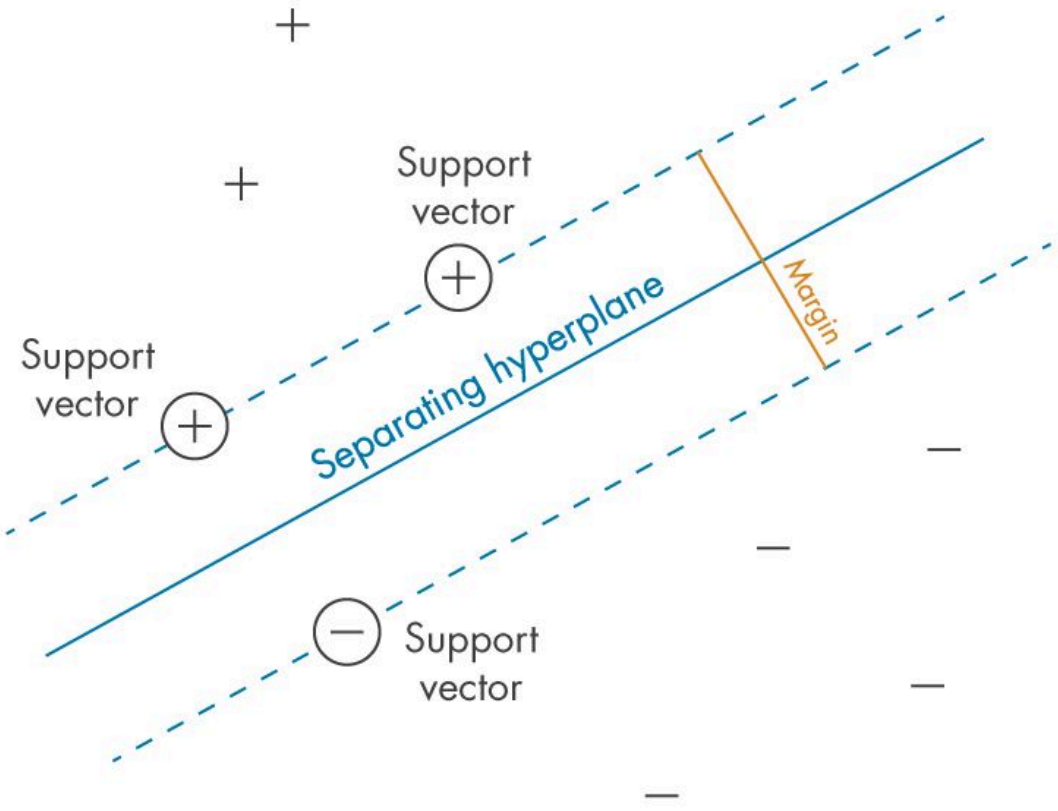
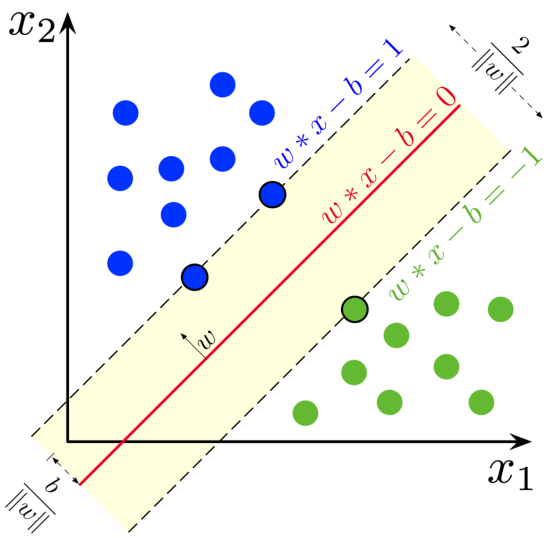
SVM - Suppor Vector Machines
MAQUINAS DE SOPORTE VECTORIAL (Aprendizaje supervisado)



El algoritmo de Máquinas de Vectores de Soporte (SVM, por sus siglas en inglés: Support Vector Machines) es un método de aprendizaje supervisado utilizado tanto para clasificación como para funciones de regresión. Su objetivo principal en problemas de clasificación es encontrar el hiperplano óptimo que mejor separa las clases en un espacio de características de alta dimensión.



Aquí hay una descripción general del funcionamiento del algoritmo SVM:

- **Definición del Espacio de Características:** El SVM opera en un espacio de características de alta dimensión donde cada instancia de datos se representa como un vector en este espacio, con cada característica del conjunto de datos representada por una dimensión.

- **Selección del Hiperplano Óptimo:** El SVM busca encontrar el hiperplano de separación que mejor divide las clases en el espacio de características. Este hiperplano es el que maximiza el margen entre las instancias de datos más cercanas de cada clase, conocidas como vectores de soporte.

- **Clasificación de Nuevos Datos:** Una vez que se ha encontrado el hiperplano óptimo, el SVM puede clasificar nuevas instancias de datos determinando en qué lado del hiperplano se encuentran. Si la nueva instancia de datos está en el lado positivo del hiperplano, se clasifica en una clase, y si está en el lado negativo, se clasifica en la otra clase.

- **Kernel Trick:** En casos donde los datos no son linealmente separables en el espacio de características original, el SVM puede emplear una técnica llamada Kernel Trick. Esto implica mapear los datos a un espacio de características de mayor dimensión donde sí sean linealmente separables, lo que permite encontrar un hiperplano de separación óptimo en ese espacio transformado.

- **Parámetro de Regularización (C):** El parámetro de regularización C en el SVM controla la suavidad del margen y la cantidad de clasificaciones incorrectas

permitidas. Valores más altos de C penalizan más fuertemente las clasificaciones incorrectas, lo que puede llevar a un hiperplano con un margen más estrecho pero con menos errores de clasificación.

El algoritmo SVM es robusto y efectivo para una amplia gama de problemas de clasificación, especialmente cuando hay un margen de separación claro entre las clases. Sin embargo, puede ser computacionalmente costoso, especialmente en conjuntos de datos grandes, y puede requerir ajuste de hiperparámetros cuidadoso para obtener el mejor rendimiento.

En las Máquinas de Vectores de Soporte (SVM), las variables W y b desempeñan papeles cruciales en la definición del hiperplano que separa los puntos de datos pertenecientes a diferentes clases.

W (Vector de Peso):

- **Definición:** El **vector de peso (W)** es un vector de alta dimensión que representa la dirección del hiperplano. Determina la orientación y la pendiente del hiperplano en el espacio de características.
- **Interpretación:** El vector **W** apunta hacia el vector normal del hiperplano. Esto significa que es perpendicular al hiperplano e indica la dirección en la que el hiperplano es más sensible a los cambios en los datos de entrada.
- **Significado:** El vector **W** es crucial para definir la frontera de decisión, que es la línea o plano que separa los puntos de datos de una clase de los de la otra. Al maximizar el margen entre el hiperplano y los puntos de datos más cercanos (vectores de soporte), las SVM buscan encontrar un hiperplano que sea lo más robusto posible a las variaciones en los datos.

b (Término de Sesgo - “bias”):

- **Definición:** El **término de sesgo (b)** es un valor escalar que representa el desplazamiento del hiperplano desde el origen. Determina la posición del hiperplano en el espacio de características.
- **Interpretación:** El término **b** desplaza el hiperplano a lo largo de su vector normal (representado por **W**). Asegura que el hiperplano no esté restringido a pasar por el origen, lo que le permite adaptarse mejor a la distribución de los datos.
- **Significado:** El término **b** es esencial para ajustar la posición del hiperplano, especialmente cuando se trata de datos no separables linealmente. Al ajustar **b**,

el hiperplano puede moverse para acomodar puntos de datos que pueden estar ligeramente fuera del margen óptimo.

Juntos, W y b definen la ecuación del hiperplano en el espacio de características:

w^T * x + b = 0

donde:

- **w^T** es la transpuesta del vector de peso **W**
- **x** es el punto de datos de entrada
- **b** es el término de sesgo

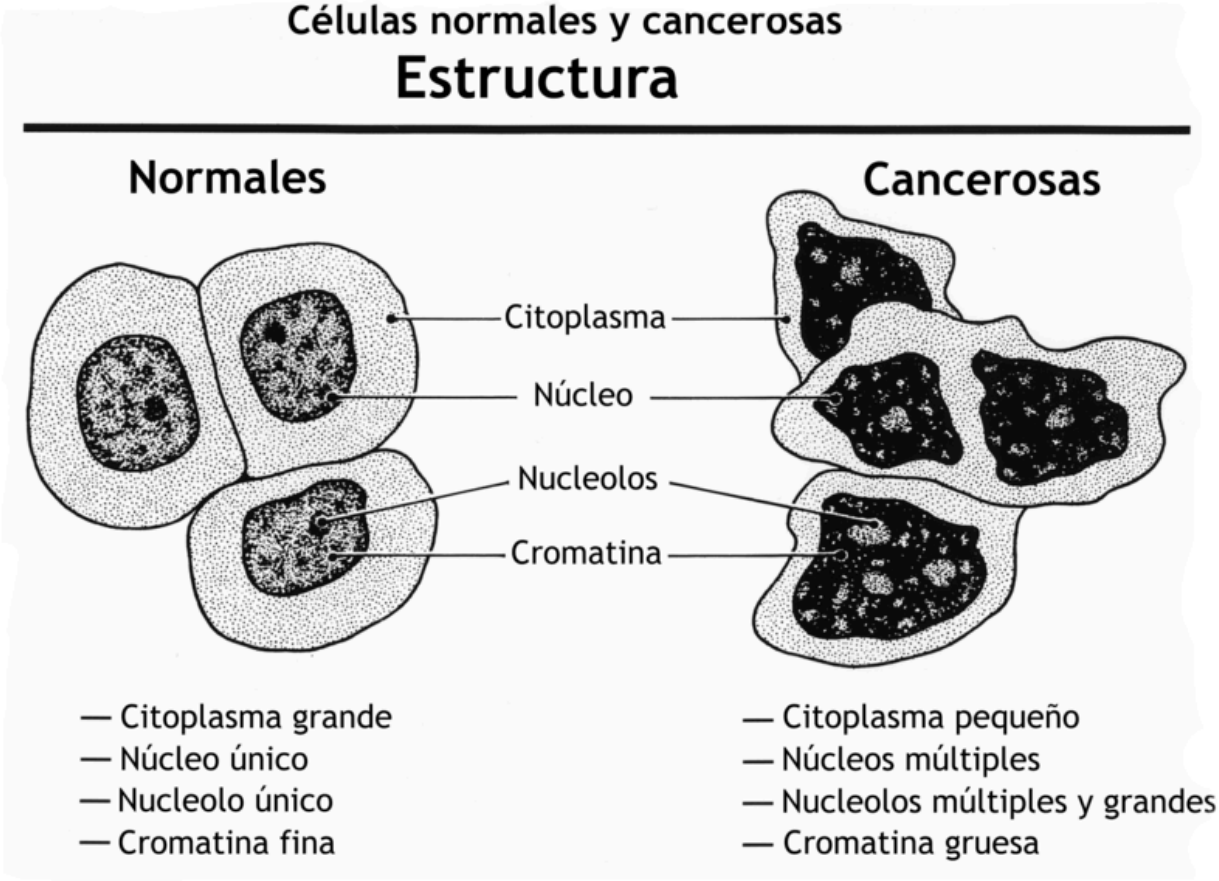
Esta ecuación representa la frontera de decisión, clasificando los puntos de datos como pertenecientes a una clase si la expresión es positiva, y a la otra clase si la expresión es negativa.

W y b son parámetros fundamentales en las SVM, determinando la orientación, la posición y la frontera de decisión del hiperplano que separa los puntos de datos. Su interacción permite a las SVM clasificar eficazmente los datos en espacios de alta dimensión.

—---------

CASO PRACTICO

Programa para identificar si una celula puede ser celula con cancer o normal.



Consideremos un conjunto de datos con está información:

- Tamaño_nucleo: Representa el tamaño del núcleo celular.
- Forma_celula: Describe la forma de la célula (por ejemplo, "round" para redonda, "irregular" para irregular).
- Densidad_nucleo: La densidad del núcleo celular.
- Clase: La etiqueta de clase que indica si la muestra es cancerosa (1) o normal (0).