Introducción a la Ciencia de Datos (Optativa)
2025
Trabajo Práctico N.º 1

1. **Adquisición de datos:**

   Para este TP vamos a suponer que tenemos un instrumento que realiza una medición cada cierto tiempo. El archivo "data_to_analyze.txt" contiene un cierto número de mediciones realizadas que fueron almacenadas en forma de texto plano. Estos son datos "raw" o sin procesar y se requiere realizar un módulo de lectura de datos estos datos.

2. **Pre procesando y analizando datos**

   Los datos adquiridos en 1) requieren una exploración y limpieza previas. Para ello se pide realizar los módulos necesarios para:

   a. Determinar:
      i. Columnas que posee el set de datos.
      ii. Tipo de dato y dominio de cada columna.
   b. Determinar: máximo valor, mínimo valor, media, mediana, moda, cantidad de datos, desviación estándar.
   c. ¿Qué resolución poseen los datos? ¿Cuántos datos diarios se posee (en teoría)? ¿Existen períodos donde la resolución varía?
   d. Eventualmente estos datos se deben comparar con datos que tienen una resolución de 30 min. ¿Cómo podría solucionar el problema de diferentes resoluciones? Utilizar alguna estrategia adecuada.
   e. Determinar el porcentaje de filas faltantes (de acuerdo a la resolución) y de datos NULLs.
      i. Implementar un algoritmo que complete la serie de tiempo con las filas faltantes (queda a criterio la elección de la resolución de los datos si es que esta varía).
   f. Determinar el 1º, 2º y 3º cuartil. ¿Hay datos fuera de rango? Si es que los hay, identificarlos.
   g. Realizar un histograma de frecuencia de los datos.
      i. ¿Se puede decir que los datos siguen una distribución normal?
   h. Graficar los datos. Observando el gráfico, existe alguna regularidad en los datos (ciclos, tendencias).
   i. Generar 1 columna extra que posea el dato de la media de los 5 días pasados a cada dato. Graficar el dato original junto a los nuevos datos generados.
   j. Si tuviera que elegir un segmento de datos para generar un modelo con algún algoritmo ¿Qué segmento elegiría? Justificar.

Introducción a la Ciencia de Datos (Optativa)
2025
Trabajo Práctico N.º 1

3) **Análisis de componentes principales (PCA)**

a) Inventar un set de datos de por lo menos dos variables, y luego:

i) Graficar y analizar la correlación que puede existir entre los datos. Utilizar algún coeficiente de correlación (Pearson por ejemplo).

ii) ¿Cuál es la diferencia entre varianza y desviación estándar?

iii) Aplicar PCA:

(1) ¿Por qué es importante estandarizar las variables antes de aplicar PCA?

(2) ¿Qué representa un valor propio grande en el contexto de PCA?

(3) Explicar la relación entre autovectores y dirección de máxima varianza.

(4) Dado un conjunto de datos con “n” variables, ¿cuántos componentes principales como máximo puede tener PCA?

(5) ¿Qué dimensión tendrán los componentes principales?

(6) Definir la varianza explicada y la proporción de varianza explicada.

(7) ¿Cuáles y cuántos componentes principales seleccionaría para la obtención de la matriz proyección (nueva matriz de datos)? Mostrar la matriz proyección obtenida.

(8) En un conjunto de dos variables altamente correlacionadas, ¿Esperaría que la primera componente principal explique la mayor parte de la varianza? Explicar.

b) Comparación de regresión lineal con y sin PCA utilizando el dataset “California Housing”. Se dispone del dataset California Housing (`sklearn.datasets.fetch_california_housing`), que contiene variables predictoras sobre características de viviendas y una variable objetivo que representa el valor medio de la vivienda. Se desea comparar el desempeño de un modelo de regresión lineal utilizando:

i) Los datos originales completos (sin reducción de dimensionalidad).

ii) Los datos transformados mediante PCA, seleccionando el número mínimo de componentes que expliquen al menos el 95% de la varianza.

iii) Pasos sugeridos:

(1) Cargar y explorar el dataset: dimensiones, nombres de variables, estadísticas básicas.

(2) Dividir los datos en train y test (por ejemplo 80% / 20%).

(3) Estandarizar los datos.

(4) Aplicar regresión lineal sobre los datos originales y calcular métricas de desempeño (MSE).

(5) Aplicar PCA y proyectar los datos en un número reducido de componentes.

(6) Ajustar un modelo de regresión lineal sobre los datos proyectados y calcular métricas.

Introducción a la Ciencia de Datos (Optativa)
2025
Trabajo Práctico N.º 1

(7) Comparar los resultados y comentar cómo afecta la reducción de dimensionalidad al desempeño del modelo.
(8) Graficar la varianza explicada acumulada por los componentes principales.