APRENDIZAJE AUTOMÁTICO

PROYECTO FINAL:

Fecha límite de entrega: 7 de junio 2019

Valoración: 25 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 python 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 todos los ficheros .py y .pdf en un zip a DECSAI.

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

1. Ajuste del mejor modelo

Este ejercicio se centra en el ajustar el 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 otras para unos datos dados. Se recomienda el uso de la librería scikit-learn en todas las fases del ajuste. Pero decisiones sin justificación y resultados sin interpretación no serán considerados válidos. Los valores fijados por defecto en la libreria no se consideran como justificados.

Es obligatorio considerar al menos uno de los siguientes modelos. Se recomienda el uso de las siguientes opciones (no son obligatorias):

- **Redes Neuronales**. Considerar tres clases de funciones definidas por arquitecturas con 0-3 capas de unidades ocultas y número de unidades por capa en el rango 0-100. 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 o el polinomial. 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 como hiperparámetros los valores que por defecto se dan en teoría y experimentar para obtener el número de árboles adecuado.

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. Pueden usarse técnicas de reducción de dimensionalidad, ej. PCA. Todos los proyectos deben justificar los siguientes apartados:

1. Definición del problema a resolver y enfoque elegido
2. Codificación de los datos de entrada para hacerlos útiles a los algoritmos.
3. Valoración del interés de la variables para el problema y selección de un subconjunto (en su caso).
4. Normalización de las variables (en su caso)
5. Justificación de la función de pérdida usada.
6. Selección de las técnica (parámetrica) y valoración de la idoneidad de la misma frente a otras alternativas
7. Aplicación de la técnica especificando claramente que algoritmos se usan en la estimación de los parámetros, los hiperparámetros y el error de generalización.
8. Argumentar sobre la idoneidad de la función regularización usada (en su caso)
9. Valoración de los resultados (gráficas, métricas de error, análisis de residuos, etc)
10. Justificar que se ha obtenido la mejor de las posibles soluciones con la técnica elegida y la muestra dada. Argumentar en términos de la dimensión VC del modelo, el error de generalización y las curvas de aprendizaje.

Puntuaciones

1. **Hasta 15 puntos**. Aquellos proyectos que seleccionen entre dos clases de modelos (funciones+algoritmo). Uno lineal y otro no lineal.
2. **Hasta 20 puntos**. Aquellos proyectos que seleccionen entre tres modelos (funciones+algoritmos). Uno lineal.
3. **Hasta 25 puntos**. Aquellos proyectos que seleccionen entre cuatro modelos (funciones+algoritmos). Uno lineal.

Se usará para ello BBDD del repositorio de la UCI (https://archive.ics.uci.edu/ml/). Se adjuntan algunos ejemplos, pero es posible elegir otras tras acordarlo con el profesor. Dentro de lo posible se intentará que cada proyecto trabaje con una base de datos distinta.

Bases de datos elegibles (entre otras):

1. Pen-Based Recognition of Handwritten digits (clasificación)
2. Page Blocks Classification (clasificación)
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 (clasificación)
16. Student Performance Data Set
17. Tennis Major Tournament Match Statistics Data Set
18. Arcene (clasificación)
19. APS Failure at Scania Trucks Data Set (clasificación)
20. Bank Marketing Data Set (clasificación)
21. Ionosfera (clasificación)
22. Diabetes (clasificación)

El uso de resultados y enfoques existentes en la literatura sobre las bases de datos está permitido y de hecho se alienta, siempre y cuando se deja manifiestamente claro que uso se hace de dicha información/resultado y cual es la aportación del proyecto sobre la misma. En caso contrario se entenderá plagio. Incluir las referencias de la bibliografía usada.