Expositor: Agustin Alvarez (Universidad Nacional de General Sarmiento, aalvarez@dm.uba.ar)
Autor/es: Agustin Alvarez (Universidad Nacional de General Sarmiento, aalvarez@dm.uba.ar); Marcela Svarc (Conicet, Universidad de San Andres, msvarc@udesa.edu.ar)

El concepto de medidas de profundidad multivariadas, como fue definido por Zou y Serfling [ZS00], al ser aplicado a distribuciones univariadas está estrechamente ligado al concepto de cuantiles en $\mathbb{R}$. Sin embargo para datos multivariado no existe una noción natural de orden. Las medidas de profundidad permiten tener una noción de orden de cuán "adentro" se encuentra un dato respecto a una distribución multivariada.

Hemos propuesto una técnica para seleccionar variables. A partir de un vector aleatorio $\mathbf{X} = (\mathbf{X_1}, \ldots, \mathbf{X_p}) \in \mathbb{R}^{\mathbf{P}}$ con $\mathbf{X} \sim \mathbf{P}$ y una medida de profundidad $D = D(\mathbf{x}, \mathbf{Q})$ definida para $\mathbf{x} \in \mathbb{R}^{\mathbf{k}}$ y Q una probabilidad Q en $\mathbb{R}^k$, para cualquier k, proponemos un método que para cada k, con $1 \leq k < p$, selecciona k variables, con el objetivo de que la profundidad de los datos restringidos (a las variables seleccionadas) sea lo más parecida posible a la profundidad de los datos originales (contemplando todas las variables). Para medir la similitud entre las profundidades de los datos restringidos y las de los originales contemplamos dos posibilidades: 1) maximizar la correlación entre las profundidades de todas las observaciones o 2) minimizar la media de las distancias al cuadrado. Sin embargo para este último caso "estandarizamos" de alguna manera las profundidades ya que medir profundidades en distintas dimensiones puede dar bastante distinto. Para la minimización 2) probamos, bajo condiciones generales, que el estimador de las variables que minimizan es consistente al minimizador poblacional.

Al implementar el programa para lograr la optimización (1) o 2)), mientras la cantidad de variables p es pequeña y la cantidad de posibles subconjuntos de tamaño k es moderada podemos realizar una búsqueda exhaustiva, sin embargo cuando la cantidad de subconjuntos resulta grande optimizamos mediante un Algoritmo genético. Para poder comparar entre subconjuntos de variables que optimizan en distintas dimensiones k, proponemos una penalización en la cantidad de variables con el fin de obtener una solución parsimoniosa y esquivar problemas de sobre-ajuste. Proponemos encontrar el parámetro de penalización mediante una convalidación cruzada de $k^\star$ grupos (o iteraciones).

Realizamos un estudio de MonteCarlo para poner a prueba el método propuesto en un ejemplo con diversas variantes y probamos también el método en un ejemplo de datos reales: los datos de bienestar de la Encuesta Permanente de Hogares entre los años 2004 y 2014.

Referencias

[ZS00] Zuo, Y. and Serfling R. (2000). "General Notion of Statistical Depth Function", in: *The Annals of Statistics*, vol **28,** (2), 461-482.