Ayudantía Nro #7 : Quasars Data

Funciones Importantes

Para poder trabajar con datos, es importante conocer algunas de las siguientes funciones:

Directory[] : Devuelve el directorio de trabajo actual del sistema, es decir, la ubicación en la que el sistema está buscando y guardando archivos por defecto. Esta función es útil para conocer la ubicación actual en la que Mathematica busca archivos y para especificar rutas de acceso en operaciones de lectura y escritura de archivos.

NotebookDirectory[]: Esta función nos permite obtener la ubicación en la que se encuentra almacenado el *notebook* actual. Es útil para acceder a archivos y recursos que están en la misma ubicación que el cuaderno, como imágenes, datos u otros cuadernos relacionados.

SetDirectory[]: Se para establecer el directorio de trabajo actual en el sistema. Permite cambiar el directorio en el que Mathematica buscará y guardará archivos, lo que es útil para organizar y acceder a diferentes ubicaciones de archivos en tu sistema. Puedes especificar una ruta de acceso como argumento para esta función, y a partir de ese punto, Mathematica considerará ese directorio como el nuevo directorio de trabajo actual.

FileNames[]: Se utiliza para obtener una lista de nombres de archivos que coinciden con un patrón específico en un directorio dado. Permite al usuario buscar y listar archivos en un directorio en función de criterios como extensiones de archivo, nombres parciales, o cualquier otro patrón que se desee aplicar.

SmoothKernelDistribution[]: La distribución de kernel suave estima la forma de la distribución de los datos alrededor de cada punto, creando una curva que se adapta a esos puntos, lo que proporciona una representación más suave y continua de la distribución de probabilidad. Podemos personalizar el ancho del kernel y otros parámetros para adaptar la suavización a nuestras necesidades específicas. Pero..

¿Qué es una Distribución de Kernel?

En el contexto de estadísticas y análisis de datos, es un método no paramétrico que permite estimar la función de densidad de probabilidad de una variable aleatoria a partir de un número finito de observaciones. En otras palabras, es una técnica utilizada para suavizar y estimar la forma de una distribución de datos a partir de un conjunto de puntos de datos. En lugar de representar los datos como una serie de puntos discretos, ésta crea una curva suave que modela la distribución de probabilidad subyacente

de los datos. Esta curva suave se conoce como “kernel” y se coloca en cada punto de datos, reflejando la probabilidad de que un valor caiga en esa ubicación.

Esta técnica es útil para visualizar y analizar datos, ya que proporciona una representación más suave y continua de la distribución subyacente, lo que puede hacer que los patrones y tendencias sean más visibles. También se utiliza en técnicas como la estimación de densidad de kernel (KDE) para estimar la densidad de probabilidad de los datos en un punto específico.

FindDistribution[]: Se utiliza para identificar automáticamente la distribución de probabilidad que mejor se ajusta a un conjunto de datos dado. Esta función examina los datos y busca la distribución que mejor se ajusta en términos de estadísticas y probabilidad, como la distribución normal, uniforme, exponencial, etc. Al hacerlo, ayuda a caracterizar y modelar la naturaleza de los datos sin necesidad de suposiciones previas sobre la distribución.

MixtureDistribution[]: Se utiliza para crear una distribución de probabilidad que combina varias distribuciones individuales. Permite modelar situaciones en las que los datos provienen de diferentes fuentes o subpoblaciones con diferentes distribuciones. La distribución resultante es una mezcla ponderada de estas distribuciones individuales, donde los pesos indican la contribución relativa de cada subpoblación.

LogLikelihood[]: Se utiliza para calcular el logaritmo de la función de verosimilitud de un modelo estadístico en función de los datos observadors y los parámetros del modelo en el contexto de estadísticas y análisis de datos, lo que es útil en el proceso de estimación de parámetros y maximización de la función de verosimilitud en estadísticas. Ahora, recordemos...

¿Qué es la Función de Verosimilitud?

En el contexto de estadística, es una medida de cuán bien un modelo estadístico representa los datos observados. Representa la probabilidad de que los datos observados sean generados por el modelo en función de los parámetros del modelo, por lo que cuantifica cuán “verosímil” o probable es que los datos provengan de un modelo específico.

La idea principal es encontrar los valores de los parámetros del modelo que maximizan la función de verosimilitud. Estos valores óptimos se conocen como estimaciones de máxima verosimilitud (MLE, por sus siglas en inglés) y representan la configuración de parámetros que hace que los datos observados sean más probables bajo el modelo.

FindMinimum[]: Esta función se utiliza para encontrar el valor mínimo de una función dada. Esto implica encontrar el punto en el espacio de parámetros en el que la función tiene su valor más bajo.

NMaximize[]: Se utiliza para encontrar el valor máximo de una función dada, lo que implica encontrar el punto en el espacio de parámetros en el que la función alcanza su valor más alto.

Chop[]: Esta función se utiliza para eliminar pequeños valores cercanos a cero en una expresión numérica. Cuando se trabaja con números *floats*, pueden surgir pequeños errores numéricos que generan valores muy cercanos a cero pero no exactamente iguales.Esta función identifica estos valores insignificantes y los reemplaza por cero, lo que simplifica y limpia los resultados numéricos.

Drop[]: Esta función elimina elementos específicos de una lista o matriz. Permite al usuario descartar elementos en una posición determinada o un rango de posiciones dentro de la lista, lo que resulta en una lista más corta o una matriz más pequeña.

Position[]: Se utiliza para encontrar las posiciones en las que un elemento o un patrón específico aparece en una lista, matriz u otra estructura de datos.

GraphicsGrid[]: Esta función organiza y muestra múltiples gráficos en una cuadrícula bidimensional, permitiendo crear una disposición de gráficos en filas y columnas, lo que facilita la comparación y visualización de múltiples gráficos en una sola vista.

Operadores Útiles:

% : Este símbolo en Wolfram Mathematica se utiliza como una variable especial que hace referencia al resultado del cálculo inmediatamente anterior, por lo que es una forma de acceder al resultado previo en una secuencia de comandos sin necesidad de asignarle un nombre o una variable específica. Por ejemplo, si realizas una operación como: `2 + 3` y luego escribimos `%*2`, el output será 10.

/. : En Wolfram Mathematica este símbolo se utiliza para realizar sustituciones en expresiones matemáti cas o fórmulas. La estructura básica es `expr /. regla` dond expr es la expresión original y `regla` es la regla de sustitución que especifica qué parte de la expresión original debe ser reemplazada por qué valor. Por ejemplo, `x^2 + 3x + 2 /. x -> 4` la expresión x^2 + 3x + 2 se evaluará sustituyendo “x” por “4”, lo que resultará en el valor 22.

Wolfram Alpha: Wolfram Alpha es un motor de conocimiento computacional desarrollado por Wolfram Research. Es una poderosa herramienta en línea que permite a los usuarios realizar cálculos matemáticos, buscar información y obtener respuestas a preguntas de manera interactiva. Wolfram Alpha utiliza un amplio conjunto de datos y algoritmos para ofrecer información detallada sobre una amplia gama de temas, desde matemáticas y ciencia hasta datos históricos y más.Para acceder a ella sólo tenemos que apretar `shift+=` dos veces y escribir lo que deseamos buscar.

Algunas definiciones...

Modelo de Distribución Bimodal

El modelo de distribución de masas bimodal es un concepto estadístico que se refiere a una distribución de datos que presenta dos modos (valores o puntos de máxima frecuencia) claramente distinguibles, en otras palabras, los datos muestran dos peaks o agrupaciones distintas en lugar de una sola.

Este fenómeno puede ocurrir cuando una población se compone de dos subpoblaciones o grupos que tienen características diferentes y, por lo tanto, generan dos conjuntos de datos con modas separadas. Por ejemplo, si se miden las alturas de hombres y mujeres en una muestra conjunta, es probable que se observe una distribución bimodal, ya que las alturas de hombres y mujeres suelen tener picos de frecuencia diferentes.

Modelo de Distribución Trimodal

Este modelo es similar al bimodal, pero en lugar de tener dos modos claramente distinguibles, presenta tres modos o picos en la distribución de datos. En otras palabras, los datos muestran tres agrupaciones distintas de valores con frecuencias máximas.

Ordinary Least Squares (OLS)

El OLS (Ordinary Least Squares) es un método estadístico utilizado en regresión lineal para estimar los parámetros de un modelo de regresión. Su objetivo es encontrar la *línea recta* que mejor se ajusta a un conjunto de datos mediante la minimización de la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos por el modelo.

En otras palabras, este método busca la línea recta que representa la relación lineal entre una variable independiente y una dependiente de manera que la suma de los cuadrados de las diferencias entre los valores reales y los valores predichos sea la más pequeña posible. Esta línea se define por dos parámetros: la pendiente (coeficiente) y la ordenada al origen.

Este método ayuda a identificar la influencia de las variables predictoras en la variable objetivo y a realizar predicciones basadas en el modelo ajustado. El método es relativamente sencillo de entender y aplicar, lo que lo hace una herramienta fundamental en análisis de datos y modelado estadístico.

