Aplicaciones de minería de datos en Ciencia y Tecnología **Examen Final**	17/12/2015
Las respuestas correctas están resaltadas	

Pregunta 1. Cuáles de las siguientes afirmaciones acerca del coeficiente de Silhouette son verdaderas o falsas:

- Cuando se lo aplica a un clustering difuso, tiene en cuenta el grado de pertenencia de cada objeto a cada cluster.
- Si se establecen puntos de corte, se puede utilizar para evaluar clusters jerárquicos.
- Es una métrica que combina conceptos de cohesión y separación.

Comentario: Al determinar un punto de corte sobre un cluster jerárquico, se definen grupos como los que se obtienen con métodos de clustering no jerárquicos, como k-medias. Sobre estos grupos se puede aplicar Silhouette.

Pregunta 2. La técnica de clustering por densidad de DBSCAN se puede aplicar cuando:

- Las nubes de punto de cada cluster no se pueden aproximar por elipsoides.
- Las nubes de punto de cada cluster no se pueden aproximar por elipsoides. Además hay puntos que pueden quedar excluidos de cualquier cluster.
- Es una técnica que solo sirve para variables categóricas multiestado no ordenadas.

Pregunta 3. ¿Cuáles de las siguientes afirmaciones son ciertas con respecto al método de clustering PAM?

- Un medoide de PAM es lo mismo que en K-medias definimos como un centroide.
- Con PAM se obtiene una estimación de la pertenencia de cada objeto al cluster asignado.
- Con PAM es posible agrupar variables categóricas multi-estado

Comentario: Son todas falsas. Un medoide no eslo mismo que un centroide, no desde el punto de vista estrictamente matemático ni desde el punto de vista conceptual. El medoide es un objeto real; el centroide, no. A su vez el centroide es representativo del centro de masa del cluster, mietras que el mededide puede quedar desplazado.

Pregunta 4. Determinar si las siguientes afirmaciones son verdaderas o falsas:

- El coeficiente de correlación cofenético es un indicador de la distorsión que introduce representar una matriz de distancia con un dendograma
- Es una alternativa al uso de mapas de calor para evaluar las calidades de los agrupamientos.
- Puede tomar valores valores negativos o positivos. Para saber que la calidad del agrupamiento es bueno, solo hace falta que su valor absoluto sea alto.

Comentario: el uso del coeficiente de correlación cofenética (ccc) y del mapa de calor son técnicas que se complementan. El mapa de claor es una visualización que sirve para identificar problemas en el agrupamiento, el ccc es un único número.

Pregunta 5. En el trabajo práctico 1, la normalización de los datos espectrales contra la

emisión a 280 nm (S280MAG) se realizó para:

- Poder analizar las intensidades independientemente del brillo absoluto o aparente de la galaxia.
- Reducir la colinealidad entre variables.
- Ninguna de las dos.

Pregunta 6. ¿Cuál es la afirmación correcta de estas cuatro?

- Un estimador de densidad kernel es una herramienta para evitar algunos de los problemas que tienen los histogramas.
- Se obtiene un estimador de frecuencias sobre el que se puede hacer diferenciación numérica.
- Las dos anteriores.
- Ninguna de las tres anteriores.

Comentario: los estimadores de densidad kernel son estimadores suavizados de la distribución, constituida por una secuencia de números en intervalos uniformes, sobre los que se puede aplicar métodos de diferenciación numérica para revelar cambios en la distribución, como pequeños hombros, que pueden pasar desapercibidas en los datos sin diferenciar.

Pregunta 7. ¿Qué se puede decir sobre la siguiente afirmación? Los mapas autoorganizativos son técnicas de agrupamiento similares a k-medias o PAM en el sentido que no se pueden establecer relaciones entre las clases.

verdadero - falso

Comentario: k-medias y PAM producen grupos disjuntos, sin relación entre ellos. En los SOM las grupos lindantes son más similares entre sí que con respecto a grupos más alejados.

Pregunta 8. Indicar si es verdadero o falso

- PMML es un estándar para implementar todos los puntos que cubre la metodología CRISP.
- Cada vez que sale una nueva versión de PMML las herramientas de análisis que usan el estándar deben incorporarla a sus productos en un lapso de seis meses.

Comentario: ambos puntos son falsos

Pregunta 9. Las políticas de gobernancia de datos pueden servir para … (marcar las correctas):

- implementar políticas de privacidad de datos.
- documentar cuando ciertos datos deben ser eliminados.
- reemplazar a CRISP para implementar estrategias de limpieza de datos.

Pregunta 10. ¿Qué se puede decir sobre la siguiente afirmación? Las correcciones de la familia FDR (false discovery rate) no se pueden usar cuando hay que hacer muchas comparaciones (por ejemplo, 50 comparaciones). En ese caso es mejor usar la corrección de Bonferroni.

verdadero - falso

Comentario: en primer lugar, hay que hacer correcciones. Luego, la corrección de Bonferroni es muy conservadora y en casos donde hay que hacer muchas correcciones, es posible que ninguna resulte significativa. En esos casos es mejor usar alguna variante

de los métodos FDR.

Pregunta 11. ¿Cuáles de estas afirmaciones son correctas con respecto al coeficiente de información máximo (MIC?

- MIC solo sirve para detectar asociaciones lineales o polinomiales entre pares de variables.
- Solo sirve para analizar variables de a pares.

Pregunta 12. Indicar si es verdadero o falso.

- El análisis de las distribuciones de grado solo es relevante para determinar si la red es de mundo pequeño o no.
- Las medidas de centralidad de grado, intermediación y autovalor son diferentes formas matemáticas para describir las mismas características de un vértice.
- El uso de pesos en las aristas afecta al cálculo de caminos mínimos entre vértices

Comentario: no solo que matemáticamente las medidas de centralidad de grado, intermediación, etc. son diferentes, sino que conceptualmente están indicando distintas características topológicas de los nodos.

Pregunta 13. Una aplicación frecuente de los grafos bipartitos es en el análisis de co-citaciones, que se usa por ejemplo para analizar la forma en que diversas personas colaboran en la autoría de trabajos o en la participación en grupos musicales.

- ¿Qué representan las aristas en estas redes?

  Las aristas representan las colaboraciones o las participaciones. Sólo pueden unir nodos de grupos diferentes, por ejemplo, un músico con una banda.

- ¿Cuáles serían los vértices?

  Un grupo de vértices representa a los autores o músicos; el otro grupo de vértices representa a los trabajos escritos o a las bandas musicales.

Pregunta 14. Para determinar la significancia de un grupo de comunidades hallados dentro de una red, se puede usar la modularidad del agrupamiento contra modelos nulos de comunidades.

Verdadero - Falso

Pregunta 15. ¿Cuáles de las siguientes afirmaciones son correctas con respecto a las ontologías?

- Las ontologías son implementaciones de grafos bipartitos.
- Una ventaja de representar el conocimiento mediante ontologías es que se pueden construir prescindiendo del experto del dominio
- En una ontología la similitud semántica se refiere a la cercanía entre términos en el grafo que se usa para representar el conocimiento del dominio.

Pregunta 16. ¿Por qué la similitud coseno no es una buena medida de similitud semántica aplicado a ontologías? (indicar verdadero/falso para cada opción)

- Es un estimador similar a cualquier otro de los disponibles para similitudes semánticas. El problema es que no existen buenos estimadores de similitud

semántica para ontologías.

- No considera la información que se puede obtener de la organización de la ontología como grafo.
- No tiene un estimador estadístico asociado