APRENDIZAJE AUTOMÁTICO

Trabajo.3: Programación

Fecha límite de entrega: 27 de Mayo 2017 (Modelos Lineales) y 7 de junio (Modelos No-lineales)

Valoración: 35 puntos

NORMAS DE DESARROLLO Y ENTREGA DE TRABAJOS

Para este trabajo como para los demás es obligatorio presentar un informe escrito con sus valoraciones y decisiones adoptadas en el desarrollo de cada uno de los apartados. Incluir en el informe los gráficos generados. También deberá incluirse una valoración sobre la calidad de los resultados encontrados. (obligatorio en pdf). **Sin este informe se considera que el trabajo NO ha sido presentado.**

Normas para el desarrollo de los Trabajos: EL INCUMPLIMIENTO DE ESTAS NORMAS SIGNIFICA PERDIDA DE 2 PUNTOS POR CADA INCUMPLIMIENTO.

- El código se debe estructurar en un único script R con distintas funciones o apartados, uno por cada ejercicio/apartado de la práctica.
- Todos los resultados numéricos o gráficas serán mostrados por pantalla, parando la ejecución después de cada apartado. No escribir nada en el disco.
- El path que se use en la lectura de cualquier fichero auxiliar de datos debe ser siempre "datos/nombre_fichero". Es decir, crear un directorio llamado "datos"dentro del directorio donde se desarrolla y se ejecuta la práctica.
- Un código es apto para ser corregido si se puede ejecutar de principio a fin sin errores.
- NO ES VÁLIDO usar opciones en las entradas. Para ello fijar al comienzo los parámetros por defecto que considere que son los óptimos.
- El código debe estar obligatoriamente comentado explicando lo que realizan los distintos apartados y/o bloques.
- Poner puntos de parada para mostrar imágenes o datos por consola.
- ENTREGAR SOLO EL CODIGO FUENTE, NUNCA LOS DATOS.
- **Forma de entrega**: Subir los ficheros .R y .pdf ... según requerido en la definición de la entrega en el Tablón docente de CCIA. Y en un zip todos ficheros en TURNITIN.

Esta práctica se puede desarrollar en solitario o en colaboración con otro compañero.

1. Ajuste de Modelos Lineales: 15 puntos

Este ejercicio se centra en el ajuste de un modelo lineal a conjuntos de datos dadas con el objetivo de obtener el mejor predictor posible. En todos los casos los pasos a desarrollar serán aquellos que nos conduzcan al ajuste y selección del mejor modelo y a la estimación del error E_{out} del modelo final. Cómo mínimo se habrán de analizar y comentar los siguientes pasos **sobre un problema de clasificación y otro de regresión**:

1. Comprender al problema a resolver.
2. Los conjuntos de training, validación y test usados en su caso.
3. Preprocesado los datos: Falta de datos, categorización, normalización, reducción de dimensionalidad, etc.
4. Selección de clases de funciones a usar.
5. Discutir la necesidad de regularización y en su caso la función usada.
6. Definir los modelos a usar y estimar sus parámetros e hyperparámetros.
7. Selección y ajuste modelo final.
8. Estimacion del error E_{out} del modelo lo más ajustada posible.
9. Discutir y justificar la calidad del modelo encontrado y las razones por las que considera que dicho modelo es un buen ajuste que representa adecuadamente los datos muestrales.

Las clases de funciones de las transformaciones no lineales pueden definirse a partir de las potencios y los productos de potencias de las variables originales. Si se usan otras transformaciones de las variables iniciales, como $\log, \sqrt{()}$, sin, etc deberá justificarse el interés de dicha elección. **Bases de datos a usar (** `http://statweb.stanford.edu/~tibs/ElemStatLearn`**). Elegir una del bloque de regresión y otra del bloque de clasificación:**

- Clasificación:
  1. Email spam
  2. Digit Recognition
  3. South African Heart Disease
- Regresión:
  1. Los Angeles Ozone
  2. Marketing (Income Data)
  3. Prostata

Recomendación: desarrollar un código en R lo suficientemente general que permita ser reusado, en su mayor parte, en el desarrollo del siguiente ejercicio. Se recomienda escribir funciones que permitan ser reusadas.

2. Ajuste de Modelos No-lineales: 20 puntos

Este ejercicio se centra en el ajuste del mejor predictor (lineal o no-lineal) a un conjunto de datos. Debemos mostrar que los distintos algoritmos proponen soluciones para los datos pero que unas soluciones son mejores que otros para unos datos dados. El criterio que usaremos en la comparación será el error medio cuadrático para regresión, la curva ROC en clasificación binaria y el número de errores en clasificación multiclase.

Los modelos no-lineales a usar son:

- **Redes Neuronales**. Considerar tres clases de funciones definidas por arquitecturas con 1,2 y 3 capas de unidades ocultas y número de unidades por capa en el rango 0-50. Definir un conjunto de modelos(arquitecturas) y elegir el mejor por validación cruzada. Recordar que a igualdad de E_{out} siempre es preferible la arquitectura más pequeña.
- **Máquina de Soporte de Vectores (SVM)**: usar solo el núcleo RBF-Gaussiano. Encontrar el mejor valor para el parámetro libre hasta una precisión de 2 cifras (enteras o decimales)
- **Boosting**: Para clasificación usar AdaBoost con funciones "stamp". Para regresión usar árboles como regresores simples.
- **Random Forest**: Usar los valores que por defecto se dan en la teoría y experimentar para obtener el número de árboles adecuado.

Si no se comparan todos los modelos se deberá de justificar adecuadamente las razones que han conducido a los modelos elegidos para la comparación, así como para los no incluidos. Si se descarta a priori algún modelo por que no se considera adecuado para el problema hay que argumentar las razones de la decisión. **Para poder puntuar al menos deben comparar adecuadamente dos modelos**.

Se habrá de buscar el mejor modelo posible para la base de datos seleccionada y se habrá de justificar cada uno de los pasos dados para conseguirlo. Los puntos señalados en el ejercicio anterior pueden servir como guía.

Se usará para ello una de las siguientes BBDD del repositorio de la UCI (https://archive.ics.uci.edu/ml/).

Bases de datos elegibles:

1. Pen-Based Recognition of Handwritten digits (clasificación)
2. Page Blocks Classification
3. Amazon Commerce reviews set (clasificación)
4. Breast Cancer Wisconsin (Diagnostic) (clasificación)
5. Communities and Crime (regresión)
6. Parkinson Telemonitoring (regresión)
7. Housing (regresión)
8. Cardiotocography (clasificación)
9. Thyroid Disease (clasificación)
10. Occupancy detection (clasificación)
11. Default of Credit Card Clients (clasificación)
12. Internet Advertisements (clasificación)

13. Human Activity Recognition Using Smartphones (clasificación)
14. Image Segmentation (clasificación)
15. Mushroom (clasificaión)
16. Student Performance Data Set
17. Tennis Major Tournament Match Statistics Data Set