Imagen que contiene Texto

Descripción generada automáticamente

|  |  |
| --- | --- |
| Colegio Universitario **IES** *Siglo 21* | |
| IEFI | |
| **Materia:** Aprendizaje Automático 1 | **Profesor:** Ricardo Piña |
| **Modalidad:** PRESENCIAL | **Fecha:** |

Reservado para el alumno

|  |  |  |
| --- | --- | --- |
| **Alumno**: Nahuel Lahoz | **Carrera: IA** | **NOTA** |
| **DNI**: 40661225 |  |

**Escala de Valoración**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Puntaje | 1-24 | 25-39 | 40-54 | 55-61 | 62-66 | 67-72 | 73-79 | 80-87 | 88-95 | 96-100 |
| Nota | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Valoración | Logro  No satisfactorio | | | Logro básico | Logro satisfactorio | | Logro  Destacado | Logro  Sobresaliente | | Logro  Excelente |

**Pregunta 1:**

En un proceso donde utilizaremos los datos para efectuar pronósticos, al imputar missing values:

a) Es preferible imputar sólo a las columnas que tengan “missing values”.

b) Es preferible imputar a todas las columnas, tengan o no “missing values”.

Justifique su respuesta:

Mi respuesta es la b. Ya que puede suceder que al dividir en train y test, como la división es azarosa; al train no le toquen valores nulos, por ende, quedarían todos en el Test y entonces no podríamos pronosticar con estos. Lo mismo si queremos pronosticar valores nuevos. Es por esto que hay que imputar todas las columnas, tengan o no missing values.

**Pregunta 2:**

Al imputar missing values a una variable numérica, Ud observa que existen un par de valores “muy grandes” y preferiría que la imputación no se viera influída por ellos:

1) Imputaría la media

2) Imputaría la mediana

3) Imputaría la moda.

Justifique su respuesta:

He elegido la respuesta 2 porque la mediana toma el conjunto de datos, los ordena de menor a mayor y luego se para en el centro (medio) de todos los datos y toma ese valor. De esta forma, si hay algunos outliers, la imputación no se ve afectada. Si uno decide imputar por la media, esta se vería afectada por los valores más grandes, lo que nos refleja una disparidad en la distribución de los datos y se perderían algunas características

**Pregunta 3:**

Al recibir un DataSet para el cual deberá seleccionar un modelo con sus correspondientes hiperparámetros decide imputar los *missing values* de las variables numéricas con la media, entonces:

1) Calculará la media de cada columna numérica e imputará esos valor a los *missing values* del DataSet y a lugo utilizará estos datos para obenter Train / Test y Validation Train/ValidationTest.

2) Dividirá en DataSet en Train y Test, calculará la media de cada columna numérica del Train Set e imputará esos valores a los *missing values* del mismo Train, del Test, y utilizará este Train Set para obtener Valiadation Train y Validation Test.

3) Calculará la media de cada columna numérica e imputará esos valores a los *missing values* del DataSet; cuando tenga una nueva observación, la incluirá en el DataSet, recalculará la media de las columnas correspondientes e imputará a todo el DataSet.

4) Dividirá el DataSet en Train y Test, luego dividirá el Train en Validation Train y Validation Test. Calculará la media del Validation Train y la imputará al mismo Validation Train y al Valiadtion Test. Calculará la media del Train y la imputará al mismo Train y a Test set, finalmente calculará la media de todo el dataset y la imputará a todo el dataset y cualquier nueva observación que llegue para pronosticar.

JUSTIFICACION

De esta manera nos aseguramos de no aprender del Test Set ni de cualquier nueva observación.

## Pregunta 4:

Explique cómo funciona kNN para imputar missing values.

Para imputar los missing values knn funciona buscando las K observaciones más cercanas a ese missing value. El valor de k se atribuye en la cantidad de vecinos (por ejemplo k=2). Si estamos imputando una columna buscara los 2 valores mas cercanos a ese valor. Y luego promedia el valor (simple o ponderado) de estas observaciones y se lo imputa al missing value.

## Pregunta 5:

Qué son las dummy variables y cómo se generan?

Las variables dummy son aquellas que sirven para identificar categorías o clases a las que pertenecen las observaciones. Cuando una variable categórica tiene más de dos categorías, se puede representar mediante un conjunto de variables ficticias, con una variable para cada categoría.

