Tarea Minería de Datos

‘Test de Evaluación’

Autor: Guillermo Villarino

La entrega de esta parte de test se realiza en la plantilla excel adjunta indicando, para cada pregunta, la única respuesta correcta.

Test_Minería_NombreApellido.xlsx

Todas las preguntas tienen el mismo valor y las respuestas incorrectas no penalizan.

Es necesario adjuntar un notebook y html con el código utilizado para las preguntas que lo requieran. Tiene que estar ejecutado sin errores.

Codigos_Test_Minería_NombreApellido.ipynb

Codigos_Test_Minería_NombreApellido.html

Las respuestas prácticas sin código de soporte no contabilizan.

El test consta de 30 preguntas con un valor de 8 puntos. Los 2 puntos restantes (opcionales) se pueden conseguir respondiendo a las 4 preguntas opcionales que aparecen al final del documento.

Pregunta 1

Para el estudio descriptivo bivariante entre dos variables de naturaleza categórica o nominal, ¿cuáles son el gráfico y el estadístico que resultan más adecuados?

a) Boxplot paralelo y tabla de contingencia

b) Gráfico de dispersión y media

c) Diagrama de mosaico y valor Chi2/V de Cramer

d) Gráfico de dispersión y coeficiente de correlación

Pregunta 2

¿Cuáles son los principales riesgos de la imputación simple por la media y de la imputación por modelos multivariantes, respectivamente?

a) Subestimación de la verdadera varianza y sobreestimación de las covarianzas

b) Carga de las colas de la distribución y subestimación de la media

c) Sobreestimación de la verdadera varianza y subestimación de las covarianzas

d) Siempre es mejor imputar por modelos

Pregunta 3

¿Qué diferencia matemática existe entre los modelos de regresión lineal y los modelos de regresión logística en términos de estimación paramétrica?

a) El modelo lineal estima los parámetros por máxima verosimilitud y el modelo logístico por mínimos cuadrados por lo que este último es más sencillo en términos matemáticos

b) El modelo logístico es más sencillo en términos matemáticos

c) La estimación paramétrica es igual en modelos lineales y logísticos

d) El modelo lineal estima los parámetros por mínimos cuadrados y el modelo logístico por máxima verosimilitud por lo que este último es más complejo en términos matemáticos

Pregunta 4

¿Qué riesgo tiene la multicolinealidad en los modelos de regresión?

a) Hace siempre imposible la estimación de parámetros

b) Provoca un aumento de la varianza de los estimadores, haciendo que sean menos robustos

c) Disminuye el valor estimado de los parámetros cambiando la interpretación del modelo

d) La multicolinealidad no es nociva para los modelos de regresión

Pregunta 5

¿Cuál es la hipótesis nula del test de Ljung-Box en series temporales?

a) La media de los residuos es nula

b) Los residuos se distribuyen acorde a una distribución normal

c) Los residuos están exentos de autocorrelaciones significativas

d) Los residuos están autocorrelacionados

Pregunta 6

En una serie temporal univariante que presenta estacionalidad y tendencia con una heterocedasticidad creciente en el tiempo, ¿qué modelo de suavizado aplicarías?

a) Modelo de suavizado doble de Holt

b) Modelo de suavizado de Holt-Winters aditivo

c) Modelo de suavizado de Holt-Winters multiplicativo

d) Modelo de suavizado simple

Pregunta 7

¿Por qué es importante que una serie temporal sea estacionaria de cara las predicciones por un modelo ARIMA?

a) En una serie estacionaria podemos calcular la media con el conjunto de observaciones y conocer la distribución de los errores

b) En una serie estacionaria, la media es 1 por lo que es más fácil predecir hacia delante

c) El método ARIMA es robusto frente a la falta de estacionariedad de la serie

d) En una serie estacionaria los parámetros siempre son menores que en una serie no estacionaria por lo que el modelo es más simple y funciona mejor

Pregunta 8

¿En el Análisis de Componentes Principales, ¿qué representan las cargas o saturaciones y las puntuaciones?

a) Las cargas son la posición de los registros y las puntuaciones la posición de las variables en el espacio de las componentes

b) Las cargas son la importancia de los factores en cada registro y las puntuaciones son la importancia de las variables en los factores

c) Las cargas son la posición de las variables y las puntuaciones la posición de los registros en el espacio de las componentes

d) Las cargas son la relación entre variables y registros en el espacio de factores

Pregunta 9

¿Qué solución de Clustering es mejor?

