INSTITUTO TECNOLÓGICO DE COSTA RICA
INGENIERÍA EN COMPUTACIÓN
I SEMESTRE 2022

Inteligencia Artificial
Tarea programada #4

Tema: Análisis de componentes principales (PCA por sus siglas en inglés) y clustering

Entrega: Un archivo .zip que contenga un archivo en formato Jupyter notebook bien documentado que incluya los ejercicios. A través del TEC-digital.

Modo de trabajo: Grupos de 2 personas máximo.

Tecnología a utilizar: Python, PyTorch, entre otras bibliotecas.

Introducción:

En este trabajo se aplicarán conceptos básicos relacionados con análisis de componentes principales para reducir la dimensionalidad de un conjunto de datos y clustering para caracterizar un fenómeno como la criminalidad por distrito en el país.

Las y los estudiantes deberán completar dos secciones de ejercicios. La primera sección está orientada a trabajar con clustering utilizando dos algoritmos k-means y DBSCAN aplicados a datos de criminalidad en Costa Rica. La segunda parte consiste en aplicar PCA a los datos de criminalidad en Costa Rica para analizar si es posible disminuir la dimensionalidad de estos antes de utilizarlos en un proceso de modelización.

El **objetivo del trabajo** es poner en práctica las **habilidades de investigación y el conocimiento adquirido durante el curso** por medio de ejercicios prácticos que permitan a las y los estudiantes experimentar con algoritmos representativos del aprendizaje automático.

Objetivos de aprendizaje:

1. Poner en práctica habilidades de investigación y documentación de resultados.
2. Aplicar el conocimiento teórico y práctico sobre análisis de componentes principales adquirido en el curso para reducir la dimensionalidad de los datos.
3. Implementar y experimentar con algoritmos de clustering aplicados a problemas de la vida real.
4. Fortalecer capacidades en los estudiantes en el uso de bibliotecas de aprendizaje automático.

Ejercicios

Sección 1. Clustering

Ejercicio 1. Implemente "de cero" los algoritmos K-Means y DBSCAN, es decir sin utilizar ninguna biblioteca.

1. Implemente de cero el algoritmo de K-Means para vectores de atributos de cualquier dimensión sin utilizar ninguna implementación de biblioteca.
2. Implemente de cero el algoritmo DBSCAN.
3. Pruebe ambos algoritmos con datos generados artificialmente.
4. Visualice ambos resultados.

Ejercicio 2. Aplicación de los algoritmos de clustering implementados en la sección anterior para caracterizar la criminalidad en Costa Rica a partir de datos del Organismo de Investigación Judicial (OIJ) y el Instituto Nacional de Estadística y Censos (INEC).

Objetivo: El objetivo del presente ejercicio es utilizar datos de criminalidad en Costa Rica combinados con datos socio-económicos asociados a distritos para demostrar cuán efectivos y precisos pueden ser los algoritmos de *clustering* en la definición de perfiles de criminalidad por distrito a nivel nacional. El presente ejercicio utilizará datos numéricos únicamente para caracterizar de forma muy simplificada (**utilizando al menos cinco características seleccionadas por las estudiantes)** los distritos de Costa Rica.

1. Baje los siguientes conjuntos de datos, publicados por las siguientes instituciones nacionales, intégrelos por medio del nombre del distrito y pre-procéselos para el ejercicio (documente muy bien todo el proceso).

   Para ambos conjuntos de datos baje los datos para todas las provincias, cantones y distritos y asegúrese que los datos están limpios antes de hacer el *join* por distrito para asegurarse que la mínima cantidad de datos no es tomada en cuenta. **Es decir deben verificar que en ambos conjuntos de datos los distritos estén escritos de la misma forma.**

   Los conjuntos de datos:

   a) El OIJ publica datos sobre criminalidad en Costa Rica que tienen como fuente las denuncias interpuestas directamente ante esta entidad nacional. Los datos recopilados por el OIJ están disponibles por provincia, cantón o distrito (deben ser bajados como hoja electrónica para contar con el dato de distrito porque en otros formatos el archivo presenta errores). Los datos están disponibles en [2]. El conjunto de datos de criminalidad del OIJ posee las siguientes columnas:

      - Delito:Tipo de Delito

- SubDelito: Tipo de SubDelito
- Fecha: Fecha del Hecho
- Hora: Rango de 3 horas del Hecho
- Victima: Descripción de la Víctima
- SubVictima: Descripción de la SubVíctima
- Edad: Grupo de Edad que pertenece la Víctima
- Genero: Género de la Víctima
- Nacionalidad: Nacionalidad de la Víctima
- Provincia: Provincia del Lugar del Hecho
- Canton: Cantón del Lugar del Hecho
- Distrito: Distrito del Lugar del Hecho

b) El INEC es la institución encargada a nivel nacional de la generación y divulgación de datos estadísticos obtenidos por medio de censos, encuestas y otros estudios sobre demografía, economía y otros. Al igual que en el OIJ los datos están disponibles por distrito. Los datos a utilizar, fueron generados por el INEC como resultado del censo realizado en el país en el año 2011. Los datos están disponibles en [1]. El conjunto de datos posee las siguientes columnas:

