MINERÍA DE DATOS Y MODELIZACIÓN PREDICTIVA

Tarea de ACP + Clustering

Pablo Flores Vidal

Noviembre 2023

Introducción

La sección de datos de National Geographic te ha encargado que como muestra de tus habilidades, y para poder optar a un empleo en su sucursal en Trödheim (Noruega) que trates de resolver y realizar lo mejor posible esta práctica. Explora el conjunto de datos y aplica las técnicas de reducción de dimensionalidad (ACP) así como de agrupamiento (clustering) vistas en clase a través de la realización de una serie de apartados (ver más abajo).

Descripción del Conjunto de Datos 'penguins'

El conjunto de datos `'penguins'` de la librería `'seaborn'` de Python contiene la siguiente información sobre diferentes especies de pingüinos:

- **species:** Es la especie de pingüino. Hay tres especies en el conjunto de datos: 'Adelie', 'Chinstrap' y 'Gentoo'.
- **island:** Representa la isla donde se recopilaron los datos. Las islas son 'Biscoe', 'Dream' y 'Torgersen'.
- **bill_length_mm:** Longitud del pico en milímetros.
- **bill_depth_mm:** Profundidad del pico en milímetros.
- **flipper_length_mm:** Longitud de la aleta en milímetros.
- **body_mass_g:** Masa corporal del pingüino en gramos.
- **sex:** Género del pingüino, con las categorías 'Male' (macho), 'Female' (hembra) o 'NaN' si la información no está disponible.

Apartados a realizar

Comienza con estas tres tareas de carga y exploración del dataset (0.5 puntos):

- Carga el conjunto de datos de Palmer Penguins utilizando el Python. Este conjunto de datos se encuentra dentro del paquete *seaborn*.
- Muestra las estadísticas básicas descriptivas sobre el conjunto de datos.
- Utiliza un gráfico de dispersión como el visto al principio del tema de clustering (ver apuntes) más alguna otra visualización relevante y apropiada para este tipo de datos para obtener información sobre la estructura de los datos.

Una vez te has familiarizado con los datos, tu objetivo es reducir el número de variables y explorar relaciones entre las características físicas de los pingüinos, así como entre las especies. Para ello:

1. Calcula la matriz de correlaciones y su representación gráfica: ¿Cuáles son las variables más correlacionadas de forma inversa entre las características físicas de los pingüinos? (0.5 pts)

2. Realiza un análisis de componentes principales (PCA) con los datos estandarizados, calculando un número adecuado de componentes (máximo 4): Estudiar los valores de los autovalores obtenidos y las gráficas que los resumen. ¿Cuál es el número adecuado de componentes para representar eficientemente la variabilidad de las especies de pingüinos? (1 pt)

3. Realiza nuevamente el análisis de componentes principales sobre los datos estandarizados, pero esta vez indicando el número de componentes principales que hemos decidido retener. Sobre este análisis, contestar los siguientes apartados: (2 pts)

   a) Comenta los gráficos que representan las variables en los planos formados por las componentes: Intenta explicar lo que representa cada componente en términos de las características físicas de los pingüinos.

   b) Sobre los gráficos que representan las observaciones en los nuevos ejes: Teniendo en cuenta la posición de las especies de pingüinos en el gráfico, ¿cuáles destacan más en cada componente?

   c) Si tuviéramos que construir un índice que valore de forma conjunta las características físicas de un pingüino, como se podría construir utilizando una combinación lineal de todas las variables: ¿Cómo podríamos construirlo? ¿Cuál sería el valor de dicho índice para una especie de pingüino representada por el conjunto de datos? Por ejemplo, ¿cuál sería el valor en la especie 'Adelie'? ¿Y en la especie 'Chinstrap'?

Ahora es el turno de las técnicas de agrupamiento (clustering)

4. Determina el Número de Grupos: Aplica el agrupamiento jerárquico al conjunto de datos y utiliza un dendrograma para sugerir un número razonable de grupos. Justifica tu elección del número de grupos. Realiza estos pasos tal y como has visto en clase y figura en los apuntes del tema. (1 pt)

5. Agrupamiento K-Means: Implementa el algoritmo de agrupamiento k-means en el conjunto de datos. Experimenta con diferentes valores de k y utiliza métricas apropiadas (por ejemplo, método del codo) para determinar el número óptimo de grupos. (1 pt)

6. Validación del Agrupamiento: Aplica métricas de validación del agrupamiento (por ejemplo, puntuación de silueta) para evaluar la calidad de los resultados del agrupamiento. Discute la efectividad del algoritmo de agrupamiento en capturar la estructura inherente de los datos tal y como se ha visto en clase. (1 pt)

7. Jerárquico versus K-Means: Compara y contrasta los resultados obtenidos del agrupamiento jerárquico y el agrupamiento k-means. Discute similitudes o diferencias en las asignaciones de los grupos. (1 pt)

8. Proporciona una interpretación de los grupos. ¿Qué representan los grupos identificados en el contexto de las especies de pingüinos? ¿Existen patrones o tendencias significativas? (1 pt)

9. Resume tus hallazgos y concluye el análisis. Discute limitaciones o desafíos encontrados durante el proceso de agrupamiento. (1 pt)