Universidad EAFIT

Maestría en Ciencias de Datos
Examen 2 - Matemáticas Aplicadas en Ciencia de Datos
Semestre 2026-1

Profesor: Diego Fonseca

Plazo de entrega: Martes 24 de Febrero de 2025 a las 11:59pm
Modalidad: Individual o grupos de máximo 3 personas
Importante: Cualquier intento de plagio será llevado a las instancias pertinentes de la universidad.

Instrucciones

- Este examen se puede realizar en grupos de máximo 3 estudiantes.
- El estudiante líder (el que entrega el examen) debe marcar correctamente su archivo con su trabajo. La primera línea de texto debe contener el nombre y apellidos completos, y código de estudiante (o cédula en caso de no tener código) de todos los integrantes del grupo.
- Puede entregar un archivo de PDF apoyado con un archivo de Colab o Notebook de Jupiter, o solo el notebook o el notebook en PDF. Independiente del formato, este debe nombrarse de la siguiente manera: `Examen1-MatApl-ApellidoIntegrante1-ApellidoIntegrante2-ApellidoIntegrante3-Tipo de archivo`.
- Los ejercicios tienen componentes prácticas, por lo que debe suministrar el código de sus implementaciones. Se recomienda entregar un script en Google Colab o un Notebook de Jupyter. Procure comentarlos de manera clara. Todo lo que muestre debe tener un propósito, debe ser conciso en las explicaciones.
- Debe justificar sus procedimientos y soluciones, e interpretar sus resultados de acuerdo con el contexto de cada ejercicio. No basta con llegar a las soluciones; es fundamental explicarlas y argumentarlas.
- Los archivos con sus soluciones se suben al buzón de la actividad de EAFIT Interactiva.

Ejercicio 1 [1.7 pts]

Se te proporciona el archivo `datos_ejercicio1.csv` (disponible en la plataforma EAFIT Interactiva), el cual contiene n observaciones x_i en $\mathbb{R}^d$ (cada fila es un vector de características). Tu objetivo es agrupar estas observaciones en k clusters utilizando un procedimiento que, en esencia, aprovecha la información de los eigenvalores y eigenvectores de una matriz derivada de la similitud entre las observaciones. Sin embargo, no se te indicará que se trata de clustering espectral; deberás descubrir la estructura latente mediante el siguiente proceso:

1. **Construcción de la Matriz de Similitud:** Define la matriz de similitud S de la siguiente forma:

$$S_{ij} = \begin{cases} \dfrac{1}{\|x_i - x_j\| + \varepsilon} & \text{si } x_j \text{ es uno de los } k_{\text{NN}} \text{ vecinos más cercanos de } x_i, \\ 0 & \text{en otro caso,} \end{cases}$$

   donde $\varepsilon > 0$ es una constante pequeña para evitar divisiones por cero (pore ejemplo $\varepsilon = 10^{-5}$), y k_{NN} es un parámetro que tú deberás seleccionar (por ejemplo $k_{\text{NN}} = 10$).

2. **Construcción de la Matriz Laplaciana:** Calcula la matriz de grado D con $D_{ii} = \sum_j S_{ij}$ (D es una matriz diagonal, el resto de entradas son cero) y define la Laplaciana como:

$$L = D - S.$$

3. **Análisis de Eigenpares:** Calcula los eigenvalores y eigenvectores de L. Ordena los eigenvalores de menor a mayor y, mediante un Scree Plot, determina un valor adecuado de k basándote en la presencia de una brecha en la secuencia de eigenvalores, es decir, ese valor adecuado será el primer k para el cual la brecha $\lambda_{k+1} - \lambda_k$ es grande a comparación de las brechas para k anteriores.
   Sugerencia: Pueden haber muchos valores propios, así que se sugiere empezar graficando los primeros 10 valores propios, si no se encuentra la brecha, se procede con los siguientes 10, así sucesivamente.

4. **Construcción del Espacio de Embedding:** Selecciona los k eigenvectores correspondientes a los k menores eigenvalores de L y forma la nueva representación de cada observación:

$$y_i = \big(v_1(i), v_2(i), \ldots, v_k(i)\big).$$

5. **Agrupamiento:** En el espacio de dimensión k, aplica un método de agrupamiento (por ejemplo, k-means, para este se pueden apoyar en alguna librería que ya lo tenga implementado) para los datos $\{y_i\}_{i=1}^n$ y así asignar cada observación y_i a uno de k clusters. Finalmente, asigna a cada observación x_i el cluster correspondiente a y_i.

