Universidad de Granada

GRADO EN ESTADÍSTICA
PROPUESTA DE TEMA PARA TRABAJOS FIN DE GRADO
CURSO ACADÉMICO 2021/2022

Título del trabajo: Regularización y selección de variables en regresión lineal

Tutora: María Dolores Martínez Miranda

Correo electrónico: mmiranda@ugr.es

Cotutor/a:

Departamento responsable: Estadística e I.O.

Perfil y número de estudiantes al que va dirigido (máximo 2): 1

(Rellenar sólo en caso de que la propuesta esté realizada a través de estudiante)

Estudiante que propone el trabajo (Nombre, Apellidos, DNI):

Amanda Estela Figueredo

Tipo de trabajo *(consultar (*))* **1, 7, 8, 9**

Competencias (*estas son las mínimas; consultar (**) si se considera añadir otras*)

Competencias generales: G02, G03, G04, G05, G06, G07, G08, G09.

Competencias específicas: E01, E02, E03, E04, E05, E08, E09, E10.

Resultados de aprendizaje *(estos son los mínimos; añadir otros si se considera)*

- Adquirir competencias globales ligadas al desarrollo y aplicación de los conocimientos del Grado.
- Adquirir competencias ligadas a la búsqueda y organización de información y documentación relevante sobre el tema objeto de estudio.
- Aplicar el "pensamiento estadístico" y tener capacidad para enfrentarse a las distintas etapas de un estudio estadístico (desde el planteamiento del problema hasta la exposición de resultados).
- Saber presentar, de forma escrita y oral, la memoria, los resultados y las conclusiones del trabajo realizado.

Antecedentes y resumen del tema propuesto:

Los métodos de regresión lineal múltiple permiten describir de una manera relativamente sencilla la relación de dependencia estadística entre una variable de respuesta y un conjunto de variables explicativas. Cuando el conjunto de variables explicativas es muy grande existen al menos dos importantes razones que hacen que los estimadores mínimo-cuadráticos habituales no resulten convenientes en la práctica:

1. Pérdida de precisión en la predicción: Si el número de observaciones (n) es mucho mayor que el de variables explicativas (p), los estimadores tenderán a tener poca varianza y las predicciones una precisión aceptable. Sin embargo, si n no es mucho mayor que p, el ajuste por mínimos cuadrados tendrá mucha variabilidad, lo que puede causar un sobre-ajuste y que las predicciones futuras sean muy poco precisas.
2. Pérdida de interpretabilidad: Cuando p es muy grande puede darse el caso que alguna (o muchas) de las variables explicativas sean irrelevantes a la hora de explicar la variable respuesta. Incluir variables irrelevantes añade complejidad innecesaria al modelo, complicando la interpretación.

Una forma de tratar el primer problema consiste en reducir la varianza a costa de incrementar un poco el sesgo. Para el segundo problema la solución está en reducir el número de variables explicativas (lo que a su vez resolvería también el primer problema). En esta línea se introducen los denominados métodos de selección de variables y los métodos de regularización (regresión ridge, lasso y sus generalizaciones, etc.).

Este trabajo supone una introducción del estudiante a los métodos más relevantes para la se-

lección variables y en general la selección de modelos en el contexto de la regresión lineal múltiple. Si bien de los estudios del grado el estudiante ya conoce algunas de las técnicas clásicas (selección forward, backward, stepwise, subconjunto óptimo) ahora tiene la oportunidad y la tarea de profundizar en este tema cubriendo los aspectos teóricos y prácticos, con ayuda de ilustraciones con datos reales y simulados, y su análisis con R.

Breve descripción de las actividades presenciales y no presenciales a realizar:

Actividades presenciales (15-30%)	Planteamiento, orientación y supervisión	18 horas
	Exposición del trabajo	2 horas
	Otras:	
Actividades no presenciales (70-85%)	Preparación del trabajo	180 horas
	Elaboración de la memoria	100 horas
	Otras:	
Total (12 ECTS)		300 horas

Objetivos que se pretenden alcanzar:

- Tomar conciencia del problema de la alta dimensionalidad en la regresión lineal múltiple, tanto desde el punto de vista teórico como práctico.
- Conocer los métodos clásicos y algunos de los métodos modernos más relevantes para reducir la dimensión, haciendo especial hincapié en los métodos selección de variables. Ser capaz de identificar sus ventajas e inconvenientes.
- Reconocer algunas de estas técnicas dentro del contexto y formulación del Machine Learning.
- Realizar aplicaciones con datos reales y simulados en el entorno de análisis y programación estadística R.

Bibliografía básica para la puesta en marcha del trabajo:

FARAWAY, J. (2014). Linear Models with R. Chapman & Hall/CRC Texts in Statistical Science.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer Series in Statistics. Springer, New York.

HASTIE, T., TIBSHIRANI, R. and TIBSHIRANI, R. (2020). Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons. Statistical Science, Vol. 35, No. 4, 579-592.

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.

Tipo de trabajo (*):

1. Estudio de profundización en algún tema concreto de Estadística, o como proyecto de aplicación de la misma a estudios o problemas de otros ámbitos científicos o sociales.
2. Realización completa de todas las fases de un proyecto estadístico, bien con auxilio de prácticas en empresas o con prácticas propuestas y dirigidas por el tutor.

3. Estudio de casos, teóricos o prácticos, relacionados con la Estadística.
4. Elaboración de un informe o un proyecto de naturaleza profesional.
5. Elaboración de un plan de empresa.
6. Simulación de encargos profesionales.
7. Trabajos bibliográficos sobre el estado actual de una temática relacionada con la Estadística.
8. Creación y/o empleo de herramientas informáticas para su uso en Estadística.
9. Trabajos de inicio a la investigación.
10. Trabajos cuya finalidad sea la divulgación de la Estadística en diversos contextos.
11. Trabajos sobre Historia de la Estadística.
12. Trabajos relacionados con la docencia de la Estadística.

Competencias (**)

Competencias generales:

G01. Poseer los conocimientos básicos de los distintos módulos que, partiendo de la base de la educación secundaria general, y apoyándose en libros de texto avanzados, se desarrollan en la propuesta de título de Grado en Estadística que se presenta.

G02. Saber aplicar los conocimientos básicos de cada módulo a su trabajo o vocación de una forma profesional y poseer las competencias que suelen demostrarse por medio de la elaboración y defensa de argumentos y la resolución de problemas dentro de la Estadística y ámbitos en que esta se aplica directamente.

G03. Saber reunir e interpretar datos relevantes para emitir juicios que incluyan una reflexión sobre temas relevantes de índole social, científica o ética.

G04. Poder transmitir información, ideas, problemas y sus soluciones, de forma escrita u oral, a un público tanto especializado como no especializado.

G05. Haber desarrollado aquellas habilidades de aprendizaje necesarias para emprender estudios posteriores con un alto grado de autonomía.

G06. Saber utilizar herramientas de búsqueda de recursos bibliográficos.

G07. Poder comunicarse en otra lengua de relevancia en el ámbito científico.

G08. Poseer habilidades y aptitudes que favorezcan el espíritu emprendedor en el ámbito de aplicación y desarrollo de su formación académica.

G09. Fomentar y garantizar el respeto a los Derechos Humanos, a los principios de accesibilidad universal, igualdad, y no discriminación; y los valores democráticos, de la cultura de la paz y de igualdad de género.

Competencias específicas:

E01. Conocer los fundamentos básicos del razonamiento estadístico, en el diseño de estudios, en la recogida de información, en el análisis de datos y en la extracción de conclusiones.

E02. Conocer, saber seleccionar y saber aplicar, técnicas de adquisición de datos para su tratamiento estadístico.

E03. Conocer los fundamentos teóricos y saber aplicar modelos y técnicas estadísticas en estudios y problemas reales en diversos ámbitos científicos y sociales.

E04. Saber seleccionar los modelos o técnicas estadísticas para su aplicación en estudios y problemas reales en diversos ámbitos científicos y sociales, así como conocer herramientas de validación de los mismos.

E05. Comprender la importancia de la Investigación Operativa como metodología de optimización, toma de decisiones y diseño de modelos particulares para la resolución de problemas en situaciones específicas.

E06. Comprender y utilizar básicamente el lenguaje matemático.

E07. Conocer los conceptos y herramientas matemáticas necesarias para el estudio de los aspectos teóricos y prácticos de la Probabilidad, la Estadística y la Investigación Operativa.

E08. Conocer y saber utilizar aplicaciones informáticas de análisis estadístico, cálculo numérico y simbólico, bases de datos, visualización gráfica y optimización, que sean útiles para la aplicación y desarrollo de las técnicas estadísticas.

E09. Conocer los conceptos básicos y habilidades propias de un ámbito científico o social en el que la Estadística o la Investigación operativa sean una herramienta fundamental.

E10. Tomar conciencia de la necesidad de asumir las normas de ética profesional y las relativas a la protección de datos y del secreto estadístico, como premisas que deben guiar la actividad profesional como profesionales de la Estadística.