UNIVERSIDAD COMPLUTENSE MADRID

ntic master

TEST DE EVALUACIÓN

Master data science, big data & business analytics

Minería de Datos y Modelización Predictiva

Carlos Balañá Arias

La entrega de esta parte de test se realiza en la plantilla Excel adjunta indicando, para cada pregunta, la única respuesta correcta. Todas las preguntas tienen el mismo valor y las respuestas incorrectas no penalizan.

Esta parte de test supone el 40% de la calificación del módulo.

1. Pregunta 1

¿Cuáles son algunos de los principales desafíos que enfrenta la minería de datos?

a) Privacidad y seguridad, Volumetría y Despliegue en producción.
b) Calidad de los datos, Reparación de errores en fuentes origen y Desbalanceo de clases
c) Volumetría, Interpretabilidad y Privacidad y Seguridad.

2. Pregunta 2

Ante un modelo con muchas variables que modeliza linealmente una variable dependiente cuando en realidad ésta tiene un comportamiento polinómico. Y que además es capaz de explicar de forma muy pobre la variabilidad de la variable objetivo, se dice que presenta:

a) Un problema de varianza asociado a la generalización del modelo en el conjunto de test.
b) Un problema de sesgo debido a una elección poco adecuada del modelo apropiado.
c) No presenta problemas de sesgo o varianza, tan solo necesita profundizar en la fase de ajuste de los parámetros del modelo.

3. Pregunta 3

¿Por qué el cálculo del coeficiente V de Cramer es equivalente al coeficiente de correlación para variables categóricas?

a) El coeficiente V de Cramer a través del coeficiente χ^2, mide la fuerza en la relación de dependencia entre 2 variables categóricas, al comparar las frecuencias observadas con las frecuencias esperadas asumiendo independencia.
b) El coeficiente V de Cramer mide la covarianza normalizada para representar la relación entre 2 variables categóricas
c) El estadístico χ^2 mide la relación de dependencia entre ambas variables categóricas y el V de Cramer se calcula a su vez como el estadístico χ^2 dividido por la raíz cuadrada del número de observaciones por la dimensión menor de la tabla de contingencia.

4. Pregunta 4

Al hacer el tratamiento de outliers en la fase de Data Preparation, la diferencia entre la transformación y la sustitución de los outliers radica en que:

a) Al aplicar una transformación modificamos la distribución subyacente de la variable sobre la que estamos actuando mientras que al aplicar sustitución buscamos reemplazar aquellos valores considerados atípicos sin modificar la distribución.
b) Al aplicar sustitución, en general se suele modificar características relevantes como la simetría y curtosis de la distribución, mientras que al transformar la variable mantenemos la distribución subyacente de la variable.
c) En ambos métodos modificamos exclusivamente los valores considerados atípicos mientras, pero difieren en que la forma en que se detectan los outliers.

5. Pregunta 5

Uno de los supuestos más importantes que se han de cumplir para que un modelo de regresión lineal pueda ser aplicable a la hora de estimar el valor de una variable objetivo es:

a) La variable objetivo de presentar homocedasticidad en sus valores, es decir, para distintas muestras tomadas aleatoriamente, la varianza debe ser la misma.
b) La varianza que presenten los residuos estimados por el modelo de regresión lineal tenga varianza constante.
c) Los errores del modelo deben presentar heterocedasticidad para evitar introducir sesgo en la estimación del error cometido en el cálculo de los coeficientes.

6. Pregunta 6

Al incluir variables categóricas en un modelo de regresión lineal hay diversas codificaciones que se pueden emplear para convertirlas en numéricas. Una de ellas consiste en incluir variables dummy para las distintas categorías de la variable. En relación con los modelos de regresión, al aplicar esta técnica:

a) Debemos incluir de forma binaria tantas variables como categorías tenga la variable para que todas queden representadas o podría haber problemas de pérdida de información.
b) Debemos incluir tantas variables como categorías menos una ya que su información irá embebida en el coeficiente intercepto y hará que no haya colinealidad entre las nuevas variables.
c) Debemos incluir tantas variables como categorías ya que, si no lo hacemos así, la matriz del cálculo de coeficientes no sería invertible impidiendo su estimación.