Las variables dummy generadas toman valores de 0 y 1, donde los valores indican la presencia o ausencia de algo (por ejemplo, un 0 puede indicar sexo masculino y 1 puede indicar sexo femenino).

Estas se crean ya que a muchos de los algoritmos de sklearn debemos pasarle sólo números en las variables X.

## Pregunta 6:

A qué tipo de variables transformaría en Dummy Variables?

1)Numéricas  
2)Categóricas  
3)Categóricas Nominales

4) Categóricas Ordinales

5)Todas

Por categóricas entendemos aquellas que son ordinales y nominales.

## Pregunta 7:

Existen dos formas de crear las dummy variables que dan resultados ligeramente distintos. Explique las diferencias entre ellas.

La primer forma corresponde a Pandas “pd.get\_dummies”. Esta es una forma práctica y que nos ayuda mas que todo a describir el conjunto de datos; no tiene mucha utilidad a la hora de pronosticar, pero si nos produce facilidad a la comprensión visual.

En cambio, creando con Scikit-Learn, estamos mas aptos para pronosticar y es mucho mas útil ya que es adaptable al flujo de Machine Learning, el único problema que tiene es que nos desacomoda el orden de las variables en el DataFrame pero no es tan importante esto, ya que para pronosticar no es necesario que se mantenga cierto orden.

**Pregunta 8:**

Explique por qué puede ser útil normalizar las variables numéricas.

Generalmente los modelos de ML calculan parámetros asociados a sus variables, esto los hace muy sensibles a las escalas de las mismas. Debido a esto necesitamos normalizar las variables para que queden entre 0 y 1. Entonces cambiamos los valores de las columnas numéricas en el conjunto de datos para usar una escala común, sin distorsionar las diferencias en los rangos de valores ni perder información.  El modelo de Árbol no se ve afectado por este fenómeno por ejemplo.

**Pregunta 9:**

Es posible que al estandarizar una variable, se obtengan observaciones con valores mayores que 1? Explicar.

Si es posible ya que al estandarizar lo que se intenta conseguir es que:

1. La media sea 0
2. El desvío standard sea 1, 𝜎= 1

Pero no sus valores, ya que pueden estar más allá del intervalo [0,1]. Los valores que obtenemos nos darán un poco más grandes que 1 o un poco mas chicos que 0 pero no nos vamos a encontrar con valores 10 o 20 veces mas grandes, todos se van a encontrar relativamente cerca.

**Pregunta 10:**

Es posible que al normalizar una variable, se obtengan valores mayores que 1 en el mismo conjunto de datos? Explicar.

Al contrario de la estandarización no es posible obtener en el Train set valores mayores que 1 ya que el objetivo de la normalización es llevar dicho valores en el rango de [0, 1]. Sin embargo lo que si puede suceder es que aparezcan valores mayores a ese rango en el TEST Set.

**Pregunta 11:**

Cuál es la función de Costo J que se utiliza en Regresión Logística. Explique cómo se obtiene.

La función J es la siguiente:

La función J se obtiene desde la función sigmoide, que tiene su imagen acotada entre 0 y 1, esta posee forma de S.

A la función básica de la función sigmoide le podemos agregar el parámetro w0 que nos va a ayudar a desplazar dicha función de izquierda a derecha, luego podemos agregar el parámetro w1 que nos ayuda a hacer más o menos empinada la función. Estos dos parámetros son de gran ayuda para conseguir una aproximación a las observaciones que tenemos.

Luego se busca sumar todas las distancias que tenemos entre la sigmoide que calculamos y nuestras observaciones ubicadas en 0 y 1. Con esto buscamos obtener la función Costo Promedio, que busca las sumatorias de todas las distancias, elevadas al cuadrado para que den siempre positivo y que a su vez se divide por 1 la cantidad de observaciones para obtener dicho promedio.

Ahora, necesitamos una función que aplicada a h(x), nos sirva para decir si el resultado obtenido esta más cerca de 1 o de 0, por lo que elegimos la función logarítmica. Para evitar problemas con la función logarítmica, la multiplicamos por -1 para que cuando nuestra observación este cerca de 1, nos dé el resultado deseado. Por otro lado, cuando el valor es 0, el logaritmo tiende a -∞ por lo que debemos restarle 1 a h(x) para que de esta manera el resultado se quede en los parámetros deseados.

Por último, se deja de lado la idea de las distancias por secciones y se la reemplaza por la asignación de 0 y 1 para la variable a predecir y obtener el resultado más conveniente.