta función nunca aumenta durante las iteraciones del algoritmo y por qué esto no garantiza encontrar el óptimo global. (5 pts)

2. Compare k-means y Modelos de Mezclas Gaussianas (GMM) matemáticamente. Específicamente, describa qué suposiciones hace cada modelo sobre la forma y el tamaño de los clústeres, y proporcione un ejemplo simple de datos donde GMM superaría a k-means. (5 pts)

1.2. Agrupamiento Jerárquico. (10 pts)

Para un conjunto de datos $D = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$ con $x^{(i)} \in \mathbb{R}^d$:

1. En el agrupamiento jerárquico aglomerativo:

   (a) Defina matemáticamente las métricas de distancia de enlace simple, enlace completo y enlace promedio entre clústeres. (5 pts)

   (b) Para el conjunto de datos en $\mathbb{R}^2$ con puntos: (1,1), (2,1), (5,3), (6,3) y (10,5), muestre las primeras dos fusiones que ocurrirían usando enlace simple y explique cómo el resultado podría diferir con enlace completo. (5 pts)

1.3. Agrupamiento Basado en Densidad. (10 pts)

Para algoritmos de agrupamiento basados en densidad:

1. Defina formalmente los conceptos de puntos **directamente alcanzables por densidad**, **alcanzables por densidad** y **conectados por densidad** en DBSCAN. (5 pts)

2. Utilizando un pequeño ejemplo de conjunto de datos, ilustre cómo DBSCAN puede identificar clústeres de forma arbitraria mientras explica por qué k-means fallaría en el mismo conjunto de datos. (5 pts)

2. Sección aplicada. (70 pts)

1. Para este proyecto, desarrollarás un sistema de recomendación de películas basado en técnicas de agrupamiento de imágenes. El sistema recomendará películas similares analizando y agrupando características visuales de pósters o fotogramas de películas. Para entrenar su sistema de recomendación utilicen el archivo *movies_train.csv*, y para la evaluación empleen el archivo *movies_test.csv* siguiendo el formato de *sample_submission.csv*. Ambos archivos están disponibles en el siguiente enlace de Drive.

2. Puedes obtener los datos de las siguientes fuentes:

   a) MovieLens Dataset (25M o más pequeño) [*Recommended*]

   b) Movie Genre from Poster Dataset

   c) TMDb API (para descargar pósters adicionales si es necesario)

3. Extrae características visuales de los pósters de películas utilizando técnicas tradicionales de computer vision. E.g.: a) Histogramas de color (RGB, HSV), b) Descriptores de textura (GLCM, LBP), de forma o bordes (HOG, SIFT), c) Momentos de imagen (Hu, Zernike).

4. Reduce la dimensionalidad de los vectores de características extraídos utilizando al menos dos técnicas. Ej.

   a) PCA (Análisis de Componentes Principales)

   b) LDA (Análisis Discriminante Lineal) si tienes etiquetas de género

   c) SVD (Descomposición en Valores Singulares)

   Analiza y compara los resultados obtenidos con ambos métodos.

5. Implementa dos algoritmos de agrupamiento distintos para organizar las películas según sus características visuales. Compara los resultados y justifica cuál es más adecuado para agrupar pósters de películas. E.g.:

   a) Un método de particionamiento (K-means, K-medoids)

   b) Un método jerárquico (Agrupamiento jerárquico)

   c) Un método basado en densidad (DBSCAN, OPTICS)

   d) Un método basado en distribución (GMM)

6. Desarrolla un visualizador simple pero funcional que permita:

   a) Buscar películas por similitud visual (seleccionando un póster o subiendo una imagen)

   b) Mostrar películas representativas de cada grupo (cluster)

   c) Visualizar la distribución de películas en un espacio bidimensional según sus características visuales

   d) Filtrar resultados por género, año u otros metadatos disponibles

7. Evalúa la calidad de las agrupaciones y recomendaciones mediante:

   a) Métricas internas: silhoutte score, rand index, información mutua

   b) Análisis de coherencia de género dentro de cada grupo (¿las películas dentro de un mismo grupo comparten géneros similares?)

   c) Ejemplos concretos de recomendaciones generadas y su relevancia

** Evita usar capturas de pantalla para mostrar resultados como precisión, puntuación F1, pérdida o error. En su lugar, asegúrate de que todos los resultados estén correctamente formateados y presentados dentro del documento.*

** El documento debe tener **un máximo de 8 páginas** y puede incluir cualquier número de apéndices que se consideren apropiados.*

Uso de bibliotecas: Para preprocesamiento/métodos/métricas distintas a las que requieren explicitamente implementación, eres libre de utilizar bibliotecas.

2.1. Rúbrica de Evaluación

Criterios	Excelente	Bueno	Aceptable	Deficiente
Extracción de Características Visuales (15 pts)	Implementación de dos o más técnicas con análisis profundo. Características altamente discriminativas.	Buena implementación de dos técnicas. Características útiles para el clustering.	Implementación básica con análisis limitado. Características parcialmente útiles.	Implementación incorrecta. Características inadecuadas.
Algoritmos de Clustering (15 pts)	Implementación experta de dos algoritmos con optimización rigurosa. Comparación crítica y justificación sólida.	Implementación correcta con buena selección de parámetros y análisis comparativo.	Implementación básica con configuración estándar y análisis limitado.	Implementación deficiente o uso incorrecto. Sin análisis o justificación.
Sistema de Recomendación (10 pts)	Sistema sofisticado que aprovecha eficazmente los clusters. Recomendaciones de alta calidad.	Sistema funcional que utiliza apropiadamente los clusters. Recomendaciones razonables.	Sistema básico con uso limitado de los clusters. Recomendaciones simples.	Sistema deficiente. Recomendaciones irrelevantes o sin relación con el clustering.
Visualizador Interactivo (10 pts)	Visualizador completo e interactivo. Representación clara de clusters y recomendaciones. Interfaz intuitiva.	Visualizador funcional con capacidades interactivas básicas. Buena representación.	Visualizador simple con funcionalidad limitada. Representación básica.	Visualizador deficiente o no funcional. No permite explorar resultados.
Evaluación del Sistema (10 pts)	Evaluación rigurosa con múltiples métricas. Análisis profundo. Comparaciones sistemáticas.	Buena evaluación con métricas apropiadas. Análisis comparativo adecuado.	Evaluación básica con métricas estándar. Análisis limitado.	Evaluación superficial o incorrecta. Métricas inadecuadas.
Votación Mejor Recomendación (5 pts)	5 votos = 5 puntos.	4 votos = 4 puntos.	3 votos = 3 puntos.	1–2 votos = 1–2 puntos.
Informe Técnico (5 pts)	Informe ejemplar con visualizaciones profesionales. Análisis crítico y reflexivo.	Buen informe con estructura clara. Resultados bien presentados.	Informe básico con análisis superficial.	Informe incompleto o desorganizado. Análisis deficiente.

Cuadro 1: Rúbrica para el Sistema de Recomendación Basado en Clustering