Proyecto AA

Alberto Díaz

Curso 2022-2023

Evaluación

- Ejercicios en cada tema
- 3 prácticas, aproximadamente
  - Es obligatorio entregar las prácticas

- La evaluación, tanto en la convocatoria ordinaria como en la extraordinaria, se basa en un trabajo, aplicando alguna(s) de las técnica(s) desarrolladas en las prácticas a un conjunto de datos de vuestra elección
  - Defensa del proyecto: 20% de la nota
  - Trabajo escrito sobre el proyecto: 80% de la nota

Proyecto

- El objetivo del proyecto es trabajar sobre un problema de clasificación con un conjunto de datos que hay que seleccionar en la plataforma Kaggle (https://www.kaggle.com/datasets)
  - Hay que enviar una propuesta al profesor de la asignatura (enviando un correo con un enlace al conjunto de datos propuesto) para que dé su visto bueno.
  - Has de utilizar el código de las prácticas que has hecho en la asignatura para desarrollar un sistema de aprendizaje automático sobre el conjunto de datos en cuestión.
  - Se debe dar una estimación de la efectividad del sistema, comparando distintas técnicas de aprendizaje automático estudiadas en la asignatura.
  - Hay que aplicar regresión logística, redes neuronales y árboles de decisión.
  - Adicionalmente, se pueden también aplicar implementaciones y técnicas no desarrolladas en la asignatura.
  - No puede haber dos proyectos que utilicen el mismo conjunto de datos.

Dataset

- Elegir dataset
  - Elegir tarea
    - tarea de clasificación para poder comparar
      - ¿Cuántas clases hay?
- Elegir columnas
  - Atributos
    - ¿cuántos hay?, ¿de qué tipo es cada uno?

Preprocesamiento de los datos

- Preprocesamiento datos, visualizar ayuda a decidir
  - (real_world_data)
    - Normalización
    - Bucketización
    - Categorización
      - One hot encoding
    - Outliers, valores raros
    - Missing values
    - Correlaciones

Subolecciones para experimentación

- Datos
  - Entrenamiento y validación
    - Para ajustar cada técnica de aprendizaje
  - Test
    - Comparación final entre las distintas técnicas ajustadas

- Subcolecciones con misma distribución de clases

Aplicar técnicas de clasificación

- Reg logística, redes neuronales, árboles de decisión

- Para cada técnica
  - Explicación adaptaciones realizadas en código prácticas
  - Explicación experimentos realizados para ajustar el modelo
  - Resultado final con test

- Métricas
  - Accuracy/Precision/Recall/F1

  - Python
    - classification_report
    - plot_confusion_matrix

- Tiempos

Memoria (notebook) y Presentación

- Título y autores (grupo y autores en los nombres de los ficheros)
- Índice, secciones con marcadores
- Presentación del dataset
  - ¿cuál es el problema?, ¿cuáles son los datos?
- Preprocesamiento
- Dataset final
- Separación en entrenamiento, validación, test
- Reg logística, redes neuronales, árboles de decisión (3 apartados)
  - Adaptación, ajuste, resultados
- Comparación
  - Comparar tiempos y resultados de las distintas técnicas, discutir cual es mejor y porqué.
- Conclusiones
- Bibliografía

Panda courses

- Kaggle courses
  - Python

  - Pandas
  - Data Cleaning
  - Data Visualization

  - Intro to Machine Learning
  - Intermediate Machine Learning