- Provincia, Cantón y Distrito
- Población de 15 años y más
- Tasa neta de participación
- Tasa de ocupación
- Tasa de desempleo abierto
- Porcentaje de poblacion economicamente inactiva
- Relación de depedencia económica

2. Seleccione las variables que a utilizar en el ejercicio (**al menos cinco variables que incluyan la cantidad de delitos por distrito, dato que debe calcular**). Documente el motivo de la selección de acuerdo al problema en estudio.

3. **Utilice el algoritmo K-Means implementado en el ejercicio 1** para caracterizar los datos usando las variables seleccionadas.

4. Utilice el **método del codo** para seleccionar el mejor K y vuelva a ejecutar el algoritmo usando el K recomendado.

5. **Utilice el algoritmo DBSCAN implementado en el ejercicio 1** para caracterizar los datos usando las variables seleccionadas.
6. Investigue sobre la mejor forma de evaluar los algoritmo K-Means y DBSCAN y **documente su investigación** (e incluya en el cuaderno de Jupyter a presentar al menos dos de los métodos de evaluación encontrados). Aplique uno de los métodos a la evaluación de los clusteres resultantes de los ejercicios anteriores.
7. **Compare los resultados de ambos algoritmos y genere y documente sus conclusiones** (incluya al menos cuatro conclusiones importantes).
8. Incluya referencias bibliográficas en formato APA.

Sección 2. PCA

Aplique el método de PCA a los datos del OIJ integrados con los datos del INEC.

1. Cargue el archivo, muestre estadísticas del conjunto de datos y luego escale los atributos.
2. Aplíquele la técnica de PCA para reducir la dimensionalidad del conjunto de datos.
3. Grafique la varianza explicada.
4. Realice al menos tres conclusiones importantes sobre el ejercicio realizado.
5. Documente apropiadamente el código generado.
6. Incluya referencias bibliográficas en formato APA.

Fuentes de datos

[1] Instituto Nacional de Estadísticas y Censos (2011). Censo 2011: Indicadores económicos, según provincia, cantón y distrito. Recuperado de http://inec.cr/documento/censo-2011-indicadores-economicos-segun-provincia-canton-y-distrito

[2] Organismo de Investigación Judicial (2018). Estadísticas poiciales. Recuperado de https://sitiooij.poder-judicial.go.cr/index.php/apertura/transparencia/estadisticas-policiales

Rúbrica

Rubro	Puntos
1) Clustering	
Ejercicio 1. Implemente K-Means y DBSCAN.	
Se implementó de cero el algoritmo de K-Means para vectores de atributos de cualquier dimensión sin utilizar ninguna implementación de biblioteca.	10
Se implemento de cero el algoritmo DBSCAN para vectores de atributos de cualquier dimensión sin utilizar ninguna implementación de biblioteca. .	10
Se probó ambos algoritmos con datos generados artificialmente.	2
Se visualizaron ambos resultados.	3
Ejercicio 2.	
Se cargaron e integraron apropiadamente los datos del INEC y del OIJ	5
Se seleccionaron las variables a utilizar en el ejercicio (al menos cinco variables que incluyan la cantidad de delitos por distrito). Se documentó el motivo de la selección de acuerdo al problema en estudio.	2
Se utilizó el algoritmo K-Means implementado en el ejercicio 1 para caracterizar los datos usando las variables seleccionadas.	3
Se utilizó el método del codo para seleccionar el mejor K y se volvió a ejecutar el algoritmo usando el K recomendado.	**2**
Utilice el algoritmo DBSCAN implementado en el ejercicio 1 para caracterizar los datos usando las variables seleccionadas.	**3**
Se investigó sobre la mejor forma de evaluar los algoritmos K-Means y DBSCAN y se **d**ocumentó la investigación y se aplicó uno de los métodos a la evaluación de los clusteres resultantes de los ejercicios anteriores.	**5**
Se compararon los resultados de ambos algoritmos y se generaron y documentaron conclusiones (incluya al menos tres conclusiones importantes).	**3**
2) PCA	
Se desplegaron estadísticas y se escalaron los datos.	**2**
Se aplicó la técnica de PCA para reducir la dimensionalidad del conjunto de datos.	**20**
Grafique la varianza explicada.	**3**
Realice al menos tres conclusiones importantes sobre el ejercicio realizado.	**3**
Consideraciones para todos los ejercicios	
Todas las secciones de los ejercicios deben estar bien documentadas.	**5**
Cuide la redacción y ortografía	**1**
Se incluye una sección de referencias en formato APA.	**3**