Proyecto DMAKit Versión 2

Contexto

Se trata de una actualización de la antigua plataforma previamente desarrollada. Se debe contemplar el análisis de conjuntos de datos desde los puntos de vista estadístico y descriptivo, además de facilitar la identificación de patrones y el desarrollo de modelos predictivos basándose en la aplicación de técnicas de machine learning.

Esta nueva versión será en formato escritorio, con el fin de facilitar el uso de recursos computacionales asociados al consumo de GPU y otros elementos de interés para poder trabajar con sistemas paralelos, y optimizar los cálculos.

Conjuntos de datos a trabajar

La nueva plataforma deberá permitir trabajar con conjuntos de datos de los siguientes tipos

1. **Vectorial**. Son los tipos de datos más simples, y representan archivos csv o tabulados. También se debe permitir el ingreso de archivos Excel para poder ser trabajados. Importante, al trabajar con archivos Excel se debe permitir seleccionar la hoja que desea trabajar.
2. **Grafos**. Son tipos de datos que pueden venir en formato pickle o similares. Además, también se pueden incorporar en formato de archivos csv, en donde se consideran matrices de adyacencia. De manera adicional, se puede tener archivos en formato csv que permitan caracterizar a los nodos y a las aristas. En el caso de entrenamiento de modelos, es posible recibir una carpeta con un conjunto de imágenes para el entrenamiento de modelos predictivos.
3. **Imágenes**. Se debe permitir el procesamiento de los datos en formato imagen o matriciales. En el caso de entrenamiento de modelos, es posible recibir una carpeta con un conjunto de imágenes para el entrenamiento de modelos predictivos.
4. **Series de tiempo o señales**. Son el mismo formato de los archivos Excel o csv.
5. **Texto**. Para trabajar con NLP.

Análisis y desarrollos en conjuntos de datos del tipo vectorial

Para los conjuntos de datos del tipo vectorial se debe efectuar los siguientes análisis:

1. Visualización de datos y resumen de información.
    a. Se debe permitir generar un resumen de la cantidad de datos en términos de filas y columnas
    b. Se debe generar un resumen de los tipos de datos
    c. Se debe hacer histogramas y gráficos de frecuencia dependiendo sea el tipo
2. Plots y visualización (usando plotly o highcharts, también seaborn)
    a. La idea es habilitar una herramienta de plots, contemplando los siguientes:
        i. Gráficos de barra
        ii. Gráficos de torta
        iii. Scatter plot
        iv. Log plots
        v. 3D plots
        vi. Line plots
        vii. Bubble chart

        viii. Parallel coordinates
        ix. Spiderweb
    b. Para todos los plots se debe permitir guardar las figuras
    c. También debe permitir agrupar o plotear con respecto a algo: Por ejemplo
        i. Scatter plot de A v/s B coloreado por columna C
3. Estadísticas descriptivas
    a. Se contempla una tabla resumen de los atributos continuos con los estadísticos básicos
    b. Se contempla plot estadísticos
        i. Histogramas
        ii. Boxplots
        iii. Heatmap
    c. Debe permitir hacer análisis de correlación:
        i. Correlaciones simples
        ii. Mutual information
4. Aplicación de test estadísticos
    a. Debe permitir aplicar test estadísticos para una muestra
        i. Evaluar la media
        ii. Evaluar la proporción
        iii. Evaluar la varianza
        iv. Evaluar si una distribución es normal o no aplicando diferentes pruebas
    b. Debe permitir aplicar test estadísticos para comparar dos muestras
        i. Comparar medias
        ii. Comparar proporciones
        iii. Comparar radios de varianzas
        iv. Test no paramétricos
    c. Debe permitir aplicar test estadísticos de comparación múltiple
        i. ANOVA
        ii. ANOVA no paramétrico
        iii. Análisis post hoc
5. Técnicas de procesamiento de datos
    a. Escalar conjuntos con variados escaladores (min max, max absolute, etc)
    b. Eliminar nulos
    c. Evaluar outliers
        i. Mediante IQR
        ii. Mediante z-score
        iii. Mediante métodos de scikit-learn (islation forest, etc)
    d. Técnicas de codificación y/o representación numérica
        i. Ordinal encoder
        ii. One Hot encoder
        iii. Frecuencias de categorías
        iv. Codificar por lenguaje natural (utilizando modelos pre-entrenados)
6. Entrenamiento de modelos de clasificación por machine learning clásico
    a. Estrategia de procesamiento
        i. División en entrenamiento y validación

