UNIVERSIDAD DE GRANADA

UNIVERSIDAD DE GRANADA

Departamento de Ciencias de la Computación e Inteligencia Artificial

Inteligencia de Negocio

Guión de Prácticas

Práctica 2:
Visualización y Segmentación.

Curso 2020-2021

Cuarto Curso del Grado en Ingeniería Informática

Sobre esta práctica

1. Objetivos y Evaluación

En esta segunda práctica de la asignatura Inteligencia de Negocio veremos el uso de visualización para analizar un *dataset*, y las técnicas de aprendizaje no supervisado para análisis relacional mediante segmentación. Se trabajará con un conjunto de datos sobre el que se aplicarán distintos algoritmos de agrupamiento (clustering). A la luz de los resultados obtenidos se deberán crear informes y análisis lo suficientemente profundos.

La práctica se calificará hasta un **máximo de 2 puntos**. Se valorará el acierto en los recursos de análisis gráficos empleados, la complejidad de los experimentos realizados, la interpretación de los resultados, la organización y redacción del informe, etc.

2. Entrega

La fecha límite de entrega será el martes **1 de diciembre** de 2020 hasta las **23:59**. La entrega se realizará a través de una tarea en Prado. En un único fichero `zip` se incluirá la documentación, los scripts de Python empleados y cualquier otro archivo que el alumno considere relevante. El nombre del archivo `zip` será el siguiente (sin espacios): `P2-apellido1-apellido2-nombre.zip`. La documentación tendrá el mismo nombre pero con extensión `pdf`. Es decir, la alumna "María Teresa del Castillo Gómez" subirá el archivo `P2-delCastillo-Gómez-MaríaTeresa.zip` que contendrá, entre otros, el archivo `P2-delCastillo-Gómez-MaríaTeresa.pdf`.

Apartado 1

Visualización (0.5)

En esta primera parte se aplicarán los conceptos del seminario de visualización para realizar expresamente.

Sobre los datos del *dataset* de mamografías de la práctica anterior, se deberán de realizar distintas visualizaciones y análisis, ajustándose al siguiente esquema:

1. **Visualización de las medidas:** Se mostrará de forma visual los resultados de las medidas para cada preprocesamiento de la práctica anterior, analizando las diferencias entre modelos dado un preprocesamiento, y también analizando las diferencias entre preprocesamientos de un mismo modelo.

2. **Gráficas de curva ROC**: Se incluirá la curva ROC de los distintos modelos presentados en la práctica anterior, sobre la misma figura.

3. **Análisis de los atributos**: se incluirá una sección realizando una serie de representaciones para mostrar la relación entre los distintos atributos y el atributo objetivo (la severidad). Paralelamente, se usarán dichas representaciones para analizar la posible relación entre ellos.

Apartado 2

Segmentación (1.5)

1. Descripción del Problema: accidentes mortales de tráfico en España

Una compañía aseguradora quiere comprender mejor las dinámicas en accidentes de tráfico en España. Para ello, a partir de diversas variables que caracterizan el accidente, se pretende encontrar grupos de accidentes similares y relaciones de causalidad que expliquen tipos y gravedad de los accidentes. Para ello se cuenta con los datos publicados por la Dirección General de Tráfico (DGT) en https://sedeapl.dgt.gob.es/WEB_IEST_CONSULTA/subcategoria.faces que incluye información desagregada (microdatos) de más de 30 variables entre los años 2008 y 2015. En esta práctica, nos centraremos en los datos para el año 2013 (89.519 accidentes). En la web de la asignatura se incluye el conjunto de datos —procesado a partir de la fuente original— sobre el que se trabajará en esta práctica.

2. Tareas a Realizar

La práctica consiste en aplicar y analizar técnicas de agrupamiento para descubrir grupos en el conjunto de datos bajo estudio. El trabajo se realizará empleando bibliotecas y paquetes de Python, principalmente numpy, pandas, scikit-learn, matplotlib y seaborn. Se recomienda consultar los siguientes enlaces:

- http://scikit-learn.org/stable/modules/clustering.html
- http://www.learndatasci.com/k-means-clustering-algorithms-python-intro/
- http://hdbscan.readthedocs.io/en/latest/comparing_clustering_algorithms.html
- https://joernhees.de/blog/2015/08/26/scipy-hierarchical-clustering-and-dendrogram
- http://seaborn.pydata.org/generated/seaborn.clustermap.html

Nos interesaremos en segmentar los accidentes seleccionando previamente grupos de interés según las variables categóricas. Por ejemplo, analizar solo accidentes donde ha habido colisión de vehículos, solo en carreteras urbanas, en determinadas condiciones ambientales, etc. También se puede considerar la hora y/o día de la semana para seleccionar subgrupos de interés (fines de semana, horario de mañana, etc.). Queda a elección libre del alumno escoger dos casos de uso y realizar el estudio sobre cada uno de ellos. Será necesario también aplicar una normalización para que las métricas de distancia y la visualización funcionen correctamente. Deberán justificarse las decisiones tomadas respecto al tratamiento de las variables.

Para elegir los atributos a utilizar se recomienda apoyarse en visualizaciones tal y como se explicó en el seminario de visualización, pero no es necesario incluir las gráficas, sólo alguna que se considere de interés.

En cada caso de estudio se analizarán 2 algoritmos distintos de agrupamiento (siendo al menos uno de ellos K-means obteniéndose para cada algoritmo las métricas de rendimiento Silhouette [1] y el índice Calinski-Harabaz [2]. Además, es necesario gráficas de los centroides como la de la Figura 2.1 para ayudar a interpretar el significado de cada grupo.

Adicionalmente, sobre el algoritmo K-means se analizará el efecto del parámetro k usando unos pocos valores.

A partir de los resultados obtenidos se deberán extraer conclusiones sobre los grupos de población. Se valorará el acierto en la selección de casos de estudio que mejor reflejen los grupos encontrados en los datos.

3. Esquema de la Documentación

La documentación de esta parte deberá ajustarse al siguiente esquema (debe respetarse la numeración y nombre de las secciones):

1. **Introducción**: se hablará sobre el problema abordado y todas las consideraciones generales que se deseen indicar.

2. **Caso de estudio X**: se incluirá una sección por cada caso de estudio analizado. En ella se explicará en detalle en un primer apartado qué caso se analiza y por qué. Se incluirá una tabla comparativa con los resultados de los algoritmos de clustering. Cada sección contendrá las visualizaciones necesarias para analizar el problema. Se añadirá un apartado final titulado "Interpretación de la segmentación" que incluirá las conclusiones generales a las que haya legado el alumno a la luz de los resultados en el correspondiente caso de estudio. En cada sección podrán incluirse extractos de los scripts que el alumno considere relevante para destacar el trabajo realizado.

[1] se recomienda ver la https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html y https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

[2] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.calinski_harabasz_score.html

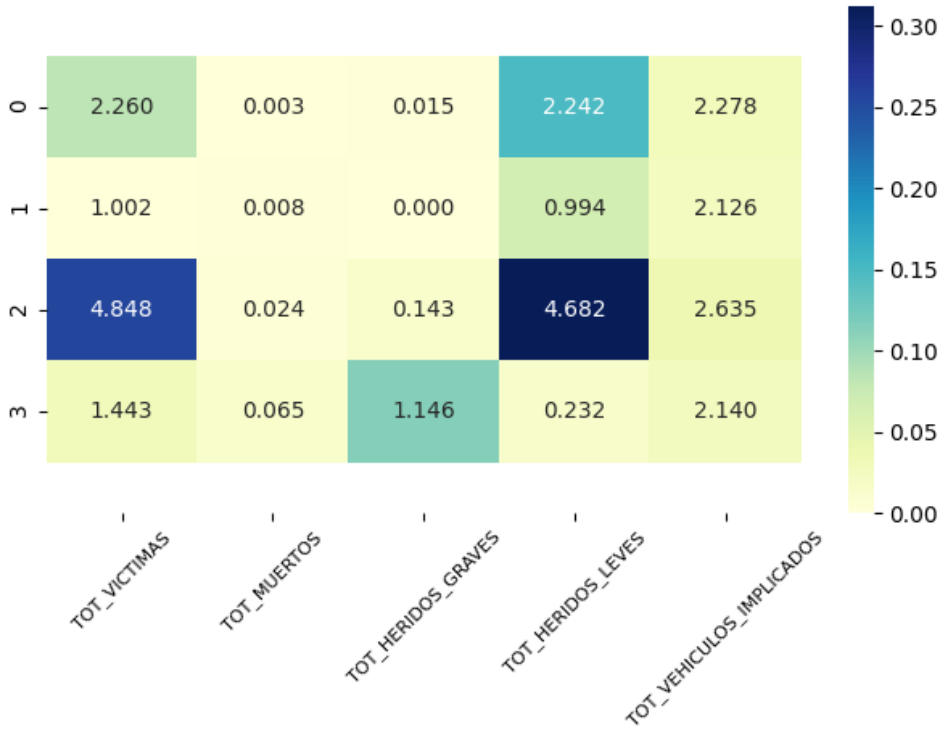

Figura 2.1: Centros de los grupos

3. **Contenido adicional**: opcionalmente, cualquier tarea adicional a las descritas en este guión puede presentarse en esta sección.

4. **Bibliografía**: referencias y material consultado para la realización de la práctica.

La primera página de la documentación incluirá una portada con el nombre completo del alumno, grupo de prácticas y dirección email. También se incluirá una segunda página con el índice del documento.