Maximum Likelihood Estimator(MLE)

El MLE (Maximum Likelihood Estimator) es otro método estadístico utilizado para estimar los parámetros de un modelo estadístico de manera que maximice la verosimilitud de los datos observados. En

otras palabras, busca encontrar los valores de los parámetros que hacen que los datos observados sean los más probables bajo el modelo propuesto.

En términos simples, el MLE busca determinar los valores de los parámetros del modelo que hacen que los datos observados sean los más probables de haber ocurrido. Esto se logra maximizando la función de verosimilitud, que mide cuán probable es que los datos observados se hayan generado bajo el modelo.

El MLE es ampliamente utilizado en estadísticas y econometría para ajustar modelos a datos y estimar los parámetros desconocidos. Es una técnica fundamental en la estimación de parámetros y se aplica en una variedad de contextos, desde la regresión lineal hasta el ajuste de distribuciones de probabilidad. El resultado del MLE es el conjunto de valores de los parámetros que mejor se ajusta a los datos disponibles según el modelo estadístico seleccionado.

¿OLS v/s MLE?

Como sabemos, OLS (Ordinary Least Squares) y MLE (Maximum Likelihood Estimator) son dos enfoques comunes para estimar parámetros en modelos estadísticos, pero difieren en la forma en que realizan las estimaciones y en las suposiciones subyacentes, pero, ¿cuál es mejor? En general, OLS es un método específico utilizado en el contexto de regresión lineal y se basa en la minimización de la suma de cuadrados de los residuos. MLE, por otro lado, es un enfoque más general que se utiliza en una variedad de modelos estadísticos y se basa en la maximización de la verosimilitud de los datos. La elección entre OLS y MLE depende del tipo de modelo y de las suposiciones adecuadas para los datos en cuestión.

Características	OLS	MLE
Método de Estimación	Busca minimizar la suma de los cuadrados de las diferencias entre los valores observados y los valores estimados por el modelo lineal, es decir, encuentra la línea de mejor ajuste a los datos mediante una minimización de la suma de residuos al cuadrado.	Busca encontrar los valores de los parámetros del modelo que maximizan la verosimilitud de los datos observados. Se basa en supuestos sobre la distribución de probabilidad subyacente de los datos y maximiza la probabilidad de que los datos observados provengan de esa distribución.
Suposiciones	Asume que los errores (residuos) en el modelo de regresión lineal siguen una distribución normal con media cero y varianza constante. Además, no requiere ninguna suposición específica sobre la distribución de los datos en sí.	Implica supuestos específicos sobre la distribución de probabilidad de los datos. La elección de la distribución de probabilidad depende del tipo de modelo estadístico que se está estimando. Puede implicar suponer que los datos siguen una distribución normal, exponencial, binomial, etc.
Flexibilidad	Más adecuado para modelos lineales, como la regresión lineal.	Es un enfoque más general que se puede aplicar a una amplia variedad de modelos estadísticos, incluidos los no lineales.

Ejercitación & Aplicación

Ejercicio 1: Acostumbrándonos a la manipulación de directorios y archivos.

(a) Utiliza la función Directory para mostrar el directorio de trabajo actual.

(b) Utiliza la función NotebookDirectory para obtener el directorio del cuaderno actual.

(c) Cambia el directorio de trabajo a un directorio específico utilizando SetDirectory.

(d) Utiliza FileNames para obtener una lista de nombres de archivos en el directorio actual.

Ejercicio 2: Análisis y Optimización de Datos

(a) Genera una lista de 1000 ventas diarias aleatorias, utilizando una distribución normal con media ($\mu = 100$) y desviación estándar ($\sigma = 20$). Almacena estos datos en una lista llamada “ventasDiarias”.

(b) Crea un histograma de las ventas diarias. Personaliza el gráfico con etiquetas de ejes y un título.

(c) Crea una distribución de kernel suave a partir de “ventasDiarias” y visualiza la distribución estimada.

(d)Utiliza FindDistribution para intentar identificar las mejores tres distribuciones de probabilidad que mejor se ajusten a las ventas diarias. Almacena las distribuciónes identificada en una variable llamada “mejorDistribucion”. Graficalas y comparalas en un mismo gráfico.

(e) Calcula la log-verosimilitud de los datos de ventas diarias con respecto a la distribución identificada en el punto d.

(f) Halla los parámetros de la distribución que maximizan la verosimilitud y el valor máximo de la función de densidad de probabilidad de la distribución.

(g) Elimina los valores muy pequeños en la función de densidad de probabilidad, si es necesario. Elimina los elementos atípicos en los datos de ventas diarias.