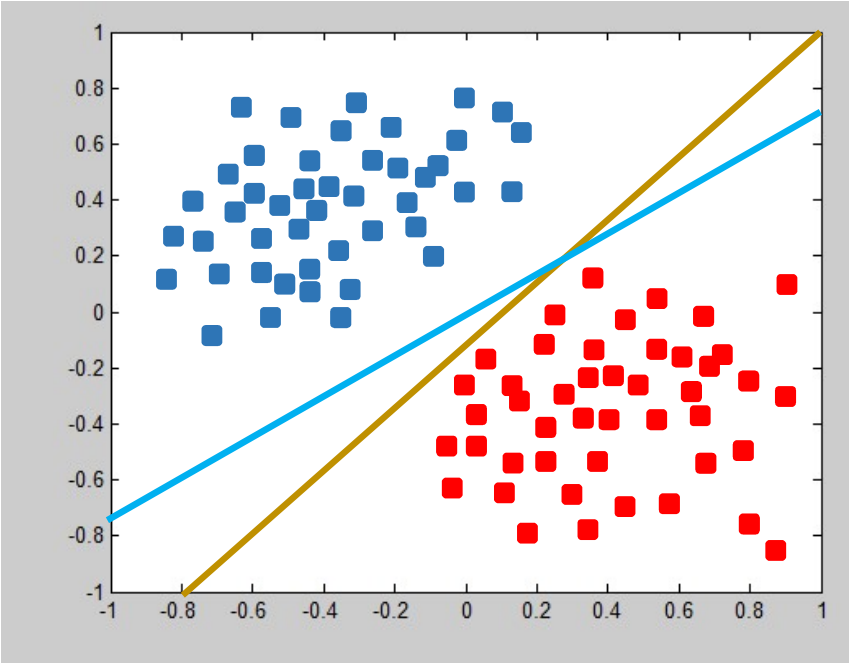
Máquinas de Soporte Vectorial
SVM (Support Vector Machine)

(Caso Linealmente Separable)

Aprendizaje Automático

Dr. Luis Eduardo Falcón Morales
ITESM
Campus Guadalajara



Dada dos clases linealmente separables como se muestra en la imagen, existirá una infinidad de rectas (hiperplanos) que separen a ambas clases.

¿Cuál podríamos considerar como "**la mejor**" de todas estas rectas de separación?

Aplicaciones:

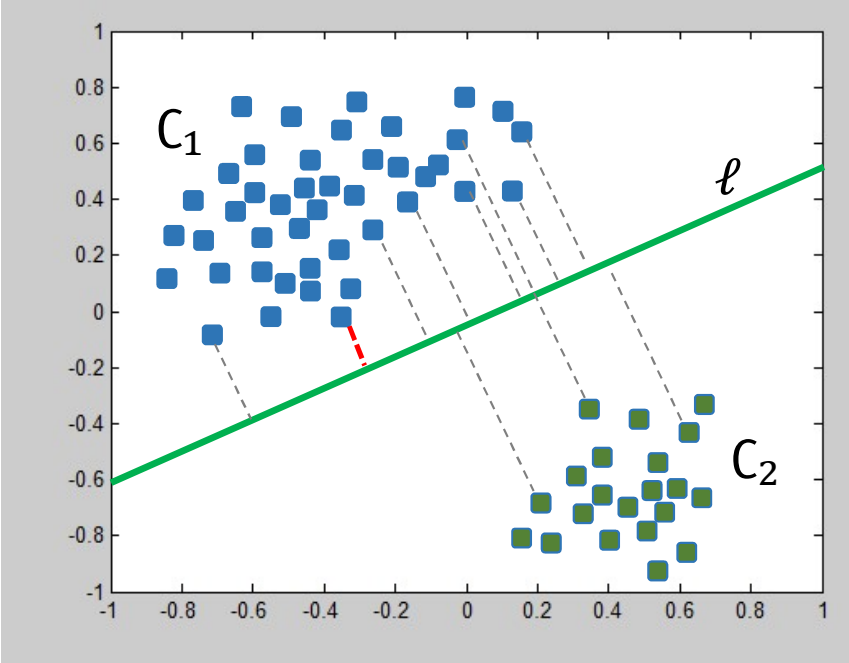
- Variable Dependiente Categórica (Clasificación)
- Variable Dependiente Numérica (Predicción)

Usualmente las Máquinas de Soporte Vectorial (SVM) involucra los siguientes tres conceptos:

- Clasificador de Margen Máximo (o Hiperplano de Margen Máximo)
- Clasificador de Vectores de Soporte
- Máquinas de Vectores de Soporte

En ocasiones estos términos se manejan bajo el mismo concepto de SVM.

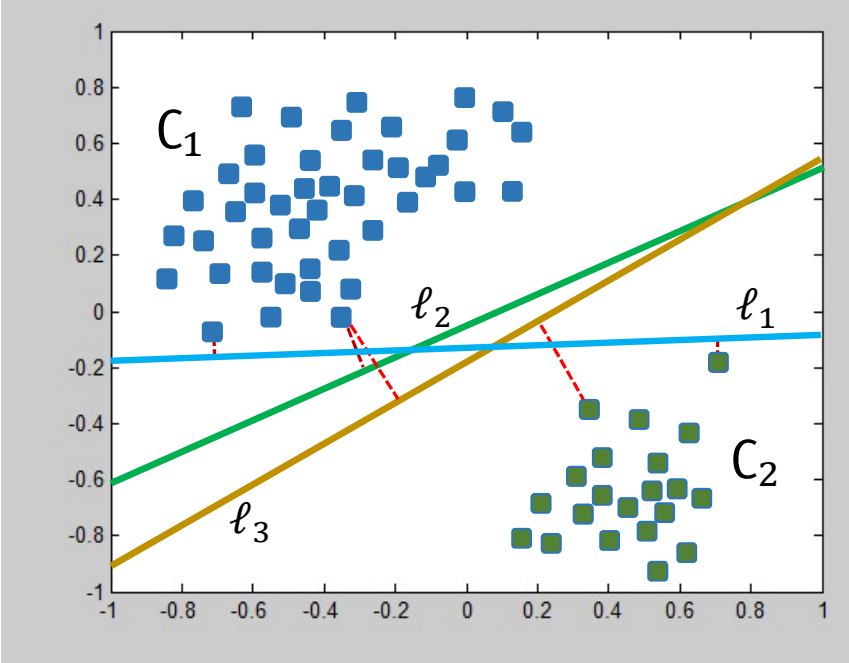
Hiperplano de Separación



Dado un hiperplano ℓ que separa dos clases C_1 y C_2, definimos su **margen** como la mínima de las distancias de ℓ a cualquiera de los puntos $x \in \Omega = C_1 \cup C_2$, es decir, si $d(\ell, x)$ es la distancia (ortogonal) del hiperplano ℓ al punto x, entonces:

$$\textbf{margen} = \min_{x \in \Omega} d(\ell, x)$$

En la imagen de la izquierda, de todas las distancias de ℓ a los puntos de $\Omega = C_1 \cup C_2$, la que se muestra con la línea punteada en rojo es la mínima, es decir, dicho valor es el margen de ℓ.



El hiperplano ℓ_3 mostrado en la imagen es el que tiene el margen con el mayor valor.

Definimos el **Hiperplano de Margen Máximo (HMM)**, al hiperplano separador de dos clases con el máximo de los márgenes.

Observa que si ℓ es un HMM de dos clases C_1 y C_2, entonces existe al menos un punto de cada clase con el mayor de los márgenes, sobre el conjunto de todos los hiperplanos separadores.

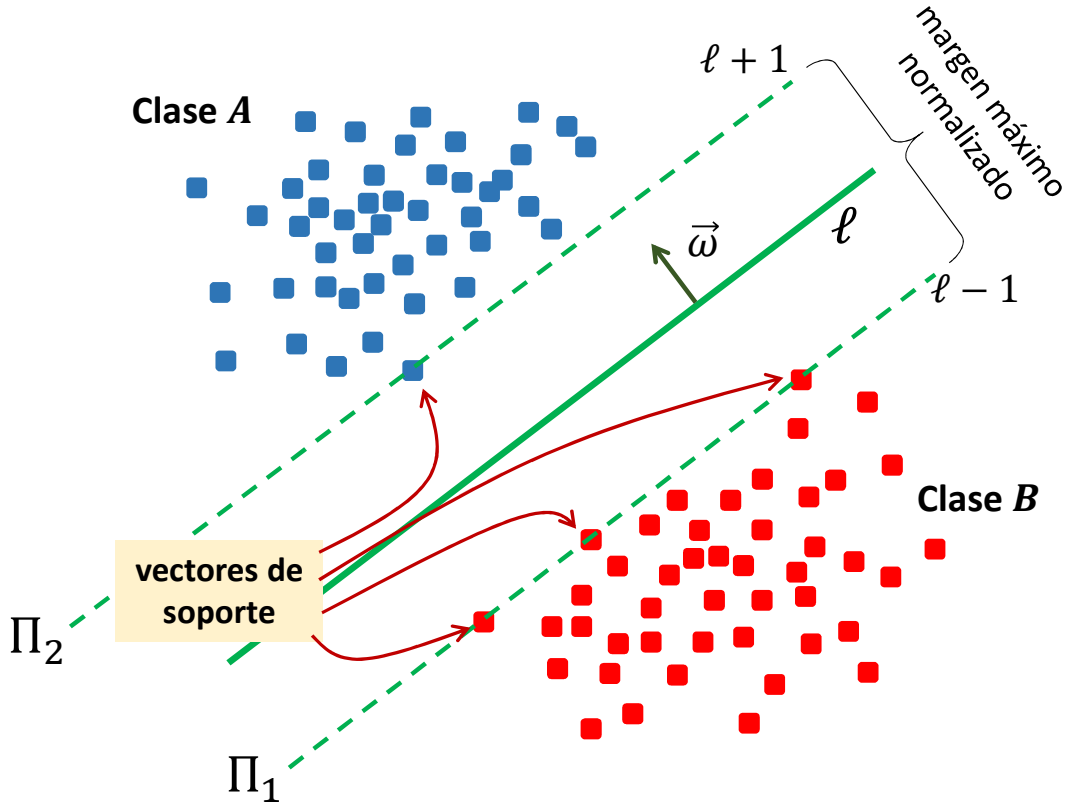
Es decir, el hiperplano de margen máximo se encuentra a la misma distancia de ambas clases.

Los puntos de cada clase que definen el hiperplano de margen máximo se llamarán **vectores de soporte**.

Hiperplano de Margen Máximo (HMM)

$$a'x + b'y + 1 = 0$$

Así, el algoritmo HMM buscará separar ambas clases por un hiperplano de manera que dicho hiperplano se encuentre a la misma distancia de ambas clases.



Sin pérdida de generalidad podemos suponer que los hiperplanos Π_1 y Π_2 que contienen a los vectores de soporte están trasladados una unidad del hiperplano ℓ, HMM, como se muestra en la figura.

Observa además que el hiperplano HMM es aquel cuya distancia entre los hiperplanos Π_1 y Π_2 que contienen a los vectores de soporte es máxima.

Hiperplano o Clasificador de Margen Máximo

Caso: Clases Linealmente Separables

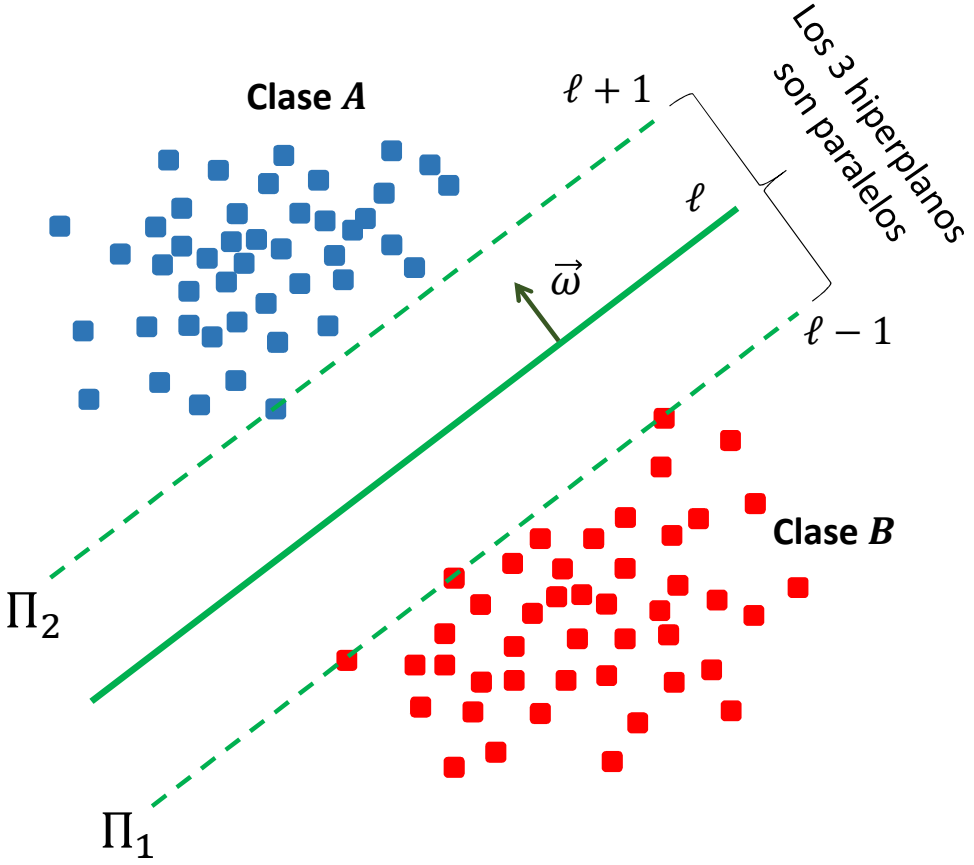
Datos de entrada:

Supongamos que tenemos un conjunto de n puntos m dimensional $\vec{x}_k = (x_{k1}, x_{k2}, \dots, x_{km}) \in \mathbb{R}^m$, $k = 1,2, \dots, n$ los cuales pertenecen a dos clases diferentes linealmente separables.

Denotemos como A y B a dichas clases.

Supongamos que cada punto $\vec{x}_k$ está asociado a una etiqueta $y_k \in \{-1, +1\}$, para $k = 1, 2, \dots, n$. Por ejemplo, para $k = 1, 2, \dots, n$

$$y_k = \begin{cases} +1 & si \quad \vec{x}_k \in Clase\ A \\ -1 & si \quad \vec{x}_k \in Clase\ B \end{cases}$$



$$\ell : \vec{\omega} \cdot \vec{x} + b = 0$$
$$\Pi_1 : \vec{\omega} \cdot \vec{x} + b = -1$$
$$\Pi_2 : \vec{\omega} \cdot \vec{x} + b = +1$$

El objetivo es encontrar los coeficientes $\omega_j,\ j = 0, 1, 2, \ldots, m$ del hiperplano HMM que separe ambas clases, es decir,

$$\ell : \omega_1 x_1 + \omega_2 x_2 + \cdots + \omega_m x_m + \omega_0 = 0$$

o bien,

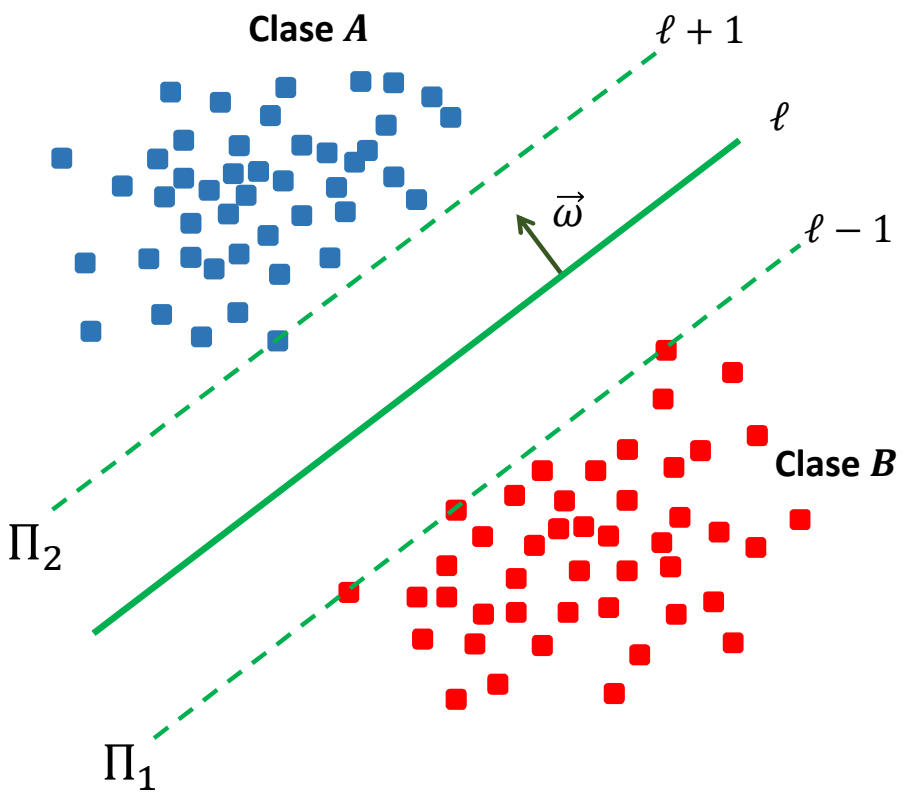
$$\ell : \vec{\omega} \cdot \vec{x} + b = 0$$

donde $\vec{\omega} = (\omega_1, \omega_2, \ldots, \omega_m),\ b = \omega_0,$ $\vec{x} = (x_1, x_2, \ldots, x_m).$

Y de la figura de la izquierda tenemos entonces que:

$$\vec{\omega} \cdot \vec{x} + b \geq +1, \qquad \forall \vec{x} \in A$$

$$\vec{\omega} \cdot \vec{x} + b \leq -1, \qquad \forall \vec{x} \in B$$



$$\ell : \ \vec{\omega} \cdot \vec{x} + b = 0$$
$$\Pi_1 : \ \vec{\omega} \cdot \vec{x} + b = -1$$
$$\Pi_2 : \ \vec{\omega} \cdot \vec{x} + b = +1$$

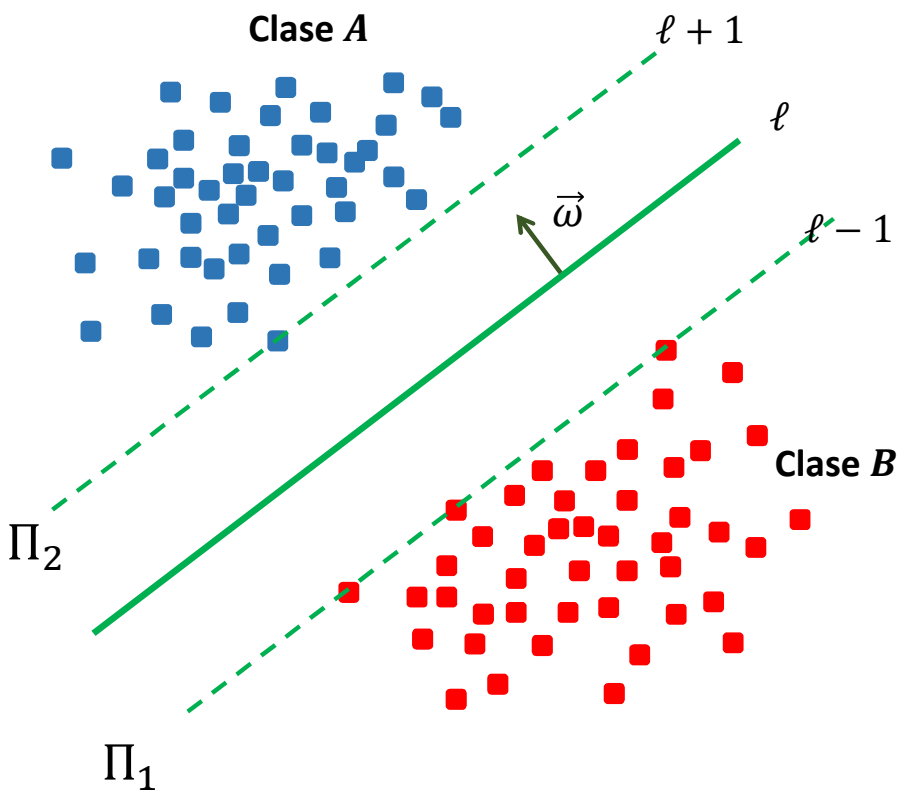
O de manera equivalente podemos decir que

$$y_k(\vec{\omega} \cdot \vec{x}_k + b) \geq +1,$$

para $k = 1,2, \ldots, n$, donde

$$y_k = \begin{cases} +1 & si \quad \vec{x}_k \in Clase\ A \\ -1 & si \quad \vec{x}_k \in Clase\ B \end{cases}$$

Donde la clase A es la clase positiva definida por el vector $\vec{\omega}$.



$$\ell : \ \vec{\omega} \cdot \vec{x} + b = 0$$
$$\Pi_1 : \ \vec{\omega} \cdot \vec{x} + b = -1$$
$$\Pi_2 : \ \vec{\omega} \cdot \vec{x} + b = +1$$

Por otro lado, recordemos que la distancia D de un hiperplano $\vec{\omega} \cdot \vec{x} + b = 0$ al origen está dada como:

$$D = \frac{|b|}{\|\vec{\omega}\|}$$

Entonces, si suponemos que el hiperplano Π_1 pasa por el origen, la distancia entre Π_1 y Π_2 que se desea maximizar está dada como (ver Figura):

$$D = \frac{2}{\|\vec{\omega}\|}$$

Así, maximizar la distancia $\frac{2}{\|\vec{\omega}\|}$ será equivalente a minimizar la magnitud $\|\vec{\omega}\|$, o bien $\frac{1}{2}\|\vec{\omega}\|^2$.

En resumen, para obtener el HMM se desea resolver el siguiente problema de optimización cuadrática con la función de costo $J = \frac{1}{2}\|\vec{\omega}\|^2$:

$$\min_{\vec{\omega},b} J(\vec{\omega}) = \frac{1}{2}\|\vec{\omega}\|^2$$

sujeto a las restricciones

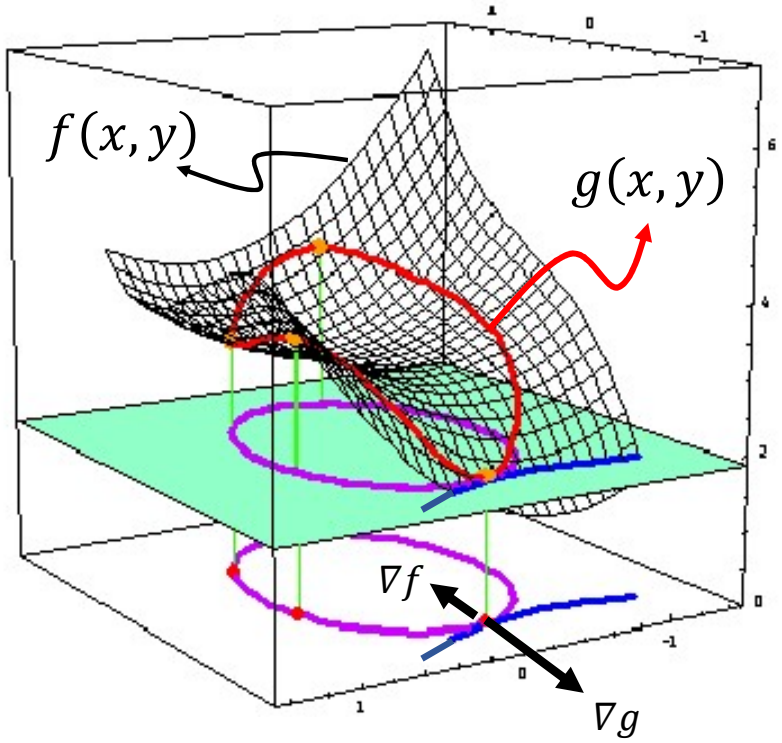
$$y_k(\vec{\omega} \cdot \vec{x}_k + b) \geq +1,$$

donde $k = 1,2,\ldots,n,\ \vec{\omega} = (\omega_1, \omega_2, \ldots, \omega_m),$

$$y_k = \begin{cases} +1 & si \quad \vec{x}_k \in A \\ -1 & si \quad \vec{x}_k \in B \end{cases}$$

De la función de costo $J = \frac{1}{2}\|\vec{\omega}\|^2$ obtenemos que el gradiente $\nabla J = \vec{\omega}$; que la matriz Hessiana $\nabla^2 J = I_{m\times m}$, es decir, es la matriz identidad y por lo tanto es definida positiva (es decir, sus eigenvalores son positivos). Lo anterior implica que la función de costo es convexa y como las restricciones son afinas, existirá una solución única para el problema de optimización por el Teorema de Karush-Kuhn-Tucker (KKT).

Interpretación Geométrica de los Multiplicadores de Lagrange



Dada la función $z = f(x, y)$,

sus máximos y mínimos y que además

satisfacen la restricción $g(x, y) = 0$

cumplen la siguiente igualdad:

$$\nabla f(x, y) = \lambda \nabla g(x, y)$$

Suponiendo que ambas funciones son diferenciables.

Multiplicadores de Lagrange

Este método se aplica cuando se desean encontrar los máximos y los mínimos de una función f sujeta a una o varias restricciones.

Por ejemplo, en particular si se desea maximizar la función $w = f(x, y)$ sujeta a dos restricciones $g(x, y) = 0$, $h(x, y) = 0$, primero se forma la restricción del Lagrangiano:

$$F(x, y, \lambda_1, \lambda_2) = f(x, y) - \lambda_1 g(x, y) - \lambda_2 h(x, y)$$

y después se buscan los puntos críticos de f. Los puntos críticos de f se encuentran mediante la resolución del sistema de ecuaciones formado por las derivadas parciales de F:

$$F_x(x, y, \lambda_1, \lambda_2) = 0$$
$$F_y(x, y, \lambda_1, \lambda_2) = 0$$
$$F_{\lambda_1}(x, y, \lambda_1, \lambda_2) = 0$$
$$F_{\lambda_2}(x, y, \lambda_1, \lambda_2) = 0$$

Si se cumplen las condiciones KKT (Karush-Kuhn-Tucker) el problema de optimización con restricciones se podrá resolver como un problema de optimización con igualdades.

Optimización mediante el método de multiplicadores de Lagrange

Queremos resolver el problema de minimización siguiente:

$$\min_{\vec{\omega},b} J = \frac{1}{2}\|\vec{\omega}\|^2$$

sujeto a las restricciones

$$y_k(\vec{\omega}\cdot\vec{x}_k + b) \geq +1,$$

donde $k = 1,2,\ldots,n$; además de que:

$$y_k = \begin{cases} +1 & si \quad \vec{x}_k \in A \\ -1 & si \quad \vec{x}_k \in B \end{cases}$$

a partir del conjunto de datos de entrada:

$$\{\underbrace{(x_{k1}, x_{k2}, \ldots, x_{km}}_{\vec{x}_k}, y_k)\}_{k=1}^{n}$$

Calculemos y simplifiquemos primeramente el Lagrangiano:

$$\mathcal{L}(\vec{\omega}, b, \lambda_1, \lambda_2, \ldots, \lambda_n) \equiv \frac{1}{2}\|\vec{\omega}\|^2 - \sum_{k=1}^{n} \lambda_k\{y_k(\vec{\omega}\cdot\vec{x}_k + b) - 1\}$$

$$= \frac{1}{2}\|\vec{\omega}\|^2 - \sum_{k=1}^{n}\{\lambda_k y_k \vec{\omega}\cdot\vec{x}_k + \lambda_k y_k b - \lambda_k\}$$

$$= \frac{1}{2}\|\vec{\omega}\|^2 - \sum_{k=1}^{n} \lambda_k y_k \vec{\omega}\cdot\vec{x}_k - \sum_{k=1}^{n} \lambda_k y_k b + \sum_{k=1}^{n} \lambda_k$$

$$= \underbrace{\frac{1}{2}\vec{\omega}\cdot\vec{\omega} - \vec{\omega}\cdot\left\{\sum_{k=1}^{n} \lambda_k y_k \vec{x}_k\right\} - b\left\{\sum_{k=1}^{n} \lambda_k y_k\right\} + \sum_{k=1}^{n} \lambda_k}_{\text{Lagrangiano } \mathcal{L}}$$

Es decir,

$$\mathcal{L}(\vec{\omega}, b, \lambda_1, \lambda_2, \dots, \lambda_n) = \frac{1}{2}\vec{\omega}\cdot\vec{\omega} - \vec{\omega}\cdot\left\{\sum_{k=1}^{n}\lambda_k y_k \vec{x}_k\right\} - b\left\{\sum_{k=1}^{n}\lambda_k y_k\right\} + \sum_{k=1}^{n}\lambda_k$$

Todos las $\vec{x}_k$ que no sean vectores de soporte no cumplen la igualdad de la restricciones, pero entonces deberán satisfacer la última restricción con $\lambda_k = 0$.

Ahora, las condiciones necesarias (condiciones de Karush-Kuhn-Tucker) para un mínimo local $\vec{\omega}^*$ en el problema de optimización propuesto son, para $k = 1, 2, \dots, n$:

$$\frac{\partial \mathcal{L}}{\partial \vec{\omega}} = 0, \qquad \frac{\partial \mathcal{L}}{\partial b} = 0, \qquad \lambda_k \geq 0, \qquad \lambda_k\{y_k(\vec{\omega}\cdot\vec{x}_k + b) - 1\} = 0,$$

Las condiciones de KKT son el análogo de que en los máximos y mínimos las parciales sean cero.

Calculando entonces las derivadas parciales de $\mathcal{L}$ con respecto a $\vec{\omega}$ y b primeramente:

$$\frac{\partial \mathcal{L}}{\partial \vec{\omega}} = \vec{\omega} - \sum_{k=1}^{n}\lambda_k y_k \vec{x}_k \qquad \frac{\partial \mathcal{L}}{\partial b} = -\sum_{k=1}^{n}\lambda_k y_k$$

Igualando a cero y despejando:

$$\vec{\omega} = \sum_{k=1}^{n}\lambda_k y_k \vec{x}_k \qquad \sum_{k=1}^{n}\lambda_k y_k = 0$$

Resolviendo este sistema de ecuaciones llegaremos a la solución óptima buscada.

De las expresiones anteriores observamos lo siguiente:

- Como $\vec{\omega} = \sum_{k=1}^{n} \lambda_k y_k \vec{x}_k$ y los multiplicadores $\lambda_k \geq 0$, entonces el vector $\vec{\omega}^*$ de la solución óptima se expresa como una combinación lineal de $n^* \leq n$ vectores $\vec{x}_k$ tales que $\lambda_k \neq 0$. Estos vectores $\vec{x}_k,\ k = 1,2, \ldots, n^*$, son los llamados **vectores de soporte** del clasificador.

- De la condición $\lambda_k\{y_k(\vec{\omega} \cdot \vec{x}_k + b) - 1\} = 0$, con $\lambda_k > 0$ vemos que los vectores de soporte descansan sobre cada uno de los hiperplanos $\vec{\omega}^* \cdot \vec{x}_k + b = \pm 1$.

- Es decir, los vectores de soporte son los vectores de entrenamiento que están más cerca del hiperplano de margen máximo, de hecho a una distancia normalizada de ± 1.

- Los vectores de entrenamiento para los que $\lambda_k = 0$ no afectan al hiperplano HMM.

- Una vez obtenido el vector óptimo $\vec{\omega}^*$, el valor de la constante b se puede obtener de cualquiera de las condiciones $\vec{\omega}^* \cdot \vec{x}_k + b = \pm 1$, donde $\lambda_k > 0$.

- En la práctica el valor de b se obtiene mediante el promedio de todas las soluciones obtenidas de $y_k(\vec{\omega} \cdot \vec{x}_k + b) = 1$, donde $\lambda_k > 0$.

- Así, la búsqueda del hiperplano de margen máximo (MMH) nos lleva a la búsqueda de los vectores de soporte.

- El hecho de que la función de costo sea una función convexa y que las restricciones estén basadas en funciones lineales, garantiza que el mínimo local encontrado será un mínimo global único.

Problema de Optimización Dual

En ocasiones conviene trabajar con el problema dual Lagrangiano, el cual se obtiene expresando la función de costo y restricciones mediante los multiplicadores de Lagrange únicamente.

Veamos, previamente obtuvimos el Lagrangiano:

$$\mathcal{L}(\vec{\omega}, b, \lambda_1, \lambda_2, \dots, \lambda_n) = \frac{1}{2}\vec{\omega}\cdot\vec{\omega} - \vec{\omega}\cdot\left\{\sum_{k=1}^{n}\lambda_k y_k \vec{x}_k\right\} - b\left\{\sum_{k=1}^{n}\lambda_k y_k\right\} + \sum_{k=1}^{n}\lambda_k$$

y las igualdades

$$\vec{\omega} = \sum_{k=1}^{n}\lambda_k y_k \vec{x}_k \qquad\qquad \sum_{k=1}^{n}\lambda_k y_k = 0$$

Que al combinarlas obtenemos:

$$\mathcal{L}(\lambda_1, \lambda_2, \dots, \lambda_n) = \frac{1}{2}\left\{\sum_{k=1}^{n}\lambda_k y_k \vec{x}_k\right\}\cdot\left\{\sum_{j=1}^{n}\lambda_j y_j \vec{x}_j\right\} - \left\{\sum_{k=1}^{n}\lambda_k y_k \vec{x}_k\right\}\cdot\left\{\sum_{j=1}^{n}\lambda_j y_j \vec{x}_j\right\} - b\left\{\sum_{k=1}^{n}\lambda_k y_k\right\} + \sum_{k=1}^{n}\lambda_k$$

$$\mathcal{L}(\lambda_1, \lambda_2, \dots, \lambda_n) = -\frac{1}{2}\left\{\sum_{k=1}^{n}\sum_{j=1}^{n}\lambda_k\lambda_j y_k y_j \vec{x}_k\cdot\vec{x}_j\right\} + \sum_{k=1}^{n}\lambda_k$$

Expresión que usaremos para escribir el problema dual.

Problema Dual de Optimización HMM: Clases Linealmente Separables

Así, como el problema de optimización (minimización) original está dada por una función de costo convexa y restricciones lineales, el problema en la forma dual de Wolfe se expresará como un problema de maximización de la forma siguiente y en términos únicamente de los multiplicadores de Lagrange:

$$\max_{\lambda_1,\ldots,\lambda_n} \left\{ \sum_{k=1}^{n} \lambda_k - \frac{1}{2} \sum_{k=1}^{n} \sum_{j=1}^{n} \lambda_k \lambda_j y_k y_j \left(\vec{x}_k \cdot \vec{x}_j \right) \right\}$$

sujeto a las restricciones:

$$\begin{cases} \lambda_k \geq 0, \text{ para } k = 1,2,\ldots,n \\ \sum_{k=1}^{n} \lambda_k y_k = 0 \end{cases}$$

Observamos que ahora las restricciones son solamente igualdades y condiciones de no-negatividad, a diferencia del problema original que involucra desigualdades.

Sin embargo, es un nuevo problema con una nueva función objetivo y una restricción en la que se vuelve a aplicar el método de multiplicadores de Lagrange.

Y a partir del conjunto de datos de entrada $\{(x_{k1}, x_{k2}, \ldots, x_{km}, y_k)\}_{k=1}^{n}$

donde $\vec{x}_k = (x_{k1}, x_{k2}, \ldots, x_{km})$, $y_k = \begin{cases} +1 & si \quad \vec{x}_k \in A \\ -1 & si \quad \vec{x}_k \in B \end{cases}$.

De la forma dual del problema de optimización anterior observamos lo siguiente:

- Los vectores de soporte se encuentran a través de los coeficientes de Lagrange no cero.
- Una vez encontrados los multiplicadores de Lagrange λ_k podemos determinar el hiperplano HMM $\vec{\omega} \cdot \vec{x} + b = 0$, mediante las expresiones:

$$\vec{\omega} = \sum_{k=1}^{n} \lambda_k y_k \vec{x}_k \,, \qquad b = y_k - \vec{\omega} \cdot \vec{x}_k$$

El coeficiente b en realidad se obtiene como el promedio para todas las k de los vectores de soporte.

- La unicidad de la representación del hiperplano mediante los λ_k en la sumatoria para $\vec{\omega}$ no necesariamente es única, sin embargo sí lo es el hiperplano HMM.
- Para determinar la clase a la que pertenecerá un nuevo dato $\vec{z}$ no se requiere encontrar de manera explícita el hiperplano HMM, sino solamente determinar, para algún $k = j_0$ de algún vector de soporte, es decir con $\lambda_{j_0} > 0$, el signo de la expresión:

$$\vec{\omega} \cdot \vec{z} + b = y_{j_0} + \sum_{k=1}^{n} \lambda_k y_k (\vec{x}_k \cdot \vec{z}) - \sum_{k=1}^{n} \lambda_k y_k \left(\vec{x}_k \cdot \vec{x}_{j_0}\right)$$

- Observa que el problema dual de optimización está definido por el producto punto entre los vectores de entrenamiento.