a) Solución con el mayor número de grupos

b) Solución con mayor variabilidad interna y menor silueta

c) Solución con menor silueta y variabilidad interna

d) Solución con mayor silueta y menor variabilidad interna

Pregunta 10

En el conjunto de datos de viviendas, la mediana de *lat* es:

a) 47.5745

b) 47.6789

c) 47.4752

d) 47.3746

Pregunta 11

En el conjunto de datos de Viviendas, el número de valores únicos en *condition* es:

a) 6

b) 5

c) 3

d) 4

Pregunta 12

En el conjunto de datos de Viviendas, la cantidad de instancias que pertenecen a la categoría 0 de la variable *waterfront* es:

a) 4852

b) 106

c) 4752

d) 4832

Pregunta 13

En el conjunto de datos de Viviendas, el porcentaje de filas faltantes no declaradas en la variable waterfront es:

a) 0

b) 2,34

c) 2,02

d) 2,12

Pregunta 14

En el conjunto de datos IPI (ventana de entrenamiento hasta el 31 de diciembre de 2017/ventana de prueba desde el 1 de enero de 2018), los resultados de la función eval_model() para un modelo aditivo de Hot-Winters son:

a) Valor p de LjungBox = 7.256e-06; MAPE = 3,14

b) Valor p de LjungBox = 7.256e-08; MAPE = 3,14

c) Valor p de LjungBox = 2,543e-05; MAPE = 5,26

d) Valor p de LjungBox = 1,584e-05; MAPE = 5,36

Pregunta 15

En el conjunto de datos IPI (ventana de entrenamiento hasta el 31 de diciembre de 2017/ventana de prueba desde el 1 de enero de 2018) los resultados de la función eval_model() para un SARIMAX(1,1,1)(1,1,1)12 son:

a) Valor p de LjungBox = 0,6974; MAPE = 0,68

b) Valor p de LjungBox = 0,0974; MAPE = 2,68

c) Valor p de LjungBox = 0,0094; MAPE = 4,68

d) Valor p de LjungBox = 0,0994; MAPE = 2,88

Pregunta 16

En el conjunto de datos de Viviendas, la cantidad de instancias con valor cero en *yr_renovated* es:

a) 4784

b) 4584

c) 4774

d) 4764

Pregunta 17

En el conjunto de datos de Viviendas (sin limpieza) el valor R^2 de un modelo con fórmula 'precio ~ lat' es:

a) 0,512

b) 0,613

c) 0,091

d) 0,453

Pregunta 18

En el conjunto de datos de Viviendas, la media de *sqft_living* es:

a) 2431.345

b) 2345.234

c) 1784.8984

d) 2077.382

Pregunta 19

En el conjunto de datos de Viviendas, la desviación típica de *price* es:

a) 371986.9

b) 372986.9

c) 373486.9

d) 373488.6

Pregunta 20

En el conjunto de datos de Viviendas, el número de valores únicos de *bedrooms* es:

a) 6

b) 3

c) 13

d) 33

Pregunta 21

En el conjunto de datos de Viviendas, el porcentaje de missings sin declarar en *sqft_lot* es:

a) 8.13

b) 7.64

c) 7.52

d) 7.54

Pregunta 22

En el conjunto de datos de viviendas (sin limpieza) el parámetro estimado para el predictor en el modelo con la fórmula *'price ~ sqft_living'* es:

a) 293.3356

b) 291.3188

c) 295.3188

d) 292.3278

Pregunta 23

En el conjunto de datos de viviendas (sin limpieza) el Odds Ratio estimado para el predictor en el modelo con la fórmula *'Luxury ~ sqft_living'* es:

a) 0.00196

b) 1.001866

c) 0.991286

d) 1.010886

Pregunta 24

En el conjunto de datos IPI (ventana de entrenamiento hasta 2017-12-31/ventana de prueba desde 2018-01-01), los resultados de eval_model() para un modelo SARIMAX(1,1,0)(1,1,0)12 son:

a) LjungBox pvalue = 5.153e-03; MAPE = 4.67

b) LjungBox pvalue = 4.613e-07; MAPE = 4.35

c) LjungBox pvalue = 4.153e-07; MAPE = 4.68

d) LjungBox pvalue = 4.653e-07; MAPE = 4.25

Pregunta 25

Si la comunalidad de una variable A en un ACP es 0.8:

a) El 80% de la variabilidad de A es capturada por las dimensiones del PCA

b) El 20% de la variabilidad de A es capturada por las dimensiones del PCA

c) El 80% de la variabilidad del sistema pertenece a A

d) El 80% de la variabilidad de PC1 es capturada por la variable A

Pregunta 26

Si las cargas de una variable A para los componentes (PC1,PC2) son (0.8,0.1):

a) A se asocia con PC2

b) A no está bien representada en el nuevo sistema

c) A se asocia con PC1

d) PC1 representa a A

Pregunta 27

En series temporales, un p-valor de la prueba de Ljung-Box de 0.8 representa que la serie residual es:

a) Falta de autocorrelación

b) Altamente auto correlacionada

c) Estacionaria

d) No estacionaria

Pregunta 28

El tamaño de una matriz de covarianza calculada a partir de un conjunto de datos que tiene N variables y M observaciones es:

a) Una matriz NxM

b) Una matriz NxN

c) Una matriz MxM

d) Una matriz MxN

Pregunta 29

Si el parámetro estimado para el predictor A en un modelo logístico para predecir una variable binaria y, es 0.4, entonces:

a) El odds ratio debe ser 1

b) El odds ratio debe ser de 0.4

c) El odds ratio debe ser de 1,39

d) El odds ratio debe ser de 1,49

Pregunta 30

En una muestra de datos, el valor de KMO es 0.8.

a) La adecuación de la muestra es buena y se podría reducir la dimensionalidad.

b) PCA tendrá 2 componentes

c) La adecuación de la muestra es mala y no se podría realizar la reducción de la dimensionalidad

d) El KMO es inaceptable

Pregunta Opcional 1

En el archivo Fuga de Clientes, imputa los valores perdidos de las variables numéricas por la mediana y los valores pedidos de las variables categóricas por la moda, obteniendo un nuevo archivo fuga_meadian_moda.

Con este archivo imputado, ajusta el modelo de fórmula 'Fuga ~ MetodoPago + Contrato + Antigüedad*FacturaMes' mediante la función logit de statsmodels.

¿Cuáles son los valores de PseudoR2 de McFadden y del logaritmo de la verosimilitud del modelo? ¿Cuántos parámetros estima el modelo? ¿Cuántos resultan significativosal 95% de nivel de confianza?

a) PseudoR = 0.2048; Log verosimilitud = -2463.6; 10 parámetros estimados; 10 parámetros significativos.

b) PseudoR = 0.2473; Log verosimilitud = -2763.1; 9 parámetros estimados; 9 parámetros significativos.

c) PseudoR = 0.2483; Log verosimilitud = -2763.1; 9 parámetros estimados; 7 parámetros significativos.

d) PseudoR = 0.2284; Log verosimilitud = -2709.5; 8 parámetros estimados; 6 parámetros significativos.

Pregunta Opcional 2

Con el mismo archivo imputado anteriormente, ajusta un proceso de selección automática de variables secuencial (SFS) con el método backward, y la configuración por parsimonia. La métrica ha de ser 'roc_auc', floating=F y 5 folders para la validación cruzada.

Nota: Tendrás que obtener la matriz explícita de diseño del modelo completo (sin el identificador de filas, claro!) y para ello puedes utilizar la función pasty.dmatrices.

Una vez finalizado el proceso, ¿Cuántos parámetros propone el método (directo de la salida)? ¿Cuál es el valor del área bajo la curva Roc resultante?

a) 11 parámetros; AUC = 0,858

b) 10 parámetros; AUC = 0,841

c) 10 parámetros; AUC = 0,847

d) 9 parámetros; AUC = 0,851

Pregunta Opcional 3

En el archivo *clientes_cluster.csv*, elimina los valores perdidos por lista. ¿Cuántos registros válidos quedan? Recodifica la variable Gender como una dummy (asegura que poner el drop_first=True) y elimina CustomerID. Escala el archivo. Ajusta un modelo de Kmeans con 4 grupos 25 inicializaciones de centroides y semilla 2025.

¿Cuál es el valor de la varianza interna de esta solución de clustering? ¿Y el valor de la silueta? ¿Qué valor tiene el centroide para el primer grupo y primera variable? ¿Y para la última variable en el último grupo?

a) Inercia = 267.702; Silueta = 0.361; Centroide 1,1 = 0.735; Centroide 4,4 = 1.131

b) Inercia = 287.654; Silueta = 0.311; Centroide 1,1 = -0.775; Centroide 4,4 = 0.113

c) Inercia = 288.527; Silueta = 0.401; Centroide 1,1 = -0.735; Centroide 4,4 = 1.013

d) Inercia = 287.502; Silueta = 0.301; Centroide 1,1 = -0.735; Centroide 4,4 = 1.113

Pregunta Opcional 4

Reduce las dimensiones del archivo anterior (escalado y sin perdidos) a 2 componentes principales. Obtén los resultados y dibuja el biplot. ¿Cuál es el valor de la carga de la variable 'Age' en la componente 1? ¿Qué variable tiene la mayor proyección sobre la componente 2?

a) Carga de Age en PC1 = 0.662767; Gender

b) Carga de Age en PC1 = -0.315729; Annual Income

c) Carga de Age en PC1 = -0.677432; Gender

d) Carga de Age en PC1 = -0.677432; Spending score