Métodos Computacionales 2025
Segundo Trabajo Práctico: Optimización y Análisis de Datos

Presentación del problema

Este trabajo práctico busca integrar los conceptos de **ortogonalidad**, **formas cuadráticas**, **descomposición en valores singulares (SVD)** y **métodos de optimización**, aplicándolos al análisis y ajuste de datos. A lo largo del recorrido, se espera que puedan reconocer cómo estas herramientas se relacionan entre sí dentro de un mismo marco conceptual: el de la modelización de datos y la búsqueda de soluciones óptimas.

Se espera que, al finalizar el trabajo, los estudiantes puedan:

1. Modelar un problema de datos utilizando representaciones matriciales.
2. Resolver sistemas sobredeterminados mediante distintos enfoques (ecuaciones normales, SVD, métodos iterativos).
3. Evaluar el impacto del condicionamiento y la ortogonalidad en la estabilidad de las soluciones.
4. Interpretar las soluciones de optimización desde una perspectiva geométrica y numérica.
5. Aplicar los conocimientos a un caso real, justificando las decisiones tomadas y analizando la calidad del modelo obtenido.

1. Modelado y cuadrados mínimos

Se dispone del conjunto de datos:

$$(x_i, y_i) = \{(0,1), (1,2), (2,2.8), (3,3.6), (4,4.5)\}.$$

Queremos ajustar un modelo cuadrático de la forma

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

usando el método de los mínimos cuadrados.

a) Escribir el sistema normal $A^T A\beta = A^T y$.

b) Crear una función en python que resuelva el sistema.

c) Graficar los datos y la curva ajustada.

d) Calcular el error cuadrático medio (MSE).

2. Interpretación mediante SVD

Antes de comenzar, recordemos que un sistema se dice **mal condicionado** cuando pequeñas perturbaciones en los datos (por ejemplo, en A o y) provocan grandes variaciones en la solución. Este fenómeno suele deberse a la colinealidad entre columnas de la matriz de diseño y puede analizarse a través de los valores singulares obtenidos por la SVD.

a) Calcular la descomposición SVD de la matriz de diseño A.

b) Verificar numéricamente que las columnas de U son ortogonales.

c) Comparar la solución obtenida por SVD con la obtenida por ecuaciones normales.

d) Analizar el efecto del mal condicionamiento y la colinealidad entre columnas.

3. Análisis cuadrático

Sea la forma cuadrática asociada al error:

$$Q(\beta) = \|A\beta - y\|^2 = \beta^T(A^TA)\beta - 2(A^Ty)^T\beta + y^Ty.$$

a) Graficar $Q(\beta_1, \beta_2)$ para un valor fijo de β_0, señalando el punto mínimo.

b) Mostrar que $Q(\beta)$ es convexa y que su mínimo se alcanza en la solución de mínimos cuadrados.

4. Optimización numérica

a) Implementar el método de descenso por gradiente para minimizar $Q(\beta)$.

b) Experimentar con distintas tasas de aprendizaje y tolerancias.

c) (Opcional) Verificar que ambos métodos convergen al mismo mínimo que el obtenido por SVD.

5. Discusión y extensiones

a) Discutir la relación entre los distintos métodos.

b) Analizar cómo cambia el ajuste si se agregan términos cúbicos o ruido a los datos.

6. Aplicación a un caso real

En esta sección se propone aplicar los conceptos desarrollados a un conjunto de datos reales. El objetivo es construir un modelo predictivo que relacione una variable de salida y con un conjunto de variables explicativas $x_1, x_2, \ldots, x_n$, utilizando las herramientas de ajuste, análisis matricial y optimización vistas en el trabajo.

Se espera que los grupos puedan:

- Formular el problema en términos matriciales, identificando claramente la matriz de diseño y la variable objetivo.
- Justificar las transformaciones o selecciones de variables que consideren necesarias (normalización, codificación de variables categóricas, etc.).

