Máster en Técnicas Estadísticas

PROPUESTA DE TRABAJO FIN DE MÁSTER (Modalidad A)
Curso 2020-2021
MÁSTER EN TÉCNICAS ESTADÍSTICAS

Título	Métodos cluster basados en la detección de modas
Director/es	Rosa M. Crujeiras Casais Jose Ameijeiras Alonso (KU Leuven)
Descripción del contenido	El objetivo principal que persiguen las **técnicas de agrupamiento** (*clustering*) es el obtener, a partir de un conjunto de datos, una serie de grupos de manera que los miembros de un mismo grupo (cluster) sean similares entre sí o más similares con respecto a los datos que pertenecen a otros grupos. Existen diversas metodologías o algoritmos que permiten crear estos grupos (véase, por ejemplo, Hennig *et al.*, 2015), como por ejemplo, los algoritmos basados en *conectividad*, donde definiendo una distancia, los datos más cercanos están más relacionados que los que están alejados; los basados en centroides, que asignan cada dato con el grupo que tiene el centro más cercano; o los **basados en distribuciones**, donde el agrupamiento se hace en función a la probabilidad que tiene un dato de pertenecer a cierto grupo. El objetivo de este trabajo se centra en la **revisión de las técnicas** pertenecientes a este último enfoque. Además, dentro de esta última **aproximación a la agrupación basada en la densidad**, se pueden distinguir dos metodologías principales dependiendo del concepto de agrupación que se adopte (Chacón, 2020): **paramétrico y no paramétrico**. Por un lado, los métodos paramétricos (métodos basados en modelos de probabilidad) supone, en general, que la densidad subyacente que genera los datos es una mixtura. En este enfoque los datos se agrupan en función de la probabilidad que tiene cada dato de pertenecer a cierta componente de la mixtura. En el caso de los métodos no paramétricos, los grupos se definen en función de donde se concentra la masa de probabilidad (Menardi, 2016). Una forma de caracterizar estas regiones es en términos de las modas (máximos relativos) de la función de densidad. En este enfoque, se asocian los datos a la moda más cercana., siendo el algoritmo *mean shift* uno de los más populares (véase https://spin.atomicobject.com/2015/05/26/mean-shift-clustering/). Finalmente, otra forma de caracterizar las regiones es a través de los conjuntos de nivel (Hartigan, 1975), donde los clusters se asocian a regiones donde la densidad excede cierto nivel. El objetivo de este trabajo será la revisión y comparativa en escenarios simulados y sobre datos reales, de las técnicas paramétricas y no paramétricas para la agrupación de datos, donde esta agrupación se realiza basándose en la estimación de la densidad (mediante mixturas o de forma no paramétrica). <u>Bibliografía</u> Chacón, J. E. (2020). The modal age of statistics. *International Statistical Review*, *88*(1), 122-141. Hennig, C., Meila, M., Murtagh, F., y Rocci, R. (Eds.). (2015). *Handbook of cluster analysis*. CRC Press, Boca Raton. Hartigan, J. A. (1975). *Clustering algorithms*. John Wiley & Sons, Inc, Nueva York. Menardi, G. (2016). A review on modal clustering. *International Statistical Review*, *84*(3), 413-433.

Recomendaciones	Es recomendable que el/la alumno/a tenga conocimientos de análisis multivariante y de estimación tipo núcleo de la densidad. También se recomienda tener un buen manejo de R dado el trabajo computacional que se deberá realizar.
Otras observaciones	La bibliografía recomendada para este tema está en inglés.