Qué se entiende por análisis multivariado?

El análisis multivariado consiste en un conjunto de métodos que pueden ser usados cuando muchas características (variables) han sido medidas para cada objeto de estudio en una o múltiples muestras, permite procedimientos descriptivos e inferenciales.

Que se entiende por análisis multivariado?

El análisis de datos multivariados comprende el estudio estadístico de un conjunto grande de variables medidas en elementos de una población con los siguientes objetivos:

- reducir dimensiones: resumir los datos originales con nuevas variables, transformando las originales, perdiendo la menor cantidad de información. El número de nuevas variables es significativamente menor al original.

- Construir tipologías de individuos: encontrar, si existen, grupos en los datos.

- Encontrar reglas de clasificación que permitan predecir y clasificar nuevas observaciones.

- Estudiar la asociación de variables y/o modalidades de dichas variables.

Que se entiende por análisis multivariado?

En una visión más tradicional las técnicas multivariantes serían aquellas que implican reducción de dimensiones, clasificación de objetos y modelos de inferencia multivariada. En el primer grupo se encuentran los componentes principales, el análisis de correspondencias y el análisis factorial. Las técnicas de clasificación lo constituyen el análisis de cluster y el discriminante.

En una visión más amplia el análisis multivariado podrá involucrar distintas técnicas asociadas con la inteligencia artificial: árboles de decisión, Bagging, Boosting, Random Forest, SVM, técnicas vinculadas al análisis de series temporales: modelos VAR, modelos de corrección de error, etc.; así como técnicas relacionadas con datos longitudinales.

Que se entiende por análisis multivariado?

Las técnicas de análisis multivariado tienen aplicaciones en todas las disciplinas; comenzaron desarrollándose para resolver problemas de clasificación en Biología, se extendieron para encontrar índices en las ciencias sociales, marketing, psicometría y han alcanzado una gran aplicación en Ingeniería y Ciencias de la Computación como herramientas para resumir la información y diseñar sistemas de clasificación automática y reconocimiento de patrones.

En la actualidad

En los últimos años los métodos multivariados están sufriendo grandes transformaciones:

- La gran cantidad de datos disponibles esta conduciendo al desarrollo de métodos de aproximación local, que no requieren hipótesis generales sobre el conjunto de observaciones.

- Se prescinde de las hipótesis de las distribuciones de los datos y cuantifica la incertidumbre mediante métodos de computación intensiva.

El desarrollo de la informática, la facilidad de procesar bases de datos de grandes dimensiones han convertido estas técnicas en herramientas de uso común en muchas disciplinas:sociología, educación, psicología, medicina, economía, finanzas, etc. A modo de ejemplo se pueden mencionar: Estudios multidimensionales de la pobreza, Modelos que estudian los determinantes del delito, Scoring de clientes en empresas financieras, Factores que contribuyen a la deserción estudiantil, etc.

Clasificación de técnicas de análisis multivariado

1. Técnicas exploratorias: componentes principales, correspondencias, clasificación, etc.

2. Técnicas explicativos, inferenciales: regresión lineal, logística, análisis factorial confirmatorio.

Historia análisis multivariado

A continuación se presenta una breve reseña histórica de las técnicas de análisis multivariado, la misma se basa en el libro Análisis de Datos Multivariantes, de Daniel Peña (2002) e incorpora algunos elementos más recientes.

- Los primeros estudios descriptivos para encontrar relaciones entre variables son debidos a **Adolfo Quetelet** (1796-1874), astrnomo belga que inició la aplicación de los métodos de probabilidad a las ciencias sociales.

- El primer método para medir la relación estadística entre dos variables es debido a **Francis Galton** (1822-1911) que introduce el concepto de recta de regresión y la idea de correlación entre variables. Galton buscaba explicar la asociación entre la altura de los padres y la altura de los hijos.

- **Kar Pearson** (1845-1926), estadístico británico creador del contraste Chi-cuadrado, obtuvo el estimador del coeficiente de correlación en datos muestrales.

- **Harol Hotelling** (1885-1973) que pasó del periodismo a la matemática y a la economía, se interesó por el problema de comparar tratamientos agrícolas en función de varias variables. Debemos a Hotelling (1931) el contraste de medias de poblaciones multivariadas que lleva su nombre, que permite comprobar si dos muestras multivariantes provienen de la misma población. En 1933 Hotelling inventó los componentes principales, indicadores capaces de resumir en forma óptima un conjunto amplio de variables y que dan lugar al análisis factorial. Posteriormente Hotelling generaliza la idea de componentes principales introduciendo el análisis de correlaciones canónicas, que permiten resumir simultáneamente dos conjuntos de variables.

- El primer análisis factorial, en su visión anglosajona como técnica para explicar un conjunto de variables observadas por un pequeño número de variables latentes o no observables (factores) fue planteado por **Charles Spearman** (1863-1945), psicólogo inglés, que intentó estimar objetivamente la inteligencia de niños británicos. El análisis factorial así planteado, a diferencia de la técnica de componentes principales, presupone un modelo estadístico formal en la generación de datos. El análisis factorial fue considerado hasta los años setenta como una técnica psicométrica con poca base estadística, hasta que trabajos de Lawley y Maxwell (1971) establecieron formalmente la estimación y el contraste del modelo factorial.

- La primera solución al problema de clasificación es debida a **Ronald Fisher** (1890-1962). Fisher propone el análisis discriminante para dos poblaciones, intentando resolver un problema de discriminación en antropología. El método se basa en el análisis de varianza. La idea de Fisher es encontrar una variable indicadora, combinación

lineal de las variables originales de las medidas del cráneo, que consiga la máxima separación entre las dos poblaciones consideradas. Fisher unifica sus ideas con las medidas de distancias planteadas por **P.C. Mahalanobis** (1893-1972) y con los contrastes de medias de poblaciones multivariadas propuesto por Hotelling. Calyampudi R.Rao (1926) extiende la idea de Fisher a más de dos poblaciones.

- Las ideas anteriores refieren a variables cuantitativas, pero se aplican poco después a variables cualitativas. Pearson introdujo el estadístico que lleva su nombre para contrastar la independencia en una tabla de contingencia.

- En los años setenta **Jean Paul Benzecri** (1930) publica métodos de análisis de datos cualitativos mediante el análisis de correspondencias, análisis planteado desde un enfoque geométrico.

- En el seno de la Escuela de Estadística Francesa (Analyse des Donees) además del desarrollo de las técnicas factoriales mencionadas se encuentran las técnicas de Análisis Multiway, análisis de datos a múltiples vías, entre las que se puede mencionar el STATIS, el Análisis Factorial Múltiple (AFM), el Análisis Factorial Dinámico (AFD). El análisis factorial múltiple fue desarrollado por **Brigitte Escoffier** y **J. Pages** en 1985, permite el estudio simultáneo de tablas en las que un mismo conjunto de individuos se describe a través de varios grupos de variables. El objetivo es obtener un sistema común de ejes que permita la representación simultánea de las subtablas. El STATIS (Structuration des Tableux A Tros Indices de la Statistique) fue introducido en 1976 por L'Hermier des Plantes y desarrollado por Lavit en 1988, permite el análisis simultáneo de diferentes tablas numéricas referidas a las mismas o distintas variables y a un determinado conjunto de individuos.

- El análisis factorial dinámico (en el marco de la escuela francesa) fue desarrollado por **Renato Coppi** y **Jorge Blanco** en los años 90, es una técnica multiway que analiza matrices a tres vías (unidades, variables, ocasiones) y constituye una aproximación metodológica al análisis de la dispersión medida a través de matrices de varianzas y matrices de proximidades.

- La aparición de la computadora transforma los métodos de análisis multivariado que experimentan un gran crecimiento desde los años setenta. En el campo descriptivo hacen posible la aplicación de métodos de clasificación (análisis de conglomerados o análisis de cluster) que se basan en uso intensivo de la computadora. **Mac Queen**, en 1967 introduce el algoritmo de K-medias.

- En cuanto a los métodos de clasificación supervisada que requieren un uso intensivo de la computadora se encuentran los árboles de clasificación, entre ellos los de tipo CART creado por **Breiman,**

Friedman, Olshen, Stone en 1984. El desarrollo de este tipo de técnicas data de los años sesenta donde Morgan y Sonquist platean el algoritmo AID (Automatic Integration Detection), le suceden el CHAID (Chi-Square Automatic Integration Detection), creado por Kass en 1980, el QUEST (Quick Unbiased Efficient Statistical Tree), creado por Lhon y Shin en 1997, y el CRUISE, entre otros.

- Mejoras a estas técnicas de clasificación son el Random Forest y el Bagging introducidas por Breiman en 1994. Estas últimas técnicas pueden ser encuadradas dentro de la llamada Inteligencia Artificial, Machine Learning, y no suelen ser presentadas estrictamente como técnicas multivariadas. La técnica SVM (Support Vector Machine) es parte de la nueva generación de los sistemas de aprendizaje basado en los desarrollos recientes de la teoría de aprendizaje estadístico. Junto con las Redes Neuronales se han convertido en las herramientas más potentes en el marco de Machine Learning y Data Mining. Fueron creados por **Vapnik** (1998), **Cherkassky y Mulier** (1998).