- Aplicar métodos de resolución adecuados (ecuaciones normales, SVD, descenso por gradiente, entre otros) y explicar por qué los eligieron.
- Evaluar la calidad del modelo obtenido mediante métricas apropiadas (MSE, R^2, precisión, etc.) y discutir sus resultados.
- Interpretar los parámetros del modelo y reflexionar sobre la validez del ajuste en el contexto del conjunto de datos.

Cada grupo podrá elegir **uno** de los siguientes conjuntos de datos, incluidos en el archivo `datasets.zip`:

Stroke Prediction Dataset

Este conjunto de datos busca predecir si un paciente es propenso a sufrir un accidente cerebrovascular (stroke) en función de variables demográficas y médicas.

Atributos:

- `Id`: identificador único del paciente.
- `Gender`: "Male", "Female" o "Other".
- `Age`: edad del paciente.
- `Hypertension`: 0 si no tiene hipertensión, 1 si la tiene.
- `Heart_disease`: 0 si no tiene enfermedades cardíacas, 1 si las tiene.
- `Ever_married`: "Yes" o "No".
- `Work_type`: tipo de ocupación ("children", "Govt_job", "Private", etc.).
- `Residence_type`: "Rural" o "Urban".
- `Avg_glucose_level`: nivel promedio de glucosa en sangre.
- `Bmi`: índice de masa corporal.
- `Smoking_status`: "formerly smoked", "never smoked", "smokes" o "Unknown".
- `Stroke`: variable objetivo — 1 si tuvo un accidente cerebrovascular, 0 si no.

Traffic Prediction Dataset

Este conjunto contiene observaciones horarias del número de vehículos que pasan por cuatro intersecciones diferentes de una ciudad.

Atributos:

- `DateTime`: fecha y hora de la observación.
- `Junction`: identificador de la intersección.
- `Vehicles`: cantidad de vehículos registrados en esa hora (variable objetivo).
- `Id`: identificador único del registro.

El dataset contiene aproximadamente 48.000 observaciones. Los sensores de cada intersección recolectaron datos en períodos distintos, por lo que puede haber diferencias temporales o registros incompletos. (Lo ideal sería deshacerse de los registros incompletos)

Salary Prediction Dataset

Este conjunto de datos busca predecir el salario anual de empleados en función de sus características personales y laborales.

Atributos:

- `Age`: edad del empleado (numérica).
- `Gender`: "Male" o "Female".
- `Education_level`: nivel educativo ("High_School", "Bachelor", "Master", "PhD").
- `Job_title`: cargo o puesto de trabajo.
- `Years_of_experience`: años de experiencia laboral.
- `Salary`: salario anual (variable objetivo).

Earthquake Alert Prediction Dataset

Este conjunto está orientado a la detección temprana de alertas sísmicas. Incluye registros históricos con características geográficas y de vibración para predecir si se emitirá una alerta.

Atributos:

- `Id`: identificador único del registro.
- `Latitude`, `Longitude`: coordenadas geográficas del sensor.
- `Depth`: profundidad estimada del evento.
- `Magnitude`: magnitud del evento sísmico.
- `Station_signal`: señal registrada en la estación (numérica).
- `Alert`: variable objetivo — 1 si se emitió alerta, 0 si no.

7. Condiciones de entrega

Fecha límite de entrega: 23 de noviembre a las 23:59
Modalidad: Entrega grupal (hasta tres integrantes) vía campus.

El trabajo debe incluir:

1. **Informe**: debe presentarse en formato `.pdf`, tienen que señalar dataset utilizado e incluir gráficos, análisis y explicaciones de experimentaciones o teoremas utilizados.
2. **Código**: debe entregarse en formato `.ipynb` (notebook de Python) ejecutable y sin errores.

Aclaración

No se permite utilizar funciones de librerías que resuelvan directamente las consignas, como por ejemplo: `numpy.linalg.solve` o `numpy.linalg.lstsq`.