MÁQUINAS DE VECTOR SOPORTE

APRENDIZAJE AUTOMÁTICO SUPERVISADO

Guillermo de Jorge Botana

Dpto. Psicobiología y Metodología en Ciencias del Comportamiento

Facultad de Psicología.

Universidad Complutense de Madrid.

NOTA:

El contenido de este texto corresponde a uno de los temas de una asignatura del Máster de Metodología de las Ciencias del Comportamiento y de la Salud. Está elaborado para tener un texto base de lo que es explicado en clase. Aunque el texto es seguido y coherente, puede ser susceptible de algunas mejoras y ampliaciones. No obstante, es lo suficientemente autocontenido para llevar a cabo un estudio independiente sobre él.

He decidido publicar este texto fuera del ámbito de la asignatura por si puede resultar de utilidad para otros estudiantes o por si a otros docentes les puede facilitar la tarea.

Tabla de contenido

1. Introducción

Las Máquinas de Vector Soporte son una de las técnicas clásicas dentro del aprendizaje automático supervisado. Surgió alrededor de los años noventa y desde entonces ha sido usada prolijamente. Como veremos a lo largo del texto, su metodología de clasificación está basada encontrar el mayor margen de separación entre dos subconjuntos de datos de diferente clase.

El hecho de que se quiera encontrar este mayor margen hace que sea eminentemente un problema de optimización. Cuando hablamos de margen estamos hablando también de distancia entre algunos ejemplares, de encontrar una mayor distancia, pero respetando la constricción de que ese margen permita discriminar entre clases. De ahí que emplee procedimientos de optimización, como los multiplicadores de Lagrange.

En general, es usada para clasificaciones binarias, lo que la hace homóloga a la regresión logística, aunque también puede modificarse para clasificación multiclase.

2. Objetivos

En la introducción hemos dado ya ciertas pinceladas del procedimiento a usar. Ahora vamos a ser más sintéticos y operativos. Diremos que el objetivo de la técnica es separar ejemplares de una clase y otra introduciendo un hiperplano (la recta verde de la figura 1) colocado entre las coordenadas que representan a los ejemplares de ambos grupos. Ya tenemos la primera idea, los ejemplares serán representados por medio de coordenadas en un espacio y también en ese espacio existirá un hiperplano separador.

No obstante, ese hiperplano que introduzcamos ha de ser el óptimo, que en este caso es que deje el mayor margen entre el ejemplar más cercano de una clase y de la otra.

Con esto tenemos ya las dos directrices que definen los objetivos de la técnica: **buscar un hiperplano que separe**, que discrimine, pero buscarlo **con la mayor distancia entre ambos grupos**, en concreto, entre los ejemplares más próximos a ese posible hiperplano (las dos distancias en negro de la figura 1).

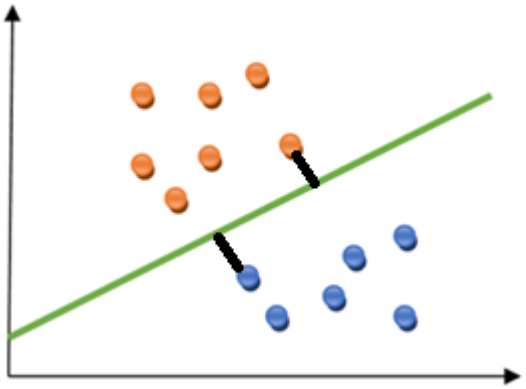

Figura 1. Elaborada a partir de https://databasetown.com/implementing-support-vector-machine-svm-in-python/

3. Procedimiento

3.1 Disposición de los datos

Al igual que los árboles de decisión y otras técnicas supervisadas, la tabla de entrada que se emplea como entrada a la técnica está formada por un conjunto de variables independientes o predictoras continuas, habitualmente escaladas, y una variable binaria a predecir, a veces llamada criterio (figura 2). Las variables predictoras se suelen notar con x y la variable criterio con y. Esa es la notación que también seguiremos en este texto. Un conjunto de valores de x corresponden con un valor binario de y (éxito o fracaso).

Se pueden considerar las x de cada fila como las propiedades que definen un ejemplar, y representarse como una coordenada en forma de vector (como un vector $\mathbf{x}$). La coordenada formada por $\mathbf{x}$ es susceptible de ser representada en el espacio n-dimensional, siendo n del mismo número que el número de componentes en $\mathbf{x}$, es decir, que el número de variables predictoras. Por ejemplo, la figura 1 muestra un espacio en el que se representan ejemplares en base a su coordenada definida por los predictores x_1 y x_2.

X1	X2	Y
2.7	2.5	0
1.4	2.3	0
3.3	2.4	0
1.3	1.8	0
3	3	0
7.6	2.7	1
5.9	2.2	1
6.9	1.8	1
8.6	3.5	1
7.7	3.5	1

Figura 2

3.2 Formalización del hiperplano de separación

Ya conocemos el objetivo final de la técnica de Máquinas de Vector Soporte. Se trata de **separar ejemplares** de una clase y otra **trazando un hiperplano entre los ejemplares de ambas clases**. Sabemos también que esos ejemplares están identificados en **un espacio** atendiendo a sus variables independientes. Habíamos dicho también que **no valía cualquier hiperplano**. El hiperplano que se trace **ha de dejar entre los ejemplares de ambos grupos el máximo margen**. En otras palabras, los dos ejemplares que queden más cerca del hiperplano han de mantener la máxima distancia con él. Ese puede ser el resumen de lo hasta ahora estudiado, pero hagamos un repaso visual de estos conceptos para luego poder estudiarlos con más detalle.

Observemos la figura 3. En ella se representan tres hiperplanos: H_1, H_2 y H_3. En este caso H_1 no separa bien a los ejemplares negros y blancos. Mal negocio. Por contra, H_2 si que lo hace. H_2 sería una solución para la clasificación, no obstante, no sería la nuestra. H_2 no es el hiperplano que deja el mayor margen. Fijémonos ahora el H_3. Ahora sí, H_3 no solamente es una solución para separar, sino que es **la mejor solución** pues lo hace dejando el mayor margen. Es decir, H_3 sería el hiperplano que deja la mayor distancia entre los ejemplares de sendas clases más cercanos a él. H_3 sería el **hiperplano definitivo**.

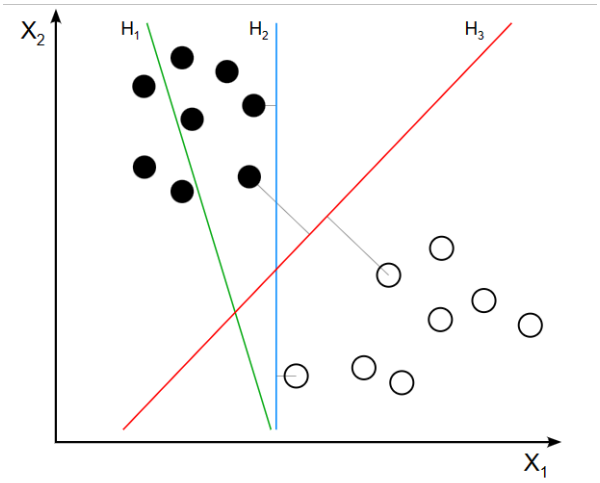

Figura 3: tomada de https://es.wikipedia.org/wiki/M%C3%A1quinas_de_vectores_de_soporte

Habiendo visto la figura y sus explicaciones, podemos intuir que lo primero que necesitamos es poseer **un aparataje matemático para definir formalmente cualquier hiperplano**. Si queremos elegir el mejor, debemos tener la capacidad de definir cualquier de ellos y luego imponer las restricciones para elegir el mejor.

Vamos a considerar el caso general de **formalización de un hiperplano** (las rectas de la figura 3 son su caso más simple) que separe los ejemplares de una clase de la otra. Un hiperplano puede ser definido como un producto escalar entre el vector de coeficientes (perpendicular al hiperplano) por el vector de variables al que se le suma un sesgo:

$$H(\mathbf{x}) = \mathbf{w}^{\mathbf{T}}\mathbf{x} + \mathrm{b} \text{ (fórmula 1)}$$

Donde:

$\mathbf{w}$ es un vector de coeficientes que es perpendicular al hiperplano (sea cual sea éste)

$\mathbf{x}$ es un vector con las variables que definen a los ejemplares

El hecho de que $\mathbf{w}$ sea perpendicular al plano se justifica en que el producto escalar $\mathbf{w}^{\mathbf{T}}\mathbf{x}$ (la proyección de $\mathbf{x}$ en $\mathbf{w}$) más el sesgo b tiene que vascular de arriba abajo para estar en un lado u otro del hiperplano (ver figura 4). El sesgo b sumado a la sombra que $\mathbf{x}$ proyecta en $\mathbf{w}$ tiene que dar por encima o por debajo de un valor que sirva para discriminar los ejemplares de ambas clases.

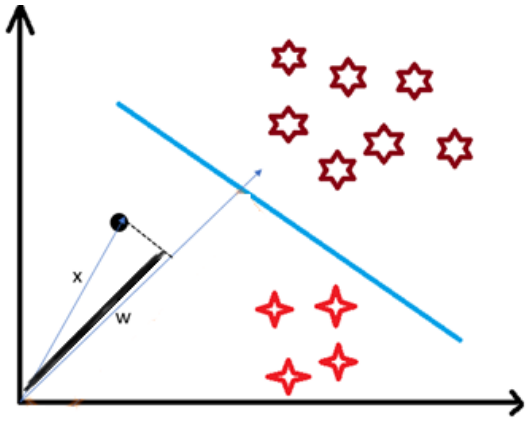

Figura 4: Elaborada a partir de https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/

Y teniendo en cuenta el margen ese valor lo podemos poner de la siguiente forma: pertenece a una clase si es 1 o por encima y a la otra clase si es -1 o por debajo. De esta forma lo que tiene que introduciendo los valores de un ejemplar concreto en $\mathbf{x}$, y multiplicado por el vector de coeficientes, más el sesgo, caiga en uno o en otro lado del hiperplano dependiendo si es de una clase u otra. Formalmente, $H(\mathbf{x})$ valdrá 1 (o por encima) si pertenece a una de las clases y -1 (o por debajo) si pertenece a la otra. Eso se puede expresar de la siguiente forma:

$$\mathbf{w}^{\mathbf{T}}\mathbf{x} + \mathrm{b} \geq 1 \text{ para los ejemplares } \mathrm{y}^{\mathrm{n}} = 1 \text{ (fórmula 2)}$$

$$\mathbf{w}^{\mathbf{T}}\mathbf{x} + \mathrm{b} \leq -1 \text{ para los ejemplares } \mathrm{y}^{\mathrm{n}} = -1 \text{ (fórmula 3)}$$

Donde:

y^{n} es el n-ésimo ejemplar del conjunto de datos.

Expresado en román paladino: si tomamos todos n los ejemplares de una clase, por ejemplo, los que cumplen que $\mathrm{y}^{\mathrm{n}} = 1$, y los pasamos por la función del hiperplano $H(\mathbf{x})$**,** el valor que toma la función debe ser sistemáticamente igual o mayor que 1. Ese es el sentido de la restricción para clasificar bien. Lo mismo se podría decir de los ejemplares $\mathrm{y}^{\mathrm{n}} = -1$. La función del hiperplano tomaría sistemáticamente valores iguales o menores que -1.

En otro orden de cosas, para familiarizarnos con el aparataje, podemos ver también las expresiones de la fórmula 2 y 3 en su expresión expandida como:

$$\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} (x_1 \quad x_2) + b \geq 1 \text{ para } \mathrm{y}^{\mathrm{n}} = 1 \quad \text{(fórmula 4)}$$

$$\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} (x_1 \quad x_2) + b \leq -1 \text{ para } \mathrm{y}^{\mathrm{n}} = -1 \text{ (fórmula 5)}$$

Y desarrollando podemos ver su verdadero aspecto lineal:

$$H(x_1, x_2) = \ w_1 x_1 + \ w_2 x_2 + b$$

$$w_1 x_1 + \ w_2 x_2 + b \geq \ 1 \ \text{ para } \mathrm{y}^{\mathrm{n}} = 1 \text{ (fórmula 6)}$$

$$w_1 x_1 + \ w_2 x_2 + b \leq \ -1 \ \text{ para } \mathrm{y}^{\mathrm{n}} = -1 \text{ (fórmula 7)}$$

Esto último homologa los hiperplanos a una regresión en la que cada variable tiene su coeficiente w. Para manejar mejor esta restricción impuesta al hiperplano, en los textos se suelen ofrecer las dos expresiones de las fórmulas 2 y 3 en una sola (luego veremos por qué):

$$\mathrm{y}^n \left(\mathbf{w}^{\mathbf{T}} \mathbf{x}_n + \mathrm{b}\right) = \ 1 \qquad n = 1, \dots, N \qquad \text{(fórmula 8)}$$

Con esta expresión sabemos que ningún valor positivo o negativo cruzará la línea representada por el hiperplano $H(\mathbf{x})$. Es decir, la función del hiperplano se adapta para dar cabida y predecir dos tipos de ejemplares, los $y = 1$ y los $y = -1$. Lo que habrá que hacer es estimar los

coeficientes **w** de manera que la restricción se cumpla. En forma de pregunta se podría formular como, ¿qué valores de **w** hacen que un posible hiperplano discrimine?

Enunciado de una manera muy formal, y de manera abstracta, diremos que discriminar significa la capacidad de **encontrar un w** cuyo producto escalar con cualquier **x** y sumado el sesgo b**, ofrezca un valor que sistemáticamente asigne 1 o más a los ejemplares de una clase y -1 o menos a los de la otra**. Dicho de otra forma, un vector perpendicular a hiperplano que sirva para acertar sistemáticamente. La consecuencia de estas afirmaciones es que el hiperplano que discrimina no es más que un ente abstracto definido por un **w** perpendicular y un sesgo. En la figura 5 se muestra una representación de la tarea. Encontrar el hiperplano $H(x_1, x_2)$ que evaluado con cualquier dupla $\{x_1, x_2\}$ haga que tome un valor de 1 o por encima o de -1 o por debajo según sea la clase.

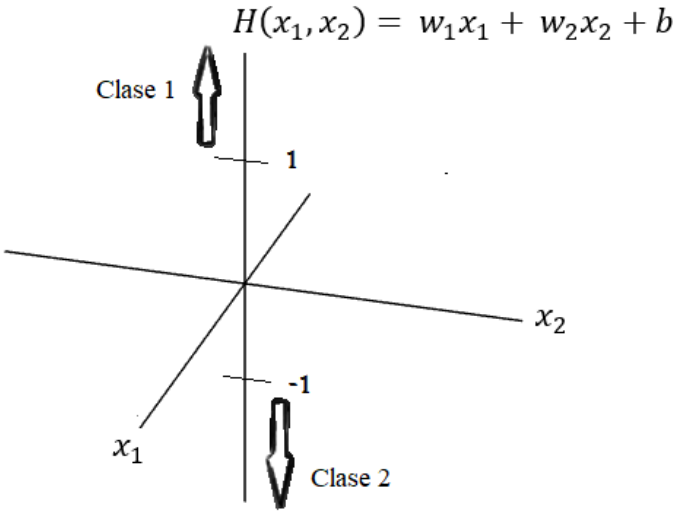

Figura 5. El reto es encontrarlos valores de w_1, w_2 y b que hagan que al introducir cualquier dupla $\{x_1, x_2\}$ la función $H(x_1, x_2)$ de valores de 1 o por encima si el ejemplar representado en esa dupla es de una categoría o de -1 o por debajo si es de otra. Esta es la primera consigna general del procedimiento de las Máquinas de Vector Soporte.

Por último, en la figura 3 se puede ver que existen varias funciones lineales con el aspecto $\mathbf{w}^{\mathbf{T}}\mathbf{x} + \mathrm{b} = 0$ que cumplen las restricciones expresadas en las fórmulas 2 y 3. Serían H_2 y H_3. No así H_1. Tenemos pues dos tentativas de hiperplanos que ya discriminan. Sin embargo, necesitamos uno solo: el óptimo, el que deja la mayor distancia entre los ejemplares más cercanos a él que representen ambas clases. Demos pues un pasito más.

3.3 El mejor hiperplano: el mayor margen

Ya hemos podido definir formalmente los hiperplanos que discriminan ejemplares de una clase y otra. Esos hiperplanos tenían unos valores en el vector $\mathbf{w}$ que hacían que introduciendo ejemplares de una clase en la función $H(\mathbf{x}) = \mathbf{w}^{\mathbf{T}}\mathbf{x} + \mathrm{b}$ daba sistemáticamente valores iguales o mayores que 1 en el caso de que fuesen ejemplares 1 o menores o iguales a -1 en el caso que fuesen ejemplares -1. El que cumpla esta restricción depende pues del vector de coeficientes $\mathbf{w}$ y del valor del intercepto b.

Sin embargo, necesitamos el hiperplano que además de cumplir esa restricción -la restricción de que sea discriminante- deje el mayor margen en torno a sí. Es decir, necesitamos H_3 (figura 3) como solución óptima, acaso diríamos única. Ya se ha definido intuitivamente que el margen es la distancia del hiperplano a los ejemplares más cercanos de cada una de las clases. Esos ejemplares más cercanos se llaman justamente **"vectores de apoyo"** (o de "soporte") por ser donde se apoya el hiperplano óptimo y su margen.

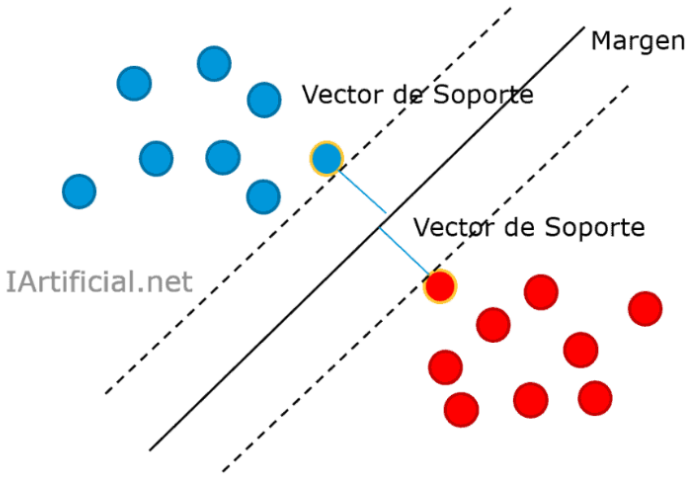

Figura 6. Tomada de https://www.iartificial.net/maquinas-de-vectores-de-soporte-svm/

Podemos ya dar una **definición formal y exhaustiva del margen** de un hiperplano: es la distancia de cada vector de apoyo y el hiperplano $H(\mathbf{x})$. Se llamará **margen máximo** si ese margen es el mayor posible que se pueda trazar entre todos los vectores de apoyo posibles (figura 6). Se puede decir también que cada uno de los márgenes hace que el hiperplano $H(\mathbf{x})$ esté cubierto de otros dos hiperplanos por cada uno de los lados.

El hecho de que se busque el margen más amplio tiene su justificación. Se hace para que en la clasificación de nuevos ejemplares se cometa el mínimo número de errores. Cuanto más margen, **menor riesgo de confusión para los ejemplares no vistos por el modelo**. Algo así como una zona desmilitarizada en la frontera que haga nítido quién es de un bando y quien es de otro. Es una forma de que el modelo **generalice mejor**. En ocasiones, la búsqueda de este margen máximo es incluso preferida a la comisión de algunos errores, si estos son anecdóticos. Aunque se comentan algunos errores en la muestra de entrenamiento, la búsqueda de un margen amplio quedará compensada cuando introduzcamos datos nuevos y se clasifiquen, pues la clasificación está menos expuesta al sobreajuste (ver figura 7). La solución no solo servirá para los datos de la muestra de entrenamiento, sino para datos nuevos. Para dar cuenta de la importancia del criterio del margen frente a la comisión de errores, los paquetes tienen un hiperparámetro habitualmente llamado c o coeficiente de regularización. Con él se permiten o no errores anecdóticos en favor del margen.

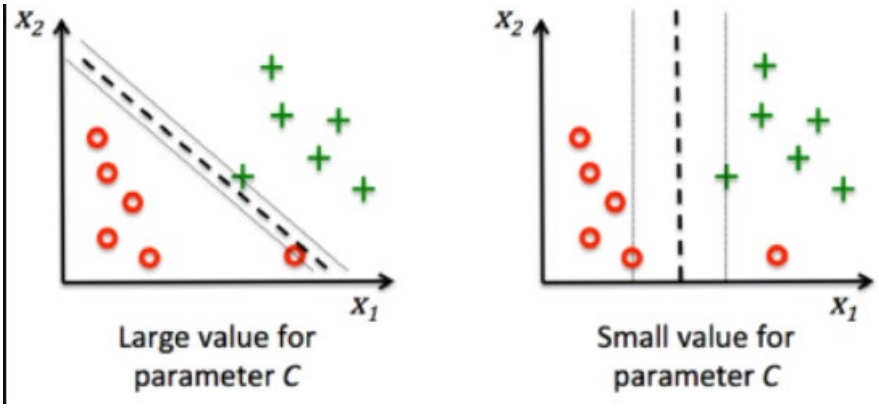

Figura 7. Tomada de https://github.com/rasbt/python-machine-learning-book-3rd-edition/blob/master/ch03/ch03.ipynb

Teniendo en cuenta el margen introducimos ya la segunda condición a la técnica. Además de buscar la discriminación por medio de la expresión de la fórmula 8, tendremos también que seleccionar el $H(\mathbf{x})$ que maximice el margen. Pondremos en orden ambos ingredientes, **discriminación** y **margen**, en breve.

3.2 Cómo se formaliza el margen

Para saber cuál es el mayor de los márgenes debemos tener una forma de calcularlo, de formalizarlo. Parece obvio. Cada hiperplano posible $H(\mathbf{x})$ tendrá asociado un margen. Por tanto, la primera cuestión que nos sale al camino es la forma de calcular ese margen. Afinando la cuestión planteada, lo que estamos buscando en primer lugar es l**a función de la distancia** d entre dos puntos separados por un hiperplano. Al tratarse de una distancia entre puntos de un espacio, la fórmula que calcula el margen está basada en nociones de algebra lineal y geometría. Está basado, afinando más, en el concepto de proyección. Observemos este gráfico:

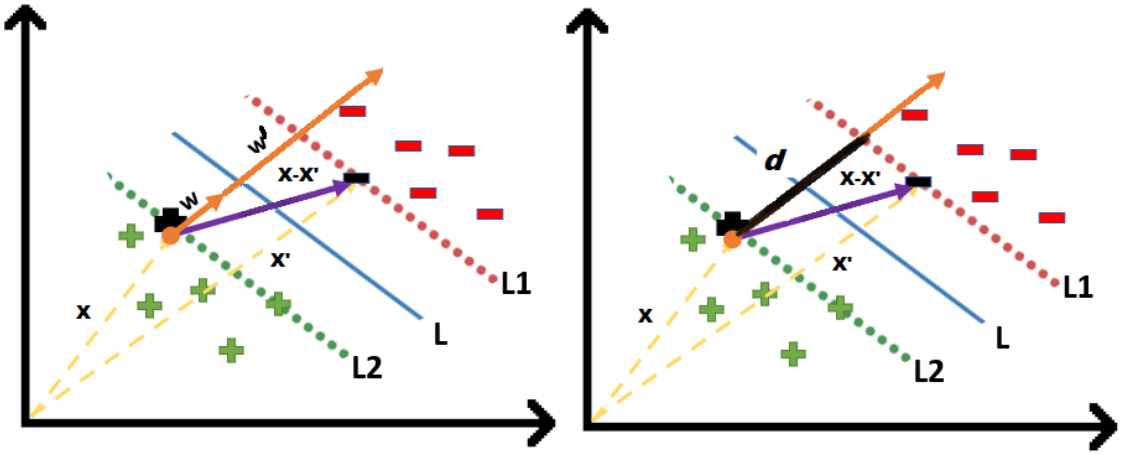

Figura 8. Los vectores que participan en las operaciones descritas en esta gráfica pueden ser también entendidas trasladando todos esos vectores al punto (0,0) de las referencias. Basada en una tomada de https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/

Los vectores de apoyo están representados por los vectores $\mathbf{x}$ y $\mathbf{x}'$respectivamente. Si quisiésemos la simple distancia entre ambos vectores de apoyo solo tendríamos que calcular la resta entre ambos vectores $\mathbf{x} - \mathbf{x}'$. Lo que ocurre es que esta medida no es una distancia pura del margen. La distancia pura del margen es la trazada desde el hiperplano que pasa por cada vector soporte y es paralelo a $H(\mathbf{x})$**.** Eso es justo lo que hay que formalizar para poder maximizarlo. Es decir, necesitamos una distancia entre ambos vectores de apoyo, pero perpendicular al hiperplano.

Si observamos la parte izquierda de la figura 8, también tenemos un vector que parte del vector soporte $\mathbf{x}$ que es además perpendicular a la línea. Es el vector $\mathbf{w}$ de los coeficientes (ver figura 4 para la justificación de su perpendicularidad al hiperplano). Imaginemos también que $\mathbf{w}'$ es el vector unitario de $\mathbf{w}$ (su norma es 1)**.** Si tuviésemos la longitud de la sombra que la resta $(\mathbf{x} - \mathbf{x}')$ hace en $\mathbf{w}'$ tendríamos justo la distancia del margen. Esa sombra la podríamos tener con la proyección de $(\mathbf{x} - \mathbf{x}')$ en el vector $\mathbf{w}'$**.** Operativamente, la forma de proyectar el vector $(\mathbf{x} - \mathbf{x}')$ sobre otro vector $\mathbf{w}'$ es calculando el producto escalar entre ambos. Por eso necesitamos que $\mathbf{w}$ pase a ser unitario. Para que la proyección de un vector sobre otro esté escalada. Recordemos también que el producto escalar ya nos ofrece un escalar de la longitud de la proyección. La proyección es de la resta en el vector unitario x_w, se expresa:

$$x_{w'} = (\mathbf{x} - \mathbf{x}')\,\mathbf{w}' \quad \text{(fórmula 9)}$$

Si expresamos la operación expandida para hacer unitario $\mathbf{w}$ **,** el vector $\mathbf{w}$ se divide entre su norma. La expresión anterior quedaría:

$$x_w = (\mathbf{x} - \mathbf{x}')\frac{\boldsymbol{w}}{\|\mathbf{w}\|} \quad \text{(fórmula 10)}$$

Esta va a ser la expresión de partida. La distancia buscada puede ser formalizada justo con esa proyección, teniéndose ya un escalar para ello (figura 8, parte derecha). Cambiamos x_w por una d de distancia y haremos más intuitiva y cómoda la fórmula:

$$d = (\mathbf{x} - \mathbf{x}')\frac{\boldsymbol{w}}{\|\mathbf{w}\|} \quad \text{(fórmula 11)}$$

Y desarrollándola más:

$$d = \frac{(\mathbf{x}\,\mathbf{w}) - (\mathbf{x}'\boldsymbol{w})}{\|\mathbf{w}\|} \quad \text{(fórmula 12)}$$

Juguemos con el numerador de la fórmula d

Como $\mathbf{x}$ es un vector cuya clase asignada es $\mathrm{y}^{\mathrm{n}} = 1$, a la expresión $\mathrm{y}^{\mathrm{n}}\,(\mathbf{w}\,\mathbf{x} + \mathrm{b}) = 1$ le corresponde:

$$1\,(\mathbf{w}\,\mathbf{x} + \mathrm{b}) = 1 \quad \text{(fórmula 13)}$$

Desarrollando tenemos:

$$(\mathbf{w}\,\mathbf{x}) = 1 - b \quad \text{(fórmula 14)}$$

Como $\mathbf{x}'$ es un vector cuya clase es $\mathrm{y}^{\mathrm{n}} = -1$, a la expresión $\mathrm{y}^{\mathrm{n}}\ (\mathbf{w}\ \mathbf{x} + \mathrm{b}) = 1$ le corresponde:

$$-1\ (\mathbf{w}\ \mathbf{x}' + \mathrm{b}) = 1 \text{ (fórmula 15)}$$

Y también desarrollando:

$$(-\mathbf{w}\ \mathbf{x}' - \mathrm{b}) = 1 \text{ (fórmula 16)}$$

$$(-\mathbf{w}\ \mathbf{x}') = 1 + b \text{ (fórmula 17)}$$

$$(\mathbf{w}\ \mathbf{x}') = -1 - b \text{ (fórmula 18)}$$

Y por tanto, sustituyendo en d:

$$d = \frac{1-b-(-1-b)}{\|\mathbf{w}\|} = \frac{1-b+1+b}{\|\mathbf{w}\|} = \frac{2}{\|\mathbf{w}\|} \text{ (fórmula 19)}$$

El margen d es:

$$d = \frac{2}{\|\mathbf{w}\|} \text{ (fórmula 20)}$$

Ya tenemos formalizado el margen que se desea maximizar. El juego con las fórmulas nos ha llevado a simplificar la distancia. El margen depende al final únicamente de los coeficientes $\mathbf{w}$ de la función del hiperplano $H(\mathbf{x})$**.** El hiperplano óptimo será el que maximiza d. Lo siguiente es juntar ambas restricciones, capacidad discriminante y distancia máxima en una sola expresión. Es lo que viene ahora.

3.3 Maximización del margen

Ya tenemos todas las formalizaciones. La primera era la que concernía a la restricción de que el hiperplano sea capaz de discriminar entre ejemplares de distintas clases. La segunda concernía a maximización de la distancia. Por tanto, tenemos la función a maximizar y la restricción. Los valores de los argumentos de la función a maximizar $d(\mathbf{w})$ tendrán que cumplir la restricción

de poder también discriminar entre clases. He aquí la ligazón entre ambas cosas. De manera formal tenemos que maximizar la función:

$$d(\mathbf{w}, \mathrm{b}) = \frac{2}{\|\mathbf{w}\|} \text{ (fórmula 21)}$$

Es decir, encontrar los argumentos de maximicen d , con la restricción:

$$\mathrm{y}^{\mathrm{n}} \left(\mathbf{w}^{\mathbf{T}} \mathbf{x}_{\boldsymbol{n}} + \mathrm{b}\right) = 1 \qquad n = 1, \dots, N \text{ (fórmula 22)}$$

Lo que también puede ser escrito:

$$\min_{W} \quad \frac{2}{\|\mathbf{w}\|}$$

$$s.a \qquad \mathrm{y}^{\mathrm{n}} \left(\mathbf{w}^{\mathbf{T}} \mathbf{x}_{\boldsymbol{n}} + \mathrm{b}\right) = 1 \qquad n = 1, \dots, N$$

O la restricción expresada mediante sumatorio

$$s.a \qquad \sum_{n=1}^{N} \left[\mathrm{y}^{\mathrm{n}} \left(\mathbf{w}^{\mathbf{T}} \mathbf{x}_{\boldsymbol{n}} + \mathrm{b}\right) - 1\right] = 0$$

De otra forma dicho. Vamos a encontrar los valores de $\mathbf{w}$ y b que maximicen el margen $d(\mathbf{w})$ teniendo la restricción de que los ejemplares de una clase estén a un lado del hiperplano y los ejemplares de la otra estén al otro lado (evaluada el hiperplano de >1 o <-1). Esto último se expresa en $\mathrm{y}^{\mathrm{n}} \left(\mathbf{w}^{\mathbf{T}} \mathbf{x}_{n} + \mathrm{b}\right) = 1$, tomando n el índice de cada uno de los ejemplares. Esta es la tarea que se ha de resolver con el algoritmo apropiado.

3.4 Mecanismo de Optimización con restricciones

Y el algoritmo apropiado va a ser tratado en este apartado. Llegados a este punto tenemos delante un problema de maximización de una función con restricciones. Dada una función de

distancia, en este caso $d(\mathbf{w}, b) = \frac{2}{\|\mathbf{w}\|}$, habrá que determinar qué valores en los argumentos $\mathbf{w}$ y b la hacen máxima, es decir, extraen su mayor valor al evaluarse con ellos. Pero esto no será de forma aislada. Los argumentos que la hagan máxima tienen que hacer también que el hiperplano sea discriminante. Es decir, los posibles argumentos tienen que estar dentro de unos valores sujetos a unas restricciones determinadas. Los argumentos tienen que hacer que se cumpla $\mathrm{y}^{\mathrm{n}} \left(\mathbf{w}^{\mathbf{T}} \mathbf{x}_{\boldsymbol{n}} + \mathrm{b}\right) = 1$ con cada uno de los n ejemplares del conjunto de entrenamiento.

La búsqueda de los valores de $\mathbf{w}$ y b que hacen máxima la distancia $d(\mathbf{w}, b)$ pero sujeto a tales restricciones se hace mediante multiplicadores de Lagrange, que proporciona un método directo. Decimos directo al no emplear algoritmos que actúan en un espacio de búsqueda de parámetros. A continuación, veremos un ejemplo sencillo, manejable, aprehensible, y lo extrapolaremos después al caso que nos ocupa, al caso de las Máquinas de Vector Soporte.

3.4.1 Concepto de multiplicadores de Lagrange (un ejemplo manejable)

Imaginemos el siguiente escenario. Se nos pide que identifiquemos los argumentos que hacen máxima la siguiente función:

$$f(x, y) = 9 - x^2 - y^2$$

Para hacerlo, podríamos calcular las derivadas parciales con respecto a x e y para luego identificar máximos y mínimos igualándolas a cero. Cero pendiente significa que ni se sube ni se baja, y por lo tanto es máximo o mínimo. Podríamos desvelar la condición de máximo o mínimo con las derivadas segundas.

Lo que ocurre es que en este caso se nos dice otra cosa (figura 9 y 10). Se impone que el máximo encontrado cumpla la siguiente restricción:

$$x + y = 3$$

Es decir, que no vale lo dicho arriba. No buscamos los argumentos de x e y que hagan máxima la función $f(x, y)$. Buscamos los argumentos de x e y que hagan máxima la función $f(x, y)$ pero que cumplan la restricción $x + y = 3$. La cosa cambia. Sean cuales sean estos

argumentos, tendrán que sumar entre los dos 3. Las figuras 7 y 8 muestran la diferencia de un máximo total y un máximo condicionado a unas restricciones. En la figura 7 se muestra el máximo condicionado con una estrella blanca y el máximo total con una flecha roja. En la figura 8 se muestra como el máximo total se encuentra en la coordenada (0,0,9) y, sin embargo, el condicionado está en un plano que cumple la restricción $x + y = 3$. El máximo hay que buscarlo en ese plano.

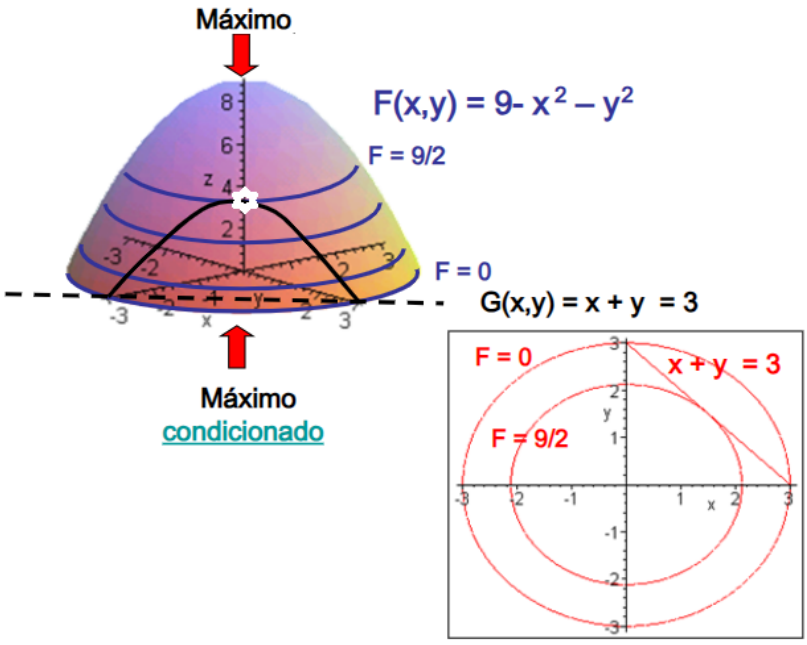

Figura 9

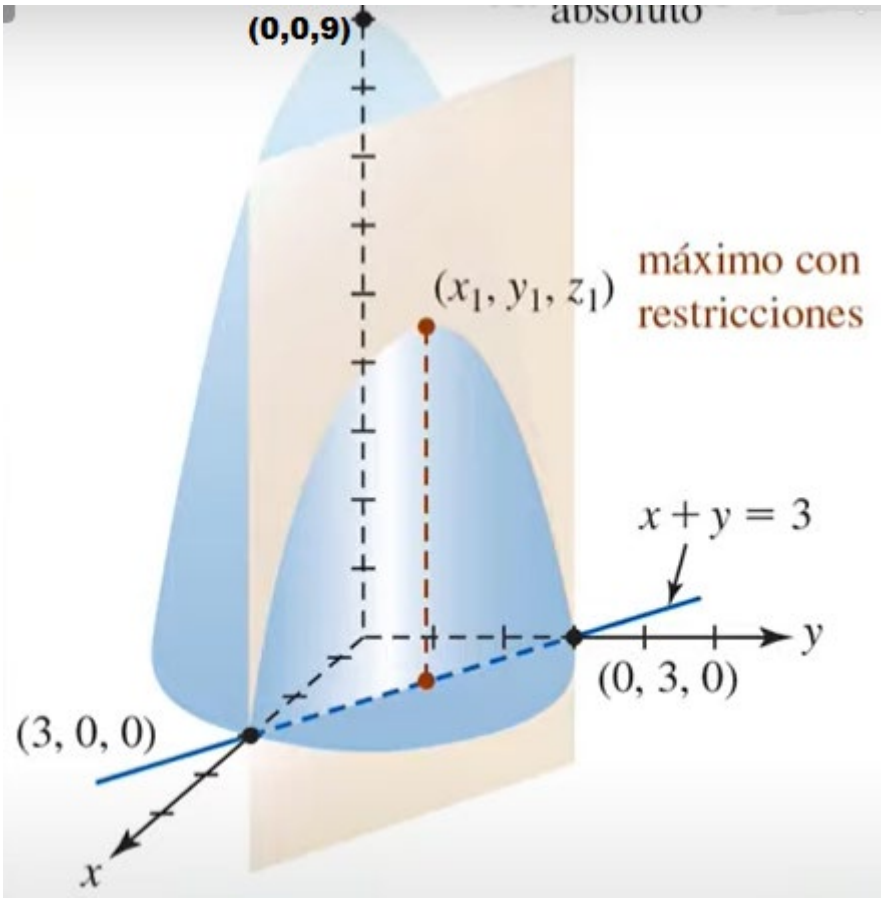

Figura 10

Justo para los casos en los que se buscan los argumentos que hacen máxima una función, pero atendiendo a restricciones, es para lo que se usan los multiplicadores de Lagrange.

La forma más fácil de aplicar los multiplicadores de Lagrange es la siguiente:

Llamemos a la función a maximizar como f y a la función que impone sus restricciones como g. De esta manera tendríamos:

$$f(x,y) = 9 - x^2 - y^2$$

Sujeto a:

$$g(x,y) = x + y - 3$$

La función de Lagrange tiene el siguiente aspecto:

$$L(x, y, \lambda) = f(x, y) - \lambda\, g(x, y)$$ (fórmula 23)

Donde λ son los multiplicadores no negativos de Lagrange. Por tanto, podemos convertir:

$$L(x, y, \lambda) = 9 - x^2 - y^2 - \lambda(x + y - 3)$$

$$L(x, y, \lambda) = 9 - x^2 - y^2 - x\lambda - y\lambda + 3\lambda$$

A continuación, se han calcular las derivadas parciales de cada una de las variables con respecto a la función y se igualan a cero:

$$\begin{cases} \frac{\partial L(x,y,\lambda)}{\partial x} = 0 \\ \frac{\partial L(x,y,\lambda)}{\partial y} = 0 \\ \frac{\partial L(x,y,\lambda)}{\partial \lambda} = 0 \end{cases}$$ (fórmula 24)

Desarrollando las derivadas nos devolverá el siguiente sistema:

$$\begin{cases} -2x - \lambda = 0 \\ -2y - \lambda = 0 \\ -x - y + 3 = 0 \end{cases}$$

Lo que visto de otra forma es:

$$\begin{cases} 2x = -\lambda \\ 2y = -\lambda \\ x + y - 3 = 0 \end{cases}$$

Dado que $2x = -\lambda$ y $2y = -\lambda$, entonces $x = y$. Aprovechemos esta información directa para ver cuánto es x e y sustituyendo cualquiera en la tercera ecuación del sistema:

$$x + x - 3 = 0$$

$$2x = 3$$

$$x = 3/2$$

Si sabemos el valor de x, sabemos también el de y:

$$y = 3/2$$

Por tanto, ya sabemos los argumentos de x e y que hacen máxima la función $f(x, y)$ bajo las restricciones que impone $g(x, y)$. Pueden ser expresados en el vector:

$$(x, y) = \left(\frac{3}{2}, \frac{3}{2}\right)$$

Y el máximo sujeto a las restricciones descritas se obtiene evaluando la función $f(x, y)$ con dichos valores tal que:

$$f\left(\tfrac{3}{2}, \tfrac{3}{2}\right) = 9 - \left(\tfrac{3}{2}\right)^2 - \left(\tfrac{3}{2}\right)^2 = \tfrac{9}{2}$$

3.4.1 Aplicación de multiplicadores de Lagrange en SVM

El ejemplo manejable de los multiplicadores de Lagrange nos ha permitido ver para qué valen y que casos puede resolver. En concreto, cuando hay que encontrar los argumentos que hacen máxima una función, pero con unas restricciones.

En el caso de las Máquinas del Vector Soporte habíamos ya identificado tanto la función a maximizar como las restricciones. En concreto, la función a maximizar era la distancia $d(\mathbf{w}, b) = \frac{2}{\|\mathbf{w}\|}$ y las restricciones eran $\mathrm{y}^{\mathrm{n}} \left(\mathbf{w}^{\mathbf{T}} \mathbf{x}_{\boldsymbol{n}} + \mathrm{b}\right) = 1$. No obstante, justificado por una mayor sencillez en los cálculos el procedimiento documentado en los manuales expresa la

función de distancia como $d(\mathbf{w}, b) = \frac{1}{2} \|\mathbf{w}\|^2$, lo que lo convierte en una tarea de minimización. Este cambio no afecta al resultado.

Al final la se ha de buscar una solución a un problema de minimización que es expresado como:

Minimizar:

$$d(\mathbf{w}, b) = \frac{1}{2} \|\mathbf{w}\|^2 \text{ (fórmula 25)}$$

Con la restricción:

$$\mathrm{y}^{\mathrm{n}} \left(\mathbf{w}^{\mathbf{T}} \mathbf{x}_{\boldsymbol{n}} + \mathrm{b}\right) - 1 \geq 0 \text{ (fórmula 26)}$$

Tomando la función de Lagrange con sus multiplicadores:

$$L(x, y, \lambda) = f(x, y) - \lambda\, g(x, y) \quad \text{(fórmula 27)}$$

Obtenemos:

$$L(\mathbf{w}, b, \boldsymbol{\lambda}) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{n=1}^{N} \lambda_n \left[\mathrm{y}^{\mathrm{n}} \left(\mathbf{w}^{\mathbf{T}} \mathbf{x}_{\boldsymbol{n}} + b\right) - 1\right] \text{ (fórmula 28)}$$

De la misma manera que antes, aplicando el procedimiento de Lagrange, obtenemos el vector de pesos $\mathbf{w}$ y el intercepto b que hacen mínima $d(\mathbf{w}, b)$ con la restricción de discriminar entre clases. Identificados ya el vector $\mathbf{w}$ y la b, se pueden llevar a cabo las predicciones. Bastará con introducir un vector ejemplar $\mathbf{x}^*$ en la ecuación de clasificación original:

$$\boldsymbol{resultado} = \mathbf{w}^{\mathbf{T}} \mathbf{x}^* + \mathrm{b}$$

Y obtener el signo del resultado. Si es positivo, pertenecerá a la clase 1. En caso de que sea negativo, pertenecerá a la clase -1. He aquí la esencia de las Máquinas de Vector Soporte.

4. Conjuntos no separables linealmente

La mayoría de las veces, los ejemplares de una y otra clase no son separables linealmente como los ejemplos anteriores (figura 11).

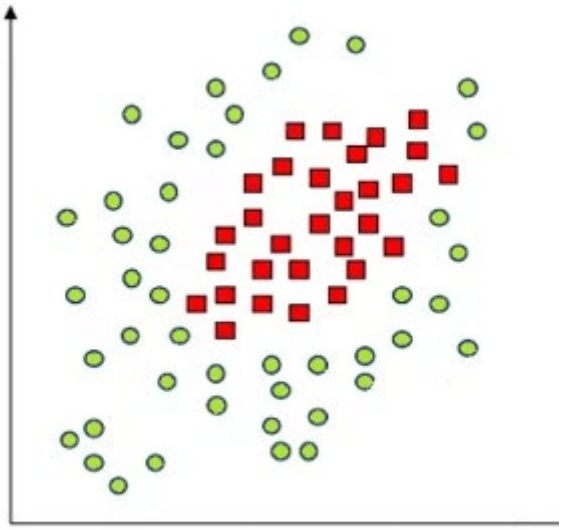

Figura 11. Tomada de https://www.iartificial.net/maquinas-de-vectores-de-soporte-svm/

Para estas ocasiones existe un truco: el uso de los llamados núcleos ("kernels"). Se trata de introducir otra dimensión al espacio donde se expresan los ejemplares. La idea es que, expