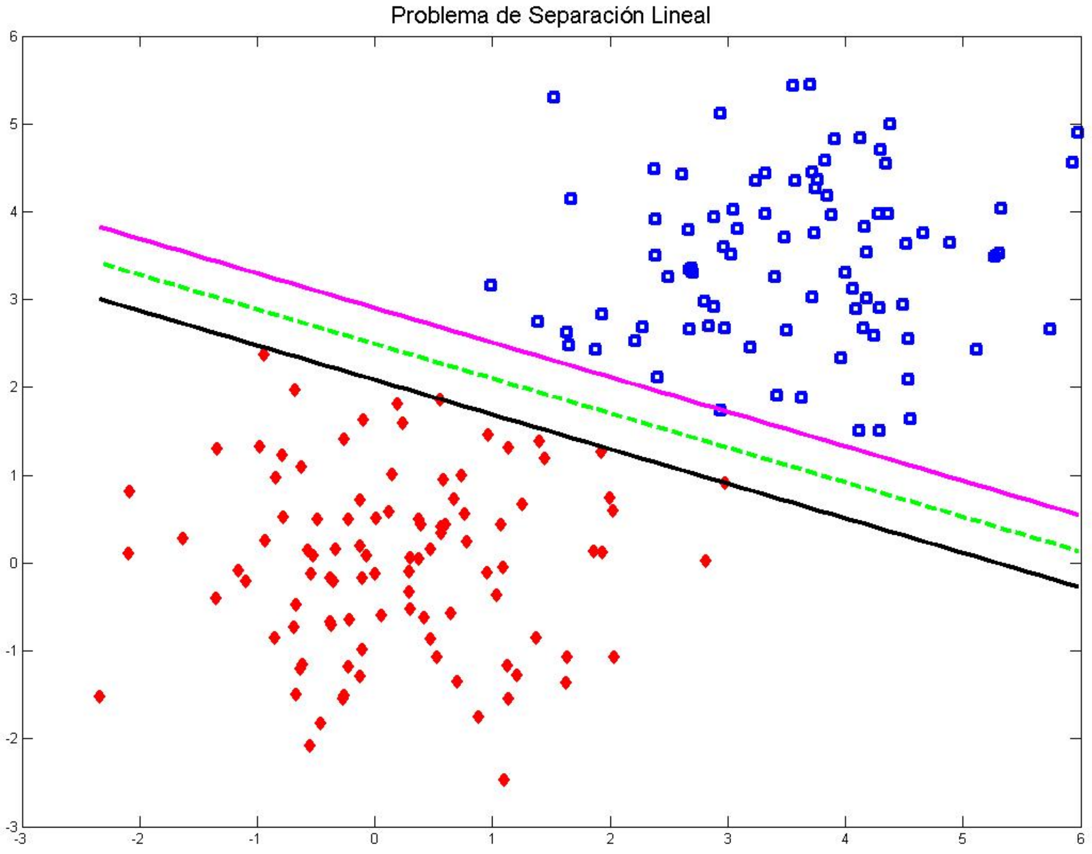
Proyecto I
Optimización Numérica
Programación Cuadrática
Clasificación con Separadores Lineales



1 Introducción

Definición. Decimos que los conjuntos

$$\mathbb{A} = \{x_1,\ x_2,\ \ldots,\ x_s\},\ \ \mathbb{B} = \{x_{s+1},\ x_{s+2},\ \ldots,\ x_{s+r}\} \subset \mathbb{R}^n,$$

son **linealemente separables** si y sólo si existe un hiperplano,

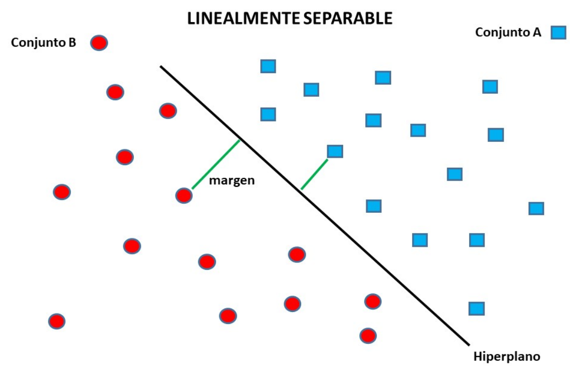
$$\mathcal{H}(w,\ \beta) = \{x \in \mathbb{R}^n \mid w^T x + \beta = 0\},$$

tal que

$$w^T x_i + \beta < 0, \quad \text{para todo} \ \ x_i \in \mathbb{A}$$

y

$$w^T x_i + \beta > 0, \quad \text{para todo} \ \ x_i \in \mathbb{B}.$$



En diversas aplicaciones se tienen dos conjuntos finitos de vectores como $\mathbb{A}$ y $\mathbb{B}$ que pueden indicar correo spam o no, o puntos en el cuerpo humano indicando células sanas o enfermas. Se necesita conocer si ambos conjuntos son linealmente separables o no. En caso de ser separables, se tiene una clasificación de los datos con el hiperplano, $\mathcal{H}(w,\ \beta)$ y al llegar un nuevo vector, y, se determina en la zona donde estará el nuevo vector, por ejemplo, se clasifica si el nuevo correo electrónico es spam o no.

2 Formulación del Modelo

Sean

$$\mathbb{A} = \{x_1,\ x_2,\ \ldots,\ x_s\},\ \ \text{y}\ \ \mathbb{B} = \{x_{s+1},\ x_{s+2},\ \ldots,\ x_{s+r}\} \subset \mathbb{R}^n,$$

para los cuales se debe determinar un hiperplano,

$$\mathcal{H}(w,\ \beta) = \{x \in \mathbb{R}^n \mid w^T x + \beta = 0\},$$

tal que

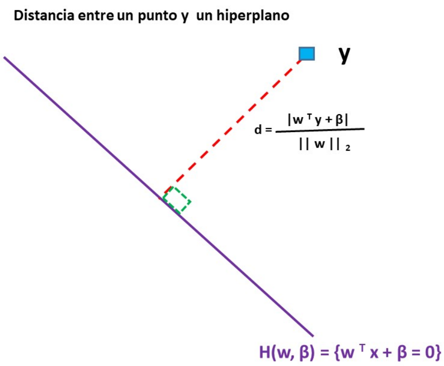
$$w^T x_i + \beta < 0, \quad \text{para todo} \ \ x_i \in \mathbb{A}$$

y

$$w^T x_i + \beta > 0, \quad \text{para todo} \ \ x_i \in \mathbb{B}.$$

Lema. Sea $\mathcal{H}(w,\ \beta) = \{x \in \mathbb{R}^n \mid w^T x + \beta = 0\}$, donde $w \neq 0$ y $\beta \in \mathbb{R}$. La distancia de y al hiperplano $\mathcal{H}(w,';\beta)$, es

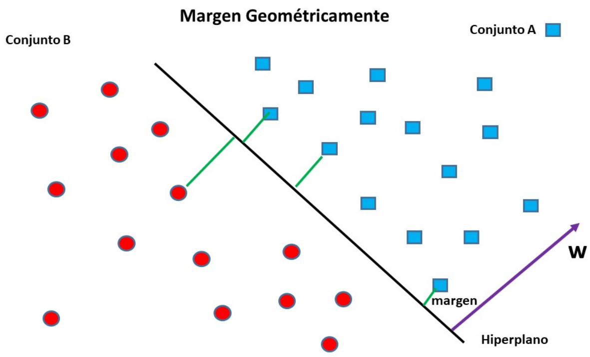
$$d(y,\ \mathcal{H}(w,\ \beta)) = \frac{|w^T y + \beta|}{\|w\|_2}.$$