7. Pregunta 7

En base a esta afirmación, seleccionar la opción más adecuada: “El coeficiente R^2 y el R^2 ajustado son las únicas métricas que permiten analizar la bondad de nuestro modelo de regresión, pero no valoran la complejidad del modelo.”

a) Es falsa, porque no es la única métrica que valora la bondad del modelo y porque una de ellas valora la complejidad del modelo.
b) Es falsa porque no es la única métrica.
c) Es falsa porque ambas tienen en cuenta la complejidad del modelo.

8. Pregunta 8

Seleccionar la opción más certera en base a la siguiente afirmación: "Los modelos de regresión logística se utilizan solo para llevar a cabo análisis de clasificación de variables dicotómicas."

a) La sentencia es correcta
b) La sentencia es incorrecta puesto que pueden clasificar entre múltiples categorías
c) La sentencia es incorrecta puesto que se trata de modelos de predicción y no de clasificación.

9. Pregunta 9

Un médico va a hacer uso del modelo para decidir si intervenir quirúrgicamente (categoría positiva) o no (categoría negativa) a sus pacientes. Hay un riesgo elevado de problemas por la intervención. En este contexto y atendiendo a los resultados de la matriz de confusión de un modelo de regresión logística, ¿qué debería primar a la hora de valorar el modelo?

a) Una alta tasa de recall (sensibilidad)
b) Una baja tasa de falsos negativos
c) Una alta tasa de precisión

10. Pregunta 10

En una regresión logística, ¿qué interpretación podemos hacer de los coeficientes de las variables?

a) El logaritmo de los odds aumenta e^{β_m} veces por cada incremento en una unidad de la variable x_m.
b) Los odds aumentan β_m veces por cada incremento en una unidad de la variable x_m.
c) Los odds aumentan e^{β_m} veces por cada incremento en una unidad de la variable x_m.

11. Pregunta 11

El método de selección de variables de tipo 'wrapper' conocido como 'backward':

a) Se trata de un método secuencial que, partiendo de un modelo vacío, va añadiendo en cada iteración la variable que más aumente la bondad del modelo
b) Se trata de un método secuencial que partiendo de un modelo con todas las variables va eliminando en cada iteración aquella variable que, por haberla quitado, haya generado mayor cambio en la bondad del modelo
c) Se trata de un método secuencial que partiendo de un modelo con todas las variables va eliminando en cada iteración aquella variable que, por haberla quitado, haya generado menor cambio en la bondad del modelo.

12. Pregunta 12

En una serie temporal con datos corto-medio plazo, ¿cuáles son las principales componentes en que se descompone la serie?

a) Tendencia, estacionariedad y ruido
b) Ruido estacionalidad y tendencia
c) Ciclo, tendencia y estacionalidad

13. Pregunta 13

Cuando tratamos de transformar en estacionaria una serie, realizamos varios pasos en los que en cada uno eliminamos una componente para hacer constante su distribución en el tiempo. Cuando aplicamos diferenciación estamos tratando de eliminar:

a) La tendencia
b) La estacionalidad
c) La heterocedasticidad

14. Pregunta 14

¿Qué tipo de suavizado exponencial es más apropiado para una serie temporal que presenta estacionalidad y tendencia?

a) Suavizado doble de Holt
b) Alisado simple
c) Suavizado de Holt-Winters

15. Pregunta 15

Un diagrama de autocorrelación ACF, ¿qué característica mostraría principalmente si tenemos una serie que presenta estacionalidad clara con periodo k?

a) Presentaría un pico significativo destacado en el retardo k
b) Mostraría hasta el retardo k coeficientes de autocorrelación significativos
c) Presentaría picos significativos en el retardo k y sus múltiplos

16. Pregunta 16

En un modelo autorregresivo de orden 3, el termino de error…:

a) … de la regresión calculada para obtener la PACF explica la variabilidad no explicada por los retardos de orden 1, 2 y 3.
b) … del modelo se presume que presenta heterocedasticidad, sin correlación y tiene una distribución normal.
c) … debe presentar una varianza con tendencia lineal creciente a lo largo del tiempo, para que el

modelo sea válido.