6. **Tareas finales:**

   (a) Implementa directamente k-means (de nuevo, para este se pueden apoyar en alguna librería que ya lo tenga implementado) en el espacio original de los datos (es decir, en los $\{x_i\}_{i=1}^n$ directamente) y compara visualmente (por ejemplo, mediante gráficos de dispersión en dos dimensiones) la agrupación obtenida con la del procedimiento anterior.

   (b) Experimenten cambiando la función de similitud. Discutan sus observaciones y justifiquen su elección.

Ejercicio 2 [1.8 pts]

En esta actividad trabajarás con imágenes de dígitos escritos a mano utilizando el conjunto de datos MNIST (disponible en Keras). Tu tarea es implementar un clasificador simple basado en la utilización del conceptos de álgebra lineal. El procedimiento es el siguiente:

1. **Selección y Preprocesamiento de Datos:**
   - Para cada dígito $d = 0, 1, \ldots, 9$, selecciona aleatoriamente k imágenes de entrenamiento. Esto te dará un conjunto de datos de tamaño $10 \times k$.
   - Convierte cada imagen a una matriz. Luego, recorta la imagen para eliminar las filas y columnas con únicamente el fondo (de modo que se preserve la mayor parte del dígito). Luego reescala la imagen resultante a un tamaño fijo de 16×16 píxeles.
   - Finalmente, convierte cada imagen reescalada a un vector de longitud 256.
2. **Construcción de Bases Latentes:**
   - Para cada dígito d, organiza los k vectores resultantes en una matriz M_d (donde cada columna corresponde a una imagen).
   - Para cada dígito calcula una base orto-normal del espacio vectorial generado por las columna de M_d, denotemos a esa base como B_d y llamemos a ese espacio generado U_d.
   - ¿Cómo se puede interpretar el espacio generado por B_d en palabras?
3. **Clasificación de una Imagen Nueva:**
   - Toma una imagen de prueba (que no esté en el conjunto de entrenamiento) y apliquele el mismo proceso: recortar, reescalar a 16×16 y convertir a un vector q de longitud 256.
   - Para cada dígito d, calcula la proyección ortogonal de q sobre el subespacio generado por B_d, obteniendo el vector proyectado p_d.
   - Define la distancia $\text{dist}(q, d) = \|q - p_d\|$. Clasifica la imagen q asignándole la etiqueta del dígito d que minimice esta distancia.
4. **Evaluación y Análisis:**
   - Utiliza un conjunto de validación (imágenes de prueba distintas a las utilizadas para formar M_d) y evalúa el porcentaje de aciertos de tu clasificador.
   - Repite la evaluación variando k (el número de imágenes de entrenamiento por dígito) y grafica la precisión del clasificador en función de k.
   - Reflexiona sobre el compromiso entre el tamaño del conjunto de entrenamiento (y, por ende, la calidad de la base B_d) y la precisión de la clasificación.
   - Adicionalmente, discute qué ventajas e inconvenientes encuentras en este método de clasificación en comparación con otros enfoques que suelen usarse para clasificar estos dígitos. (aquí no debes implementar otros métodos, pero debes ser riguroso en tu argumentación, explorar las otras opciones existentes)

Ejercicio 3 [1.5 pts]

Este ejercicio se inspira en pruebas psicotécnicas comunes en procesos de selección laboral y tiene como objetivo desarrollar soluciones algorítmicas que respondan a preguntas de razonamiento verbal.

Se le solicita que:

1. Obtenga o elabore una prueba de razonamiento verbal que contenga:
   - 10 preguntas de analogía (ej.: "A es a B como C es a _____").
   - 10 preguntas de contexto (ej.: "¿Cuál de las siguientes palabras no encaja en el conjunto {...}?").
2. Desarrolle dos algoritmos (o procedimientos), uno para responder a las preguntas de analogía y otro para las preguntas de contexto, utilizando un método para convertir frases en representaciones numéricas (use la libreria *SentenceTransformer* y también debe usar TF-IDF, la idea es ver si el rendimiento de los algoritmos depende del embeding usado).
3. Genere representaciones gráficas que ilustren las relaciones entre las frases de la prueba.
4. Evalúe el rendimiento de sus algoritmos y compare los resultados obtenidos.

Nota: La obtención de las preguntas (analogías y contexto) es parte del desafío; puede buscarlas en fuentes disponibles en Internet.