Profesor: Dr. Oldemar Rodríguez Rojas
Análisis de Datos 1
Fecha de Entrega: Viernes 5 de abril - 8 a.m.

Tarea Número 2

-
- Las tareas se pueden realizar en grupos de máximo 3 personas, tareas idénticas se les asignará cero puntos.
- Todas las tareas tienen el mismo valor en la nota final del curso, es decir, el promedio de las notas obtenidas en la tareas será la nota final del curso.
- Todos los ejercicios tienen el mismo valor.

1. [50 puntos] En este ejercicio vamos a usar la tabla de datos `beansV2.csv` que tiene los resultados de una investigación donde se utilizaron siete tipos diferentes de frijoles secos, teniendo en cuenta las características como forma, tipo y estructura según la situación del mercado. Se desarrolló un sistema de visión por computadora para distinguir siete variedades diferentes registradas de frijol seco con características similares para obtener una clasificación uniforme de semillas. Para el modelo de clasificación se tomaron imágenes de 4,767 granos con una cámara de alta resolución. Las imágenes de frijoles obtenidas por el sistema de visión por computadora se sometieron a etapas de segmentación y extracción de características; de los granos, así se obtuvieron 12 dimensiones y 4 formas. Esta tabla de datos es una muestra de un dataset más grande llamado `beansV2.csv` que se puede encontrar en el sitio web de UCI Machine Learning Repository.

   La tabla contiene 4767 filas y 17 columnas, las cuales se explican a continuación.

   - `Area:` El área de una zona de frijol y el número de píxeles dentro de sus límites. (Numérica)
   - `Perimeter:` La circunferencia del frijol se define como la longitud de su borde (Numérica)
   - `MajorAxisLength:` La distancia entre los extremos de la línea más larga que se puede dibujar de un frijol (Numérica)
   - `MinorAxisLength:` La línea más larga que se puede dibujar desde el frijol mientras se encuentra perpendicular al eje principal.
   - `AspectRation:` Define la relación entre L y l.
   - `Eccentricity:` Excentricidad de la elipse que tiene los mismos momentos que la región.
   - `ConvexArea:` Número de píxeles en el polígono convexo más pequeño que puede contener el área de una semilla de frijol.
   - `EquivDiameter:` El diámetro de un círculo que tiene la misma área que el área de una semilla de frijol.
   - `Extent:` La relación de los píxeles en el cuadro delimitador al área del frijol.
   - `Solidity:` También conocida como convexidad. La relación entre los píxeles de la capa convexa y los que se encuentran en los frijoles.

- `Roundness`: Calculado con la siguiente fórmula: $(4\pi A)/(P^2)$.
- `Compactness`: Mide la redondez de un objeto: Ed/L.
- `ShapeFactor1`: Factor de forma 1.
- `ShapeFactor2`: Factor de forma 2.
- `ShapeFactor3`: Factor de forma 3.
- `ShapeFactor4`: Factor de forma 4.
- `Class`: Clase, Seker, Barbunya, Bombay, Cali, Dermosan, Horoz y Sira.

Nota: Todas son variables numéricas salvo la variable `Class` y no tienen `NAs`. Realice lo siguiente:

a) Efectúe un ACP con solo las variables numéricas.

b) Elimine en **R** los individuos y las variables mal representadas (coseno cuadrado menor al 10 %).

c) En el plano principal encuentre 3 clústeres.

d) En el círculo de correlación determine e interprete la correlación entre las variables. Compare las correlaciones con las que se obtienen usando los gráficos que ofrecen paquetes de **R** para visualizar las correlaciones. ¿Cuál cálculo es más exacto? ¿Por qué?

e) Interprete la formación de los 3 clústeres basado en la sobre-posición del círculo y el plano.

f) Convierta la variable `Class` a código disyuntivo, efectúe de nuevo el ACP e interprete de nuevo la formación de los 3 clústeres basado en la sobre-posición del círculo y el plano. ¿Se gana interpretabilidad?

2. [50 puntos] En este ejercicio vamos a usar la tabla `water_potability.csv`, la cual contiene métricas de calidad del agua para 3276 cuerpos de agua diferentes. La tabla contiene 3280 filas y 10 columnas, estas son:

   - `ph`: El pH es una medida de la es una medida de la acidez o basicidad de una solución.
   - `Hardness`: Capacidad del agua para precipitar jabón en mg/L (miligramos por litro)
   - `Solids`: Sólidos disueltos totales en ppm (partes por millón).
   - `Chloramines`: Cantidad de cloraminas en ppm.
   - `Sulfate`: Cantidad de sulfatos disueltos en mg/L.
   - `Conductivity`: Conductividad eléctrica del agua en S/cm (microSiemens/cm).
   - `Organic_carbon`: Cantidad de carbón orgánico en ppm.
   - `Trihalomethanes`: Cantidad de trihalometanos en g/L.
   - `Turbidity`: Medida de la propiedad de emisión de luz del agua en NTU (Unidades de turbidez nefelométrica).
   - `Potability`: Indica si el agua es segura para consumo humano. (Categórica)

   Realice lo siguiente:

a) Cargue los datos en **R**, recuerde transformar la variable `Potability` en categórica, además verifique bien si hay nombres de fila o no, verifique los sepradores de datos y de decimales con un editor de texto.

b) Con **R** efectúe un ACP y dé una interpretación siguiendo los siguientes pasos: 1) elimine individuos mal representados y variables mal representadas (coseno cuadrado menor al 5 %) 2) en el plano principal identifique un cluster en cada cuadrante 3) en el círculo de correlación determine e interprete la correlación entre las variables y 4) explique la formación de los clústeres basado en la sobre-posición del círculo y el plano.

c) En el círculo de correlación, usando los componentes 1 y 3, interprete la correlación entre las variables `Conductivity`, `Trihalomethanes` y `Organic_carbon`, que están mal representadas en los componentes 1 y 2.

d) Ahora desde **R** convierta la variable `Potability` en Código Disyuntivo Completo y repita el ACP ¿Se gana interpretabilidad al convertirla en Código Disyuntivo Completo? En este caso use solamente 2 clústeres para la interpretación. Use en **R** el gráficos 3D, (con algún paquete de graficación 3D y usando 3 componentes principales) para confirmar los resultados e interpretaciones. ¿Por qué el gráfico en 3D aporta resultados más confiables?

Entregables: Debe entregar un documento autreproducible HTML con todos los códigos y salidas, incluya pruebas de ejecución de las funciones programadas. No olvide poner un título para cada pregunta. Las demostraciones las puede entregar en papel a mano.

oldemar rodríguez

CONSULTOR en M1NER14 DE D4T0S