



Regresión en alta dimensión

(Fecha última actualización 30 septiembre 2022)

Máster en Estadística Aplicada. CURSO ACADÉMICO 2022-2023	
Título	Métodos de regularización y selección de variables en R
Tipo	INVESTIGACIÓN <input type="checkbox"/> ORIENTACIÓN PRÁCTICA <input checked="" type="checkbox"/>
Profesor(es)/ email	María Dolores Martínez Miranda (mmiranda@ugr.es)
Descripción	<p>Los modelos de regresión en dimensiones elevadas suponen un reto tanto computacional como teórico. La inclusión de predictores en un modelo de regresión lineal múltiple no es gratis y existe un precio a pagar en términos de variabilidad añadida, difícil interpretación, además de la posible inclusión de predictores altamente correlados. Por otro lado, si el número de predictores es muy elevado, incluso superior al número de observaciones los estimadores mínimo-cuadráticos estándar no son adecuados.</p> <p>Seleccionar a priori el subconjunto más adecuado de predictores, o construir estimadores que lo hagan automáticamente, permitiría reducir la dimensión y simplificar la interpretación de modelo. Sin embargo la solución al problema no es sencilla, ni es única, es más, no existe una solución única que funcione en todos los casos.</p> <p>El entorno de análisis estadístico y programación R proporciona diversas herramientas para resolver problemas de regresión lineal en alta dimensión. Además de existir algunas funciones elementales en el sistema base, en CRAN se pueden encontrar varios paquetes específicos que implementan los últimos avances en este campo.</p> <p>Esta propuesta de TFM supondrá para el estudiante una aproximación computacional práctica al problema de regresión en alta dimensión poniendo el foco en los métodos de selección de variables y regularización. Una vez que haya adquirido las nociones teóricas básicas el resto del trabajo se centrará en aplicaciones prácticas en R usando datos reales y simulaciones.</p>
Objetivos particulares	<ol style="list-style-type: none"> 1. Tomar conciencia del problema de la alta dimensionalidad en la regresión lineal múltiple, tanto desde el punto de vista teórico como práctico. 2. Conocer los métodos modernos más relevantes para reducir la dimensión, haciendo hincapié en los métodos de selección de variables y técnicas de regularización. 3. Reconocer algunas de estas técnicas dentro del contexto y formulación del Machine Learning. 4. Revisar y documentar las herramientas disponibles en R para la computación de los métodos descritos. 5. Realizar aplicaciones con datos reales y simulados.
Prerrequisitos y recomendaciones	Haber cursado alguna asignatura donde se incluyan contenidos de modelos de regresión y su inferencia. Se requiere destreza en el uso del entorno de programación y análisis estadístico R, incluyendo habilidad para estudiar paquetes específicos disponibles en CRAN y realizar tareas de programación de funciones.
Plan de trabajo	<ol style="list-style-type: none"> 1. Revisión de los aspectos teóricos. 2. Implementación computacional: estudio de herramientas disponibles R y creación de funciones originales en su caso. 3. Aplicaciones con datos reales y datos simulados. 4. Reflexión sobre limitaciones y extensiones de la metodología considerada. 5. Elaboración de las conclusiones del trabajo.
Competencias generales y específicas	<p>Competencias generales:</p> <p>CG1 - Los titulados han de saber aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio.</p>



CG7 - Los titulados han de realizar una contribución a través de una investigación original que amplíe las fronteras del conocimiento desarrollando un corpus sustancial, del que parte merezca la publicación referenciada a nivel nacional o internacional.
CG8 - Los titulados deben ser críticos en el análisis, evaluación y síntesis de ideas nuevas y complejas.

Competencias específicas:

CE1 - Conocer métodos para el Análisis de Datos
CE5 - Adquirir conocimientos avanzados en Inferencia Estadística
CE17 - Adquirir capacidades de elaboración y construcción de modelos y su validación
CE18 - Ser capaz de realizar un análisis de datos
CE20 - Ser capaz de realizar una correcta representación gráfica de datos
CE21 - Conocer, identificar y seleccionar fuentes estadísticas
CE22 - Ser capaz de interpretar resultados a partir de modelos estadísticos
CE24 - Ser capaz de extraer conclusiones y redactar informes
CE25 - Ser capaz de identificar relaciones o asociaciones
CE26 - Saber utilizar con destreza entornos de programación y análisis estadístico

Bibliografía

1. Faraway, J. (2014). Linear Models with R. Chapman & Hall/CRC Texts in Statistical Science.
2. Hastie, T., Tibshirani, R. y Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics. Springer, New York.
3. Hastie, T., Tibshirani, R. y Tibshirani, R. (2020). Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons. Statistical Science, 35, No. 4, 579-592.
4. James, G., Witten, D., Hastie, T., Tibshirani, R. (2017). An Introduction to Statistical Learning: with Applications in R. Springer Texts in Statistics. Springer, New York.
5. Seber, G.A.F. (1984) Multivariate Observations. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. New York: John Wiley & Sons.