        ii. Aplicar validación cruzada
        iii. Ambas
    b. Entrenamiento de modelos mediante algoritmos (ver los habilitados en scikit learn)
        i. Debe permitir seleccionar un algoritmo
        ii. Debe permitir seleccionar hiperparámetros por algoritmo
    c. Evaluar desempeño y calcular métricas
        i. Tabla resumen
        ii. Matrices de confusión
        iii. Curvas ROC
        iv. Plot de precisión v/s recall
    d. Modo exploración de hiper parámetros
        i. Debe seleccionar 1 algoritmo
        ii. Se deben explorar múltiples hiperparámetros en un formato grilla
        iii. Se debe correr el algoritmo y generar estadísticas resumen
        iv. Se debe tener una tabla con todos los resultados y una opción de detalle por cada ejecución para ver los resultados como si fueran ejecuciones individuales
    e. Modo exploración de algoritmos
        i. Similar al caso D pero también seleccionando múltiples algoritmos
        ii. Debe tener información por default
    f. Modo exploración mediante algoritmos genéticos
        i. La idea es armar la ejecución mediante un algoritmo genético (Ver tpot)
7. Entrenamiento de modelos de regresión por machine learning clásico
    a. Estrategia de procesamiento
        i. División en entrenamiento y validación
        ii. Aplicar validación cruzada
        iii. Ambas
    b. Entrenamiento de modelos mediante algoritmos (ver los habilitados en scikit learn)
        i. Debe permitir seleccionar un algoritmo
        ii. Debe permitir seleccionar hiperparámetros por algoritmo
    c. Evaluar desempeño y calcular métricas
        i. Tabla resumen
        ii. Scatter plot error
        iii. Scatter plot realidad v/s predicción
    d. Modo exploración de hiper parámetros
        i. Debe seleccionar 1 algoritmo
        ii. Se deben explorar múltiples hiperparámetros en un formato grilla
        iii. Se debe correr el algoritmo y generar estadísticas resumen
        iv. Se debe tener una tabla con todos los resultados y una opción de detalle por cada ejecución para ver los resultados como si fueran ejecuciones individuales
    e. Modo exploración de algoritmos
        i. Similar al caso D pero también seleccionando múltiples algoritmos
        ii. Debe tener información por default
    f. Modo exploración mediante algoritmos genéticos
        i. La idea es armar la ejecución mediante un algoritmo genético (Ver tpot)

8. Entrenamiento por métodos de regresión (complementario al punto 7)
    a. Estrategia de procesamiento
        i. División en entrenamiento y validación
        ii. Aplicar validación cruzada
        iii. Ambas
    b. Métodos de regresión
        i. Lineal
        ii. Variaciones de lineal
        iii. No lineal
        iv. Entre otros
    c. Evaluar desempeño y calcular métricas
        i. Tabla resumen
        ii. Scatter plot error
        iii. Scatter plot realidad v/s predicción
    d. Modo exploración de hiper parámetros
        i. Debe seleccionar 1 algoritmo
        ii. Se deben explorar múltiples hiperparámetros en un formato grilla
        iii. Se debe correr el algoritmo y generar estadísticas resumen
        iv. Se debe tener una tabla con todos los resultados y una opción de detalle por cada ejecución para ver los resultados como si fueran ejecuciones individuales
    e. Modo exploración de algoritmos
        i. Similar al caso D pero también seleccionando múltiples algoritmos
        ii. Debe tener información por default
9. Métodos de clustering por machine learning clásico
    a. Modo un algoritmo (ver los habilitados en scikit learn)
        i. Seleccionar algoritmo
        ii. Seleccionar hyperparámetros
        iii. Ejecutar
        iv. Obtener desempeño y mostrar un resumen
    b. Modo exploración de un algoritmo
        i. Seleccionar algoritmo
        ii. Armar grilla de exploración
        iii. Explorar las combinaciones
        iv. Generar una tabla resumen con un detalle por cada ejecución
    c. Modo exploración
        i. Similar al punto b pero se trabaja con múltiples algoritmos
10. Uso de modelos predictivos desarrollados
    a. Cada modelo de clasificación o regresión entrenado debe ser exportado
    b. Se debe permitir el uso de los modelos, esto implica, aplicar el mismo pipeline con el que se desarrolló (escalar, etc)
    c. Generar reporte de predicciones
11. Técnicas de transformación (ver los habilitados en scikit learn)
    a. PCA lineal
    b. Kernel PCA

    c. Non negative matrix factorization
    d. Factor Analysis
    e. Incremental PCA
12. Clustering por distancias:
    a. Calcular distancia seleccionando una opción de interés
    b. Armar distribución
    c. Armar matriz de adyacencia y generar el grafo
    d. Crear el clustering mediante comunidades
    e. Reportar mediante resumen los grupos, rendimiento, visualización.
13. Redes neuronales y Deep learning (para métodos de predicción)
    a. Habilitar back propagation (scikit learn)
    b. RNN arquitecturas (tensorflow)
    c. LSTM arquitecturas (tensorflow)
    d. CNN 1D
    e. CNN 2D

Análisis y desarrollos en conjuntos de datos tipo grafo

Estos serán los iniciales, puede que aparezcan más. No obstante, de momento son:

1. Clustering de comunidades
2. Descripción de grafos (centralidades, grados, etc)
3. GCN y GNN

Análisis y desarrollos en conjuntos de datos tipo imágenes

Para conjuntos de imágenes se debe trabajar en el desarrollo de modelos predictivos aplicando CNN 3D

Mientras que para una única imagen, se debe permitir hacer operaciones del tipo:

- Filtro
- Binarización
- Convolución

Análisis y desarrollos en conjuntos de datos del tipo series de tiempo

1. Se debe permitir plotear las series de tiempo
2. Clustering de series de tiempo (ver tslearn)
3. Modelos predictivos de series de tiempo (ver tslearn)
4. Modelos de pronóstico como autoregresión

Análisis y desarrollos en conjuntos de datos de texto

Se debe permitir

1. Modelos de sentimental análisis
2. Desarrollo de autoencoders mediante word2vec o doc2vec (gensim libray)