17. Pregunta 17

En un modelo ARIMA (p,d,q), las siglas AR hacen referencia a la parte de modelo autorregresivo y la p su orden. Las siglas MA hacen referencia a la parte de medias móviles del modelo y q su orden. La sigla I hace referencia a:

a) A la parte integrada del modelo siendo esta una transformación logarítmica de orden d.
b) A la parte integrada del modelo siendo ésta una diferenciación de orden d para reducir la colinealidad de los errores
c) A la parte integrada del modelo siendo ésta una diferenciación de orden d para hacer estacionaria la serie que presenta tendencia.

18. Pregunta 18

El análisis de componentes principales (PCA) maximiza sobre las nuevas componentes la varianza de los datos proyectos. ¿Cómo se lleva a cabo en el proceso matemático del análisis de componentes está maximización?

a) La ortogonalidad de los autovectores obtenidos de la matriz de correlación o covarianza, maximiza la varianza proyectada de los datos sobre dichas componentes.
b) Al obtener los autovectores de la matriz de correlación (o covarianza), obtenemos componentes ortogonales y sus autovalores representan la varianza proyectada de los datos. Por tanto, al seleccionar las componentes con mayores autovalores, maximizamos la varianza.
c) La matriz de correlación o covarianza recoge la información conjunta entre las distintas variables y de esta forma nos aseguramos de que la varianza proyectada sea máxima en nuestras componentes principales.

19. Pregunta 19

Las rotaciones Varimax y Promax que se pueden realizar en un análisis factorial, tienen su principal diferencia en:

a) La rotación Varimax al rotar las componentes deshaciendo la ortogonalidad consigue aumentar la varianza explicada por las componentes, mientas que Promax aumenta la varianza explicada sin deshacerla.
b) La rotación Promax es un punto intermedio entre las rotaciones ortogonales Varimax y Equamax.
c) La rotación Varimax mantiene la ortogonalidad mientras que la Promax es una rotación oblicua que sacrifica parte de la ortogonalidad por una mejora en la varianza explicada por sus componentes.

20. Pregunta 20

Una de las principales diferencias entre el análisis de componentes principales (PCA) y el análisis factorial (AF) es:

a) El análisis PCA busca principalmente reducir la dimensionalidad del conjunto de variables mientras que el análisis factorial busca sacar a la luz la estructura latente que subyace de los datos para su interpretación.
b) La rotación es una operación casi obligatoria en el PCA ya que permite ajustar las cargas. En el AF no es necesario y no suele hacerse.
c) En el cálculo del análisis AF se trabaja con la varianza total de los datos mientras que en PCA se trabaja con la varianza compartida entre variables ignorando la especifica de cada una.

21. Pregunta 21

Supongamos un proyecto para la creación de un asistente virtual documental en una empresa. Tiene una base documental vectorizada donde sus documentos están debidamente vectorizados como embeddings. El proyecto utiliza modelos LLM de lenguaje natural para responder en base a la información disponible. Para determinar la distancia o similitud entre la información contenida en una pregunta de un usuario y la información vectorial de los documentos con el fin de seleccionar el documento adecuado, ¿Cuál sería la métrica de similitud más apropiada para comparar los vectores de la pregunta y la documentación?

a) Distancia euclídea
b) Índice de Jaccard
c) Coeficiente de coseno

22. Pregunta 22

Para un proyecto de una empresa se pretende segmentar a sus clientes en distintos grupos en base a un conjunto de variables o indicadores de cada cliente. Teniendo en cuenta que los grupos no se pretende que tengan potencialmente la misma forma, ¿cuál sería el algoritmo de clustering más apropiado en base a esta condición?

a) K-Means
b) Jerárquico aglomerativo.
c) Ninguno de los anteriores ya que ambos son algoritmos supervisados.

23. Pregunta 23

El criterio de enlace de un algoritmo jerárquico aglomerativo más conservador a la hora de fusionar clusters es:

a) El Simple, puesto que minimiza la distancia más corta y por tanto siempre fusiona clusters donde al menos 2 de los puntos de ambos clusters estén muy cercanos entre sí.
b) El completo ya que fusiona clusters solo cuando los puntos más alejados de ambos clusters suponen la distancia menor entre los posibles clusters a fusionar.
c) El basado en centroide ya que es un termino medio entre ambos y asegura que las fronteras más alejadas no impacten tanto como el centro del cluster a la hora de fusionar.