El margen de los conjuntos $\mathbb{A}$, $\mathbb{B}$ y $\mathcal{H}(w,\ \beta)$ se define como

$$\textbf{Min} \quad \tfrac{|w^T\ x_i + \beta|}{\|w\|_2}$$

$$x_i \in \mathbb{A} \cup \mathbb{B},$$



El problema sería

$$\begin{array}{ll}\textbf{Max} & \left(\text{Min}\left\{\frac{|w^T\ x_i+\beta|}{\|w\|_2}\mid x_i\in\mathbb{A}\cup\mathbb{B},\right\}\right)\\ (w,\ \beta) & \\ \textbf{Sujeto a} & w^Tx_i+\beta<0,\ \ x_i\in\mathbb{A}\\ & w^Tx_i+\beta>0,\ \ x_i\in\mathbb{B}\end{array}$$

Por escalamiento, podemos pedir que

$$\textbf{Min}\{|w^T\ x_i+\beta|\ \mid\ x_i\in\mathbb{A}\cup\mathbb{B}\}=1$$

El problema se convierte

$$\begin{array}{ll}\textbf{Max} & \frac{1}{\|w\|_2}\\ (w,\ \beta) & \\ \textbf{Sujeto a} & \min\{|w^T\ x_i+\beta|\ \mid\ x_i\in\mathbb{A}\cup\mathbb{B}\}=1\\ & w^Tx_i+\beta<0,\ \ x_i\in\mathbb{A}\\ & w^Tx_i+\beta>0,\ \ x_i\in\mathbb{B}\end{array}$$

por la primer restricción de desigualdad se obtiene

$$\begin{array}{ll}\textbf{Max} & \frac{1}{\|w\|_2^2}\\ (w,\ \beta) & \\ \textbf{Sujeto a} & \textbf{Min}\{|w^T\ x_i+\beta|\ \ x_i\in\mathbb{A}\cup\mathbb{B}\}=1\\ & w^Tx_i+\beta\leq-1,\ \ x_i\in\mathbb{A}\\ & w^Tx_i+\beta\geq1,\ \ x_i\in\mathbb{B}\end{array}$$

La formulación final del problema es

$$
\begin{array}{ll}
\textbf{Min} & (1/2)w^T w \\
(w,\ \beta) & \\
\textbf{Sujeto a} & w^T x_i + \beta \leq -1,\ \ i = 1,\ 2,\ \ldots,\ s \\
& w^T x_i + \beta \geq 1,\ \ i = s+1,\ s+2,\ \ldots,\ s+r,
\end{array}
$$

con las variables $w \in \mathbb{R}^n,\ \beta \in \mathbb{R}$.

3 Proyecto

Escriba en **Python** un método de puntos interiores para el problema cuadrático :

$$
\begin{array}{ll}
\textbf{Min} & \mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{c}^T\mathbf{x} \\
\textbf{Sujeto a} & \mathbf{F}\mathbf{x} \geq \mathbf{d},
\end{array} \tag{1}
$$

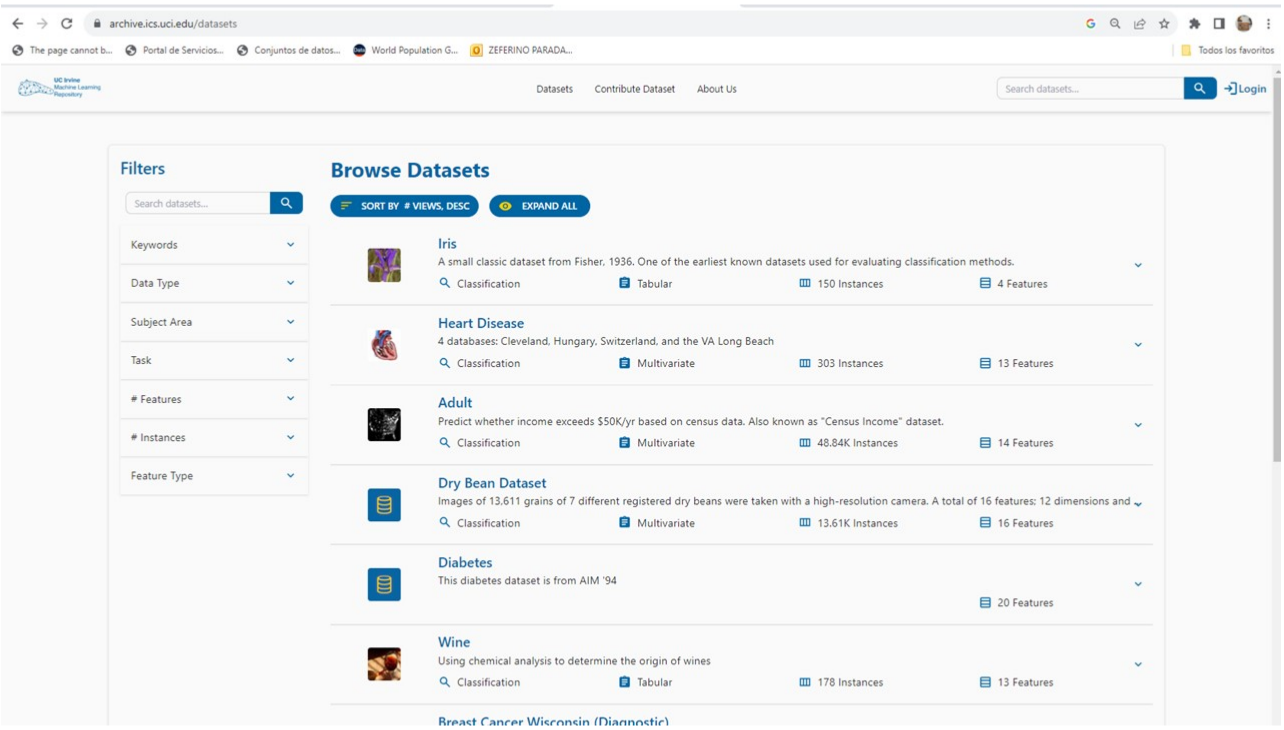
con $\mathbf{Q}$ matriz simétrica positiva definida.

def **qp_intpoint**($\mathbf{Q}$, $\mathbf{F}$, $\mathbf{c}$, $\mathbf{d}$)
$\mathbf{x}$ solución óptima del problema.
Min $(1/2) * x^T Q x + c^T x$
Sujeto a $Fx \geq d$
#
La tolerancia para las condiciones necesarias de primer orden es
$tol = 1e-05$.
Número máximo de iteraciones, **maxiter** $= 100$.
⋮
Su programa
⋮
return x, μ, $iter$
x vector solución del problema
μ multiplicador de Lagrange
$iter$ número de iteraciones

4 Datos

Usaremos los datos de la Universidad de California en Irvine:

https://archive.ics.uci.edu/datasets



Los datos son de:

1. Iris.
2. Wine.
3. Breast cancer.

Los datos pueden obtenerlos en formato de **Python**. Luego tiene que analizar los datos y separar los conjuntos $\mathcal{A}$ y $\mathcal{B}$.

5 Qué Entregar

Equipos de a lo más cuatro personas.

1. Programa: **qp_intpoint.py**.
2. Archivo, *separacion.py*, donde obtenga los valores de w y β en la pantalla y la graficaciones de: $\mathbb{A}^T w - e$ y $\mathbb{B}^T w - e$. Estas gráficas muestran la separación de los conjuntos con respecto al signo, positivo o negativo, que deben tener los puntos con respecto la hiperplano.
3. La base de datos que usaron.
4. Programas con los nombres y clave única de los integrantes del equipo y documentación de los códigos.
5. Entrega por Canvas, **miércoles 3 de abril de 2024 a las 18:00 horas**. Una entrega por cada equipo.
6. Presentación: **miércoles 3 de abril de 2024 a las 18:00 horas**.

6 Calificación

Concepto	**Comentarios y Sugerencias**	**Ponderaión**	**Calificación**
Parte Teórica		**20**	
Programación		**30**	
Resultado Numérico		**30**	
Graficación		**10**	
Presentación		**10**	
TOTAL de puntos		**100**	