UNIVERSIDAD DE GRANADA

FACULTAD DE CIENCIAS

Propuesta TFG. Curso 2024/2025

GRADO: Grado en Matemáticas

CÓDIGO DEL TFG: 270-020-2024/2025

1. DATOS BÁSICOS DEL TFG:

Título: Bootstrap y ensamblado en modelos de aprendizaje estadístico

Descripción general (resumen y metodología):

El aprendizaje estadístico (statistical learning) se refiere a un conjunto de herramientas para modelizar y comprender datos complejos. Tales herramientas suelen clasificarse en supervisadas o no supervisadas. El aprendizaje estadístico supervisado engloba métodos predictivos, en los que una de las variables que intervienen en el problema se define como variable respuesta. Dependiendo del tipo de respuesta se diferencia entre clasificación (respuesta categórica) o regresión (respuesta cuantitativa). El aprendizaje no supervisado busca descubrir relaciones y estructuras en los datos y, aunque intervienen variables (inputs) no se considera ninguna variable de respuesta (output).

Con la explosión del denominado "Big Data", el aprendizaje estadístico se ha convertido en tema de gran actualidad e interés en muchas disciplinas, siendo sus herramientas muy demandadas en el mundo empresarial.

Independientemente del método de aprendizaje estadístico que nos planteemos utilizar, la elección del modelo concreto y su nivel de complejidad es crucial. En todo problema de aprendizaje estadístico se busca reducir el sesgo y la varianza. Una mayor flexibilidad en el modelo, así como la incorporación de diversas variables relevantes al problema, permitirán reducir el sesgo, sin embargo esto conllevará un incremento de la varianza, implicando posiblemente un sobre-ajuste no deseado. Entender y encontrar el equilibrio entre varianza y sesgo es clave en la aplicación de este tipo de métodos.

Los denominados métodos de ensamblado hacen uso de distintas técnicas estadísticas y matemáticas, como el bootstrap o los algoritmos de optimización, para mejorar las capacidades predictivas de los modelos de aprendizaje estadístico, permitiendo a su vez encontrar un equilibrio óptimo entre sesgo y varianza. En términos generales, estos métodos buscan acoplar, o ensamblar, una serie de modelos de base hasta alcanzar un modelo en el que se consiga el deseado equilibrio entre sesgo y varianza. Se suele usar la terminología modelos de aprendizaje débil para los modelos de base y modelo de aprendizaje fuerte para el modelo final. Entre los métodos de ensamblado más populares se encuentran el denominado bagging, que combina el ensamblado y el bootstrap, el boosting y el stacking.

El aprendizaje estadístico no debe usarse como una "caja negra". Por lo general ningún método de aprendizaje estadístico funcionará bien en todos los casos y elegir el más adecuado requerirá un profundo conocimiento del mismo. Este trabajo supone una introducción al aprendizaje estadístico desde una perspectiva estadística matemática. Tras una adecuada contextualización se hará especial hincapié en los métodos de aprendizaje supervisados, y en los métodos de ensamblado como herramienta para encontrar un equilibrio óptimo entre flexibilidad y complejidad.

Tipología: Estudio de casos, teóricos o prácticos, relacionados con la temática del Grado.

Objetivos planteados:

1. Aprender los conceptos y terminología básica del aprendizaje estadístico, así como las diferencias entre aprendizaje supervisado y no supervisado.
2. Describir las bases y fundamentos de algunos de los métodos de aprendizaje supervisado más relevantes.
3. Comprender el problema de equilibrar sesgo y varianza y su relación con la complejidad y flexibilidad de los modelos de aprendizaje estadístico.

4. Estudiar medidas para la evaluación y validación de modelos, incluyendo validación cruzada y bootstrap.
5. Comprender la motivación y objetivos generales de los métodos de ensamblado en el contexto de los modelos de aprendizaje estadístico supervisado.
6. Conocer y describir algunos de los métodos de ensamblado más relevantes.
7. Realizar aplicaciones en R y/o Python con datos reales y/o simulados.

Bibliografía básica:

1. Bühlmann, P. and van de Geer, S. (2011). Statistics for High-Dimensional Data: Methods, Theory and Algorithms. Springer Series in Statistics.
2. Hastie, T., Tibshirani, R. and Friedman, J. (2009). The Elements of Statistical Learning. Springer.
3. James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. Springer.
4. Joshi, A.V. (2023). Machine Learning and Artificial Intelligence. Springer.
5. Kuncheva, L. (2014). Combining Pattern Classiffiers: Methods and Algorithms. Wiley
6. Mohri, M., Rostamizadeh, A. and Talwalkar, A. (2018). Foundations of Machine Learning. MIT Press.
7. Rhys, H. (2020). Machine Learning with R, the tidyverse, and mlr. New York, Manning.

Recomendaciones y orientaciones para el estudiante:

Es recomendable que el estudiante haya cursado la asignatura de Estadística Computacional del Grado en Matemáticas y tenga habilidades en temas de computación.
Para una introducción en el tema de esta propuesta se recomienda comenzar con una lectura de los libros de Hastie et al. (2009) y James et al. (2013).

Plazas: 1

2. DATOS DEL TUTOR/A:

Nombre y apellidos: MARÍA DOLORES MARTÍNEZ MIRANDA

Ámbito de conocimiento/Departamento: ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

Correo electrónico: mmiranda@ugr.es

3. COTUTOR/A DE LA UGR (en su caso):

Nombre y apellidos:

Ámbito de conocimiento/Departamento:

Correo electrónico:

4. COTUTOR/A EXTERNO/A (en su caso):

Nombre y apellidos:

Correo electrónico:

Nombre de la empresa o institución:

Dirección postal:

Puesto del tutor en la empresa o institución:

5. DATOS DEL ESTUDIANTE:

Nombre y apellidos: ALBERTO BOHORQUEZ GAVIRA

Correo electrónico: bgalberto@correo.ugr.es