UNIVERSIDAD DE GRANADA

□

Regresión en alta dimensión

Máster en Estadística. CURSO ACADÉMICO 2023-2024	
Título	Una revisión crítica del método LASSO para la selección de variables en regresión lineal
Tipo	INVESTIGACIÓN ☒ ORIENTACIÓN PRÁCTICA ☐
Número de alumnos	1
Profesor(es)/ email	María Dolores Martínez Miranda (mmiranda@ugr.es)
Descripción	Los modelos de regresión en dimensiones elevadas, con un número elevado de predictores (p>n, siendo n el número de observaciones) suponen un reto tanto computacional como teórico. Mientras que en el pasado era habitual recoger unas pocas variables para cada observación, cuidadosamente seleccionadas, hoy en día es habitual medir cualquier característica que pueda tener el más mínimo efecto en la respuesta. En muchas áreas de aplicación esto puede dar lugar a ratios del orden p/n>1000. Por supuesto muchas de estas variables tendrán un efecto casi nulo, sin embargo, es difícil conocerlo de antemano. La cuestión es cómo proceder en estos casos. El estimador por mínimos cuadrados estándar en regresión lineal múltiple no es adecuado en esta situación de alta dimensión. Incluso en situaciones menos extremas, donde la inclusión de predictores en el modelo de regresión lineal no es gratis y existe un precio a pagar en términos de variabilidad añadida, difícil interpretación, además del riesgo de multicolinealidad severa. Sería deseable por tanto disponer de un estimador que primero detectara las variables más relevantes y después estimara sus coeficientes. Uno de los métodos más popular en regresión en alta dimensión es denominado LASSO (Least Absolute Shrinkage and Selection Operator) introducido por Robert Tibshirani en 1996. Se trata de un método de regularización que además de permitir reducir la dimensión seleccionando predictores, da lugar a estimadores con buenas propiedades predictivas. Desde la introducción de LASSO se han desarrollado diversos estudios tanto teóricos como prácticos para comprender sus propiedades, lo que ha dado lugar a diversas variantes que pretenden resolver algunas de sus debilidades y limitaciones en la práctica. Este trabajo consiste en una revisión del método LASSO y de sus variantes más relevantes desde una perspectiva crítica. El estudiante realizará ejercicios de simulación en diferentes escenarios y aplicaciones prácticas con datos para evaluar y comprender el funcionamiento de LASSO y algunas de sus variantes.
Objetivos particulares	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 regularización y en especial en LASSO. Ser capaz de identificar sus ventajas e inconvenientes. 3. Reconocer algunas de estas técnicas dentro del contexto y formulación del Machine Learning. 4. Realizar experimentos de simulación utilizando el entorno de análisis y programación estadística R y los paquetes específicos más relevantes para este tema. 5. Realizar aplicaciones con datos reales.

Prerrequisitos y recomendaciones	Haber cursado alguna asignatura donde se incluyan contenidos de modelos de regresión y su inferencia. Además de destreza en el manejo del entorno de programación y análisis estadístico R, que incluya programación a un nivel medio.
Plan de trabajo	1. Revisión bibliográfica. 2. Desarrollo de la parte teórica del trabajo. 3. Experimentos de simulación y aplicaciones con datos reales utilizando paquetes apropiados de R. 4. Conclusiones del trabajo.
Competencias generales y específicas	Competencias generales: 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. 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. Bühlmann, P. y van de Geer, S. (2011) Statistics for High-Dimensional Data: Methods, Theory 2. and Algorithms. Springer Series in Statistics. 3. Freijeiro-González, L., Febrero-Bande, M. y González-Manteiga, W. (2022). A Critical Review of LASSO and Its Derivatives for Variable Selection Under Dependence Among Covariates. International Statistical Review, 90, 1, 118-145. DOI: 10.1111/insr.12469 4. 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. 5. 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. 6. 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.