ANÁLISIS MULTIVARIANTE

Introducción al análisis multivariado,.

Mg. Víctor Guevara P

Universidad Nacional José Faustino Sánchez Carrión

Mayo 2023

Normas de participación.

- Ingresar puntual
- Realizar las actividades encomendadas.

Presentación

Docente

- **Mg. Víctor Guevara**
  - Ingeniero Estadístico
  - Docente Investigador
  - Egresado Msc. ciencia de los datos URP
  - Egresado Msc. Estadística aplicada UNALM

Sílabo

Introducción

Introducción

En el análisis de datos, observamos diferentes variables (o factores) y cómo pueden afectar ciertas situaciones o resultados.

Por ejemplo:

- En marketing, puede observar cómo la variable "dinero gastado en publicidad" afecta la variable "número de ventas".
- En el sector de la salud, es posible que desee explorar si existe una correlación entre las "horas semanales de ejercicio" y el "nivel de colesterol".

Esto nos ayuda a comprender por qué ocurren ciertos resultados, lo que a su vez nos permite hacer predicciones y decisiones informadas para el futuro.

Análisis de datos

El análisis de datos basado en los tipos de variables en consideración se divide ampliamente en tres categorías:

Análisis univariado:

El más simple de todos los modelos de análisis de datos, el análisis univariado considera solo una variable en el cálculo. Por lo tanto, aunque su aplicación es bastante simple, tiene un uso limitado en el análisis de grandes datos. Por ejemplo, la incidencia de una enfermedad.

Análisis bivariado:

Como su nombre indica, el análisis bivariado tiene en cuenta dos variables. Tiene un área de aplicación ligeramente ampliada pero, sin embargo, está limitada cuando se trata de grandes conjuntos de datos. Por ejemplo, la incidencia de una enfermedad y la estación del año.

Análisis de datos

Análisis multivariante:

El análisis multivariante tiene en cuenta una gran cantidad de variables. Esto lo convierte en una herramienta complicada a la par que esencial. La mayor virtud de este modelo es que considera tantos factores como sea posible. Esto da como resultado una tremenda reducción del sesgo y da un resultado más cercano a la realidad.

Tendencias convergentes

La información disponible para la toma de decisiones se ha disparado en los últimos años y seguirá haciéndolo en el futuro, probablemente incluso más rápido.

Hasta hace poco, gran parte de esa información simplemente desaparecía. No se recogió o se desechó.

Hoy en día, esta información se recopila y almacena en almacenes de datos, y está disponible para ser "extraída" para mejorar la toma de decisiones. Parte de esa información se puede analizar y comprender con estadísticas simples, pero gran parte requiere técnicas estadísticas.

Tendencias convergentes

1. Surgimientos de Big Data

- No hay ningún factor que afecte la analítica que haya sido más publicitado y comentado que el "Big Data", ya que ha habido una explosión en los datos disponibles en la actualidad.

Tendencias convergentes

- Las fuentes son variadas: el mundo de las redes sociales y el comportamiento en línea; el Internet de las cosas, que ha traído conectividad a casi todos los tipos de dispositivos; la cantidad casi incomprensible de datos en las ciencias en áreas como la genómica, la neurociencia y la astrofísica; la capacidad de los dispositivos de almacenamiento para capturar toda esta información y software (por ejemplo, Hadoop y otros) para administrar esos datos; y finalmente el reconocimiento por parte de organizaciones de todo tipo de que conocer más a sus clientes a través de la información puede informar mejor la toma de decisiones.

Tendencias convergentes

¿Qué es Big Data?

Figure 1: Big Data:

Tendencias convergentes

2.Modelos estadísticos vs Modelos minería de datos y/o machine learning * La era de Big Data no solo ha proporcionado fuentes de datos nuevas y variadas, sino que también ha impuesto nuevas exigencias a las técnicas analíticas necesarias para tratar con estas fuentes de datos. El resultado ha sido el reconocimiento de dos "culturas" de análisis de datos que son distintas y tienen un propósito a su manera. Breiman (2001) definió estas dos culturas como modelos de datos versus modelos algorítmicos.

- Existen diferencias significativas entre estos dos enfoques, ambos operan en las mismas condiciones.
- objetivos: predicción precisa del resultado y explicación/conocimiento de cómo opera el proceso.

Tendencias convergentes

Modelos estadísticos o de datos

- El concepto de modelos de datos se alinea estrechamente con nuestra visión clásica de modelos y análisis estadísticos. Aquí, el analista suele definir algún tipo de modelo de datos estocásticos (por ejemplo, un modelo de regresión logística o múltiple), como las variables predictoras y su forma funcional.

Tendencias convergentes

Modelos estadísticos o de datos * Por lo tanto, el modelo de datos es un modelo especificado por el investigador que luego se estima utilizando los datos disponibles para evaluar el ajuste del modelo y, en última instancia, su aceptabilidad. Los desafíos para el investigador son:

- a. Especificar correctamente una forma de modelo que represente el proceso que se examina.
- b. Realizar el análisis correctamente para proporcionar la explicación y la predicción deseadas.

Por lo tanto, el investigador debe hacer algunas suposiciones sobre la naturaleza del proceso y especificar una forma de modelo que reproduzca mejor sus operaciones.

Tendencias convergentes

Modelos minería de datos y/o machine learning * Los modelos algorítmicos, también conocidos como minería de datos e incluso los términos contemporáneos de aprendizaje automático e inteligencia artificial, adoptan un enfoque diferente para comprender el proceso al cambiar el enfoque de la explicación del proceso a la predicción.

- La premisa fundamental es que el proceso que se estudia es inherentemente tan complejo que la especificación de un modelo preciso es imposible. Más bien, el énfasis está en los algoritmos, cómo pueden representar cualquier proceso complejo y qué tan bien predicen los resultados en última instancia.

Tendencias convergentes

- La explicación del proceso es secundaria a la capacidad del algoritmo para completar su tarea de predicción. Por lo tanto, los modelos de reconocimiento de imágenes no brindan información sobre cómo se pueden distinguir las imágenes, sino qué tan bien se diferencian entre las imágenes.

Tendencias convergentes

3 Inferencia causal

La inferencia causal es el movimiento más allá de la inferencia estadística a la declaración más fuerte de "causa y efecto" en situaciones no experimentales. Si bien las declaraciones causales se concibieron principalmente como el dominio de los experimentos controlados aleatorios, los desarrollos recientes han brindado a los investigadores

- a. Los marcos teóricos para comprender los requisitos para las inferencias causales en entornos no experimentales

Tendencias convergentes

3 Inferencia causal

- b. Algunas técnicas aplicables a datos no experimentales. reunidos en un entorno experimental que todavía permiten extraer algunas inferencias causales.

El paso de la inferencia estadística al análisis causal no es una técnica única, sino un cambio de paradigma que incorpora un énfasis en los supuestos que son la base de todas las inferencias causales y un marco para formular y luego especificar estos supuestos

Análisis multivariante

¿Qué es el análisis multivariante?

- Las estadísticas multivariadas se refieren a métodos que examinan el efecto simultáneo de múltiples variables.
- Los métodos multivariados son técnicas que permiten realizar el análisis estadístico de datos, cuando se han registrado muchas características sobre un conjunto de objetos o individuos.

Objetivos

- Analizar y simplificar la estructura de datos
- Clasificación y conglomeración
- Análisis de dependencia
- Inferencia estadística

Clasificación del análisis multivariante

- La clasificación tradicional de los métodos estadísticos multivariados sugerida por Kendall se basa en el concepto de dependencia entre variables (Kendall 1957). Si un interés se centra en la asociación entre dos conjuntos de variables, donde un conjunto es la realización de una variable dependiente (o variables) y el otro conjunto es la realización de un número de variables independientes, entonces la clase apropiada de técnicas serían aquellas designados como métodos multivariados de dependencia.
- Si el interés se centra en la asociación mutua entre todas las variables sin distinción entre tipos de variables, se utilizan métodos multivariados de interdependencia (Dillon 1984).

Métodos de Interdependencia

No hay distinción entre las variables. Son métodos descriptivos que sintetizan la información, mostrar la estructura de los datos o clasificar las variables.

Método	Métricas	No métricas
Análisis de componentes principales	X	
Análisis factorial	X	
Análisis de correspondencia		X
Análisis de cluster	X	
Análisis de escalamiento multidimensional	X	X

Figure 2: Métodos de Interdependencia

Métodos de dependencia

Se distinguen variables dependientes e independientes. Son métodos con finalidades explicativas.

	Var. Dependiente		Var. Independiente	
Método	Met.	No Met.	Mét.	No Mét.
Análisis discriminante		X	X	
Análisis de regresión multivariado	X		X	
Análisis de regresión logística		X	X	X
Análisis de variancia	X			X

Figure 3: Métodos de Interdependencia

Métodos multivariados

Métodos multivariados

Análisis de componentes principales (ACP).

Su propósito es el de reducir la dimensionalidad de las variables originales, tratando de explicar la mayor parte de la variabilidad total del conjunto de variables originales con el menor número posible de componentes principales, también es usada como un análisis descriptivo de los datos.

Métodos multivariados

Análisis Factorial (AF).

Permite sintetizar el fenómeno en estudio a través de analizar la estructura de correlaciones entre el conjunto de variables, se resume la información e identifica una estructura subyacente del conjunto de los datos.

Métodos multivariados

Análisis de Correspondencia (AC).

Es similar al AF, en el sentido que trata de descubrir y describir las dimensiones fundamentales de un fenómeno pero con la particularidad que las variables son categóricas que proporcionan mapas preceptúales que facilitan la interpretación y su análisis.

Métodos multivariados

Análisis de cluster .

A partir de un conjunto de variables métricas trata de agrupar el conjunto de observaciones (objetos o individuos) aplicando medidas de distancia, formando grupos que sean lo más homogéneos los individuos dentro de ellos y heterogéneos entre los grupos

Métodos multivariados

Análisis de Escalamiento Multidimensional .

Es similar al análisis de cluster, con la diferencia que trabaja con variables de escalas métricas y/o ordinales (cualitativas) , con la finalidad de formar grupos.

Métodos multivariados

Análisis Discriminante.

La variable dependiente es categórica (dos o más categorías) y las independientes son variables métricas. El método trata de construir una regla o función de clasificación en base de las variables independientes y conociendo a priori los grupos indicados por la variable categórica; y así poder asignar nuevas observaciones a uno de los grupos aplicando la regla o función de clasificación.

Métodos multivariados

Análisis de Regresión Multivariada.

Es el método que trata de estudiar la relación funcional entre un grupo de variables dependientes y un conjunto de variables independientes. La relación puede ser lineal o no lineal y las variables independientes pueden ser métricas y no métricas.

Metodología para el análisis multivariante de datos

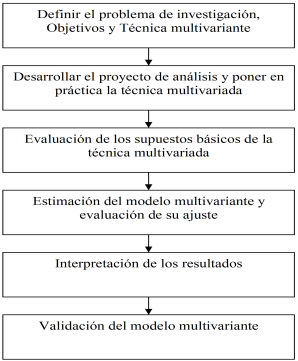

Figure 4: Metodología

Análisis previo de datos

Análisis previo de datos

Antes de aplicar cualquier técnica multivariada es preciso un análisis de los datos. Es necesario evaluar las variables individuales y sus relaciones, tales como datos faltantes, presencia de datos atípicos y supuestos.

- Análisis exploratorio de datos
- Análisis de datos faltantes (Missing)
- Detección de datos atípicos coutliers)
- Comprobación de supuestos

Análisis exploratorio de datos

- Histogramas de frecuencias
- Gráficos
- Diagrama de tallos y hojas
- Diagrama de cajas
- Cálculo de medidas estadísticas
- Medidas de asimetria
- Medidas de curtosis

Análisis de datos faltantes

Supresión de datos

- Eliminación de casos o variables
- Eliminación de datos según parejas de variables

Imputación de datos

- Sustitución por la media o mediana (hay outliers)
- Sustitución por interpolación (alta variabilidad). Uso de valores adyacentes
- Sustitución por datos constante. Uso de fuentes externas
- Sustitución por análisis de regresión

Detección de datos atípicos (outliers)

- Diagrama de cajas simples y múltiples
- Gráficos de dispersión
- Gráfico de dispersión matricial
- Gráficos de control
- Estadísticas descriptivas

Distribución Normal Multivariada

Prueba de normalidad p-variada

Con el Software R se hará la prueba de normalidad p-variada de Shapiro.

- H0: Las p variables tienen distribución normal p-variada.
- H1: Las p variables no tienen distribución normal p-variada.

EJEMPLO DE APLICACIÓN