Primer Parcial (30pts)

Programacion Estadística II

Lic. Alvaro Chirino Gutierrez

8/6/2020

Instrucciones

- Duración: 3 horas. 18:15 a 21:15
- Forma de entrega, incluir solo las respuestas en archivos separados, el nombre del archivo debe ser: respuesta_1, respuesta_2, etc.
- La entrega al correo achirino.stat@gmail.com máximo hasta las 21:15, pasado el tiempo se reduciran 5 puntos por cada 10 minutos de retraso
- Incluir en el correo su nombre completo.

Pregunta 1 (5pts).

- Liste y comente las fases del descubrimiento del conocimiento en las bases de datos (KDD)
- Describa que es un warehouse
- Describa la diferencia entre componentes principales y análisis de correspondencia
- Describa la diferencia entre k-center y los clusters jerárquicos
- En el proceso de imputar si una variable alcanza un índice influx alto, esto siginifica que: ¿esta variable es más útil para imputar otras variables?. Si, No, explique.

Pregunta 2 (5 pts):

Empleando la encuesta a hogares 2018, obtenga una tabla que contenga el porcentaje y total de hogares (expandido) por departamento y área, que tengan a algún miembro del hogar con diabetes ó hipertension arterial ó enfermedad del corazón.

Pregunta 3 (10 pts)

- PCA: Empleando la base de datos de la encuesta a hogares 2018, seleccione solo al jefe del hogar y calcule el PCA con las variables; mujer, edad, años de educación, horas trabajadas a la semana, ingreso laboral, ingreso no laboral, ingreso percapita del hogar. Identifique la cantidad de componentes a retener con el criterio de eigenvalores que superen la unidad. Calcule los componentes principales retenidos y grafique sus histogramas.
- CA: Empleando la base de datos de las ENDSA para el año 2008, para las 10 variables de violencia tranformar a binarias como (1=frecuentemente ó a veces 0=Nunca) y realizar el MCA para las 10 variables de violencia transformadas incluyendo la variable sexo. Comentar los resultados.

Pregunta 4 (10 pts)

Clustering: Empleando la base de datos de las elecciones del 20 de octubre, aplique de forma separada para los municipios y los países, en términos relativos sin considerar los blanco y nulos:

- El Cluster k-center, identifique el mejor valor de k entre 2 al 10, el mejor centro; media, mediana o medoide, y la mejor distancia; euclideana, manhattan
- El Cluster jerarquico, identifique el mejor valor de k entre 2 al 10, el mejor método; complete, single, average y la mejor distancia; euclideana, manhattan

Pregunta 5 (Opcional, 5 pts):

Programe una función que calcule el coeficiente de silueta dada una matriz de distancias y un vector de identificación del cluster