Guillermo Ortiz Macías
216787558
guillermo.ortiz7875@alumnos.udg.mx
Maestría en Ciencia de los Datos
Programación 2
Tarea Reforzamiento

1. El teorema del límite central

El teorema del límite central establece que la distribución del promedio de un gran número de muestras de una población tiende a comportarse como una distribución normal, esto sin importar si la distribución de la población es normal o no. En otras palabras, si se toma un número de muestras de una población y se calculan sus promedios, la distribución de los promedios tenderá a comportarse de forma normal, esto a medida de que el número de muestras aumenta.

Esto sucede independientemente de la distribución de probabilidad de los datos, en cualquier distribución, aún no siendo normal, el promedio de una muestra se va a comportar de forma normal.

Esto es de utilidad porque en muchas ocasiones no se tiene certeza sobre qué distribución de probabilidad siguen los datos que tenemos, pero sí se tiene certeza de que el promedio de estos datos sigue una curva normal. Esto nos permite obtener cosas como intervalos de confianza, pruebas t, ANOVA, o cualquier otra prueba estadística que dependa de una distribución normal.

2. Muestreo

El muestreo es la técnica de seleccionar un subconjunto de los datos que representan a la población. Su objetivo es tomar decisiones o hacer inferencia sobre toda la población basándose únicamente en la muestra, de forma que, al ser un set de datos más pequeño, se hace un proceso más rápido y con menor costo computacional.

Los muestreos se pueden hacer de muchas maneras, algunas de ellas son:

- **Aleatorio simple:** como su nombre lo indica, es un método aleatorio donde cada miembro de la población tiene exactamente la misma probabilidad de ser seleccionado. Es uno de los más utilizados, sin embargo puede no ser la mejor opción cuando la población no está balanceada: se tienen muchos miembros de un tipo y pocos de otro.
- **Muestreo estratificado:** este muestreo es útil cuando la población no está balanceada. Primero se divide la población en subgrupos, llamados estratos, para después hacer un muestreo en cada uno de los estratos. De esta forma cada uno de los subgrupos se representa de la misma forma en la muestra.
- **Muestreo sistemático:** Primero se hace una selección aleatoria y después se seleccionan a los siguientes miembros usando un patrón, por ejemplo seleccionar cada quinto miembro.
- **Muestreo con conveniencia:** No siempre se tiene acceso a toda la población, sino únicamente a algunos miembros. En este muestreo se seleccionan aquellos miembros que son de más fácil acceso. Este tipo de muestreo puede tener problemas de sesgos y no siempre representa correctamente a la población.

En machine learning el muestreo de los datos se utiliza mucho, su aplicación más común es el llamado "*Train-Test-Split"* que divide el set de datos en **Datos de entrenamiento** y **Datos de prueba.** Usualmente se realiza en proporción 70-30 ó 80-20, donde 70% de los datos se utilizan para entrenar el modelo, y el 30% restante para evaluarlo. Se debe tomar en cuenta que la muestra utilizada para el entrenamiento debe representar correctamente a la población, por lo que en algunos casos es necesario realizar un muestreo estratificado. Todo esto con el objetivo de prevenir un sobreajuste, reducir costos y mejorar la eficiencia del entrenamiento.

3. Diferencia entre error Tipo I y Error Tipo II

Estos errores surgen en modelos predictivos binarios donde el modelo predice si una entrada va a ser "Positiva" o "Negativa". Se dice que el modelo se equivoca cuando una entrada debió haber sido categorizada como "Positiva" pero su predicción fue "Negativa", o viceversa. De esta forma los errores se pueden clasificar como Tipo I o Tipo II:

- **Error Tipo I:** Se trata de un falso positivo. La entrada debió haber sido clasificada como "Negativa" pero se clasificó como "Positiva". Un ejemplo muy común es un modelo que dice si una persona tiene o no una enfermedad: habría un falso positivo cuando el modelo dice que la persona tiene la enfermedad cuando realmente no la tiene.
- **Error Tipo II:** Se trata de un falso negativo. La entrada debió haber sido clasificada como "Positiva" pero se clasificó como "Negativa". Usando el mismo ejemplo del modelo que diagnostica una enfermedad: el modelo dice que la persona no está enferma, pero realmente sí lo está.

Dependiendo del problema que se está resolviendo podría ser más peligroso o menos peligroso que un modelo tenga más errores de un tipo o del otro. Por ejemplo, un modelo que sea "estricto" puede estar mejor entrenado a predecir positivos de forma que se le tiene mucha confianza cuando da un resultado positivo, pero no tanta cuando da uno negativo.

En el ejemplo del modelo que diagnostica enfermedades se puede argumentar que es mejor tener muchos falsos positivos a tener muchos falsos negativos:

- Si a una persona le diagnostican una enfermedad pero realmente no la tiene solo le generaría molestias, pero
- Si a una persona no se la diagnosticaron pero sí la tiene, esto podría causar la muerte de la persona.

4. Regresión lineal y sus métricas.

La regresión lineal es un modelo de machine learning que predice el valor de una variable numérica contínua a partir de una o varias variables también numéricas. Para que el modelo funcione correctamente debe existir una relación lineal entre la variable dependiente (la que se está prediciendo) y las variables independientes (las variables de entrada).

Matemáticamente hablando una regresión lineal se ve así:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_n x_n + \varepsilon$$

Donde:

- y: variable dependiente que se quiere predecir.
- B0: Intercepción con el eje y
- Xn: Cada una de las variables independientes
- Bn: Cada uno de los coeficientes que indican cuánto cambia y con respecto a la respectiva variable independiente Xn.
- ε: Representa el error, la diferencia entre el valor predicho y el real.

Este modelo de machine learning se puede evaluar, algunas formas de evaluarlo son las siguientes:

- **P-value:** El p-valor es un indicativo que nos dice qué tan significativas son cada una de las variables independientes para predecir el valor de la dependiente. Cada uno de los coeficientes Beta del modelo indican cuánto va a cambiar la variable dependiente con respecto a la variable independiente que está actuando sobre este Beta. Si el Beta es 0, entonces la variable independiente no va a tener ningún impacto sobre la dependiente.

  El p-valor es una prueba de hipótesis en el que nos preguntamos qué tan diferente de cero debe ser un coeficiente β para que realmente represente un impacto significativo de la variable independiente sobre Y.

- **R-squared:** El R cuadrado puede tomar valores entre 0 y 1, donde 1 significa que el modelo predice bien los datos, y 0 significa que no lo hace. Esta métrica indica qué tanto de la variable dependiente es correctamente explicado por el modelo.

5. Estadística por iteración.

Son métodos estadísticos que utilizan múltiples iteraciones, o repeticiones, para estimar parámetros, realizar inferencias o ajustar modelos. Su utilidad principal sucede cuando nos enfrentamos a problemas donde las matemáticas o la estadística tradicional no son suficientes para resolverlos.

Uno de los ejemplos de estos métodos es el **bootstrapping** en el cual se calculan estadísticas de interés, como la media o la desviación estándar, mediante un proceso iterativo. El proceso es el siguiente:

- Primero se seleccionan “n” submuestras de la muestra original, donde n es el número de iteraciones.
- Para cada una de las submuestras se calcula la estadística de interés, como la media o la desviación estándar.
- Con cada uno de los resultados obtenidos de cada submuestra se construye un intervalo de confianza donde se encuentra el resultado real de la población.

La principal ventaja de estos modelos es que permiten encontrar soluciones a problemas muy complejos y difíciles de resolver con matemáticas tradicionales.

Y sus desventajas son que pueden ser más costosos computacionalmente, especialmente cuando requieren de muchas iteraciones; que la convergencia del algoritmo no está garantizada, no siempre se va a llegar a un resultado; y que tienen mucha dependencia de cómo se hace la inicialización del algoritmo.

6. Sesgo de selección.

El sesgo de selección se trata de un error sistemático donde la muestra seleccionada no representa correctamente a la población que se está estudiando. El error surge durante el muestreo de los datos al no realizarse una muestra correcta que represente a toda la población.

En machine learning este error puede significar que si un modelo se entrenó con datos sesgados solamente va a ser bueno para predecir o categorizar otras observaciones que compartan el mismo sesgo.

Para prevenir este error se tiene que realizar un buen muestreo de la población. La muestra debe incluir observaciones que representen a cada uno de los estratos distintos que existan en la población.

Una vez hecho el muestreo se puede hacer una revisión para identificar posibles sesgos y asegurarse que la muestra sea representativa de toda la población.

7. Probabilidad binomial.

La probabilidad binomial ayuda a determinar la probabilidad de que un número de éxitos ocurra en un número de ensayos. Por ejemplo: ¿cuál es la probabilidad de tener exactamente 3 zurdos en un grupo de 19 personas?

La fórmula matemática para calcular esto es:

$$P(X = x) = {}_nC_x \, p^x \, q^{n-x}$$

Donde:

- x, es el número de éxitos buscado.
- n, es el total de ensayos.
- p, es la probabilidad de que suceda el éxito.
- q, es la probabilidad de que no suceda el éxito.

Si la probabilidad de tener un zurdo es 0.13 entonces:

$$P(X = 3) = {}_{19}C_3 \, p^3 \, q^{19-3}$$

Al hacer este cálculo se tiene que la probabilidad de tener 3 zurdos en un grupo de 19 personas es del 23%

La distribución binomial calcula la probabilidad de éxito de un evento aleatorio con únicamente dos resultados posibles.

En machine learning la distribución binomial puede utilizarse por ejemplo para medir la precisión de un modelo binario. Se puede analizar el número de éxitos del modelo y compararlo con el número de predicciones realizadas. De esta forma se podría determinar la probabilidad de que el modelo tenga “n” cantidades de éxitos para un determinado número de observaciones.