ANALISIS DE DATOS DE ENTRADA

AUTOR: JAVIER CHINCHILLA MORALES

NOVIEMBRE: 2020

San Marcos

Introducción

El análisis de datos ha ido adquiriendo cada vez más importancia dentro de las distintas áreas científicas y muy especialmente dentro de las denominadas Cinecias Sociales y de la Salud.

Contenido

Análisis de datos de entrada

Selección de familia de distribuciones

Existen distintos tipos de distribución de datos, para los cuales es importante conocer cuales son y como darles utilidad, a continuación, una serie de consejos para determinar qué tipo de distribución de datos siguen nuestros datos

1. Conoce los diferentes tipos de distribución de datos: uniforme discreta, Bernoulli, binomio, binomio negativo, Poisson, geométrica, uniforme continua, normal (curva de campana), exponencial, gamma y beta.
2. Realiza una representación gráfica de tus datos.
3. Descarta primero lo que no puede ser.
4. Si hay algún pico en el conjunto de datos, no puede ser una distribución uniforme discreta.
5. Si los datos tienen más de un pico, no es Poisson o binomio.
6. Si tiene una sola curva, no hay picos secundarios, y tiene una pequeña pendiente en cada lado, podría ser una distribución Poisson o gamma. Pero no podrá ser una distribución uniforme discreta.
7. Si los datos se distribuyen de manera uniforme, y es sin inclinar hacia un lado, es seguro excluir una distribución gamma o Weibull.
8. Si la función tiene una distribución uniforme o un pico en el medio de los resultados graficados, no es una distribución geométrica o una distribución exponencial.
9. Después de que el tipo de distribución de probabilidad se ha reducido, haz un análisis de R cuadrado de cada posible tipo de distribución de probabilidad. El que tenga el mayor valor R cuadrado es probablemente el correcto.
10. Elimina un dato atípico. A continuación, vuelve a calcular R cuadrado. Si el mismo tipo de distribución de probabilidad aparece como la coincidencia más cercana, luego hay un alto grado de confianza de que se trate de la distribución de probabilidad correcta para utilizar en el conjunto de datos.

Test de ajuste de curvas

En cuanto al tema de ajuste de curvas en métodos numéricos tenemos los 2 métodos más utilizados que son la regresión y la interpolación:

Regresión: también conocida como correlación líneal es un método estadístico que estudia la relación lineal existente entre dos variables, para ello es necesario disponer de parámetros que permitan cuantificar dicha relación, La correlación lineal entre dos variables, además del valor del coeficiente de correlación y de sus significancias, también tiene un tamaño de efecto asociado. Se conoce como *coeficiente de determinación* (para mas detalle diríjase al siguiente link)

Interpolación: En ciertos casos el usuario conoce el valor de una funci´on f(x) en una serie de puntos x1, x2, · · ·, xN, pero no se conoce una expresi´on anal´ıtica de f(x) que permita calcular el valor de la funci´on para un punto arbitrario. Un ejemplo claro son las mediciones de laboratorio, donde se mide cada minuto un valor, pero se requiere el valor en otro punto que no ha sido medido. Otro ejemplo son mediciones de

temperatura en la superficie de la Tierra, que se realizan en equipos o estaciones meteorol´ogicas y se necesita calcular la temperatura en un punto cercano, pero distinto al punto de medida. La idea de la interpolaci´on es poder estimar f(x) para un x arbitrario, a partir de la construcci´on de una curva o superficie que une los puntos donde se han realizado las mediciones y cuyo valor si se conoce. Se asume que el punto arbitrario x se encuentra dentro de los l´ımites de los puntos de medici´on, en caso contrario se llamar´ıa extrapolaci´on. En este texto se dicute exclusivamente la interpolaci´on, aunque la idea es similar (para mas detalle diríjase al siguiente link)

Ajuste de curvas en ausencia de datos

A lo largo de la profesión de un ingeniero, un físico, un matemático, frecuentemente se presentan ocasiones en las que deben ajustar curvas a un conjunto de datos representados por puntos. Las técnicas desarrolladas para este fin pueden dividirse en dos categorías generales: interpolación y regresión. Se considerará aquí la primera de estas dos categorías. Más aún, como la teoría de aproximación polinomial es más adecuada para un primer curso de cálculo numérico, será la que se considere principalmente en este trabajo.

Cuando se asocia un error sustancial a los datos, la interpolación polinomial es inapropiada y puede llevar a resultados no satisfactorios cuando se usa para predecir valores intermedios. Los datos experimentales a menudo son de ese tipo. Una estrategia más apropiada en estos casos es la de obtener una función aproximada que ajuste “adecuadamente” el comportamiento o la tendencia general de los datos, sin coincidir necesariamente con cada punto en particular. Una línea recta puede usarse en la caracterización de la tendencia de los datos sin pasar sobre ningún punto en particular. Una manera de determinar la línea, es inspeccionar de manera visual los datos graficados y luego trazar la “mejor” línea a través de los puntos. Aunque este enfoque recurre al sentido común y es válido para cálculos a “simple vista” es deficiente ya que es arbitrario. Es decir, a menos que los puntos definen una línea recta perfecta (en cuyo caso la interpolación sería apropiada), cada analista trazará rectas diferentes. La manera de quitar esta subjetividad es considerar un criterio que cuantifique la suficiencia del ajuste. Una forma de hacerlo es obtener una curva que minimice la diferencia entre los datos y la curva y el método para llevar a cabo este objetivo es al que se le llama regresión con mínimos cuadrados.
Ajuste de curvas se usa para encontrar una función que responda a una muestra de datos obtenidas de alguna medición, sampleo etc.
La aplicación más elemental es para dibujar una curva en una computadora en base a algunos puntos (datos) de manera que se vea bien.
Otra aplicación más interesante es la obtener una función que en base a algunos puntos obtenidos de medición se pueda estimar otros puntos que no fueron medidos empíricamente.

Para lograr este objetivo se utilizan, entre otros, interpolación y aproximación por el método de mínimos cuadrados; en los métodos por interpolación la función pasa exactamente por los puntos observados, en cambio en el método de aproximación se busca que una función pase lo más cercanamente posible por los puntos observados.

Generación de variables aleatorias.

Buscamos métodos que nos permitan obtener valores de variables aleatorias que sigan determinadas distribuciones de probabilidad a partir de los números aleatorios generados, que siguen la distribución

Uniforme en el intervalo (0,1). Hay cuatro métodos generales de generación de variables aleatorias y una serie de métodos particulares de las distintas distribuciones. La facilidad de aplicación de dichos métodos, así como el coste computacional asociado a los mismos, varía mucho según la familia de variables aleatorias a las que se apliquen. Normalmente existen varios algoritmos que se pueden utilizar para generar valores de una determinada distribución, y diferentes factores que se pueden considerar para determinar qué algoritmo utilizar en un caso particular. Desafortunadamente dichos factores suelen entrar en conflicto unos con otros y a veces se ha de llegar a una solución de compromiso. Algunos de estos factores son los siguientes: Exactitud: se han de obtener valores de una variable con una precisión dada. A veces se tiene suficiente con obtener una aproximación y otras no. Eficiencia: el algoritmo que implementa el método de generación tiene asociado un tiempo de ejecución y un gasto de memoria. Elegiremos un método que sea eficiente en cuando al tiempo y a la cantidad de memoria requeridos. Complejidad: Buscamos métodos que tengan complejidad mínima, siempre y cuando se garantice cierta exactitud.

Conclusiones y recomendaciones

Una vez analizado el tema del análisis de datos de entrada en métodos numéricos es de importancias saber que se debe escoger un tipo de análisis de datos que se amolde a unestro conjunto de datos y luego conocer herramientas matemáticas que nos ayudan a analizar ajustes de curvas, ya sea con datos o en ausencia de, tambien como generar variables aleatorias.

Referencias bibliográficas

- Anderson, D; Sweeney, D. & Williams, T. (2019). *Fundamentos de métodos cuantitativos para los negocios*. Cengage Learning

San Marcos

www.usanmarcos.ac.cr

San José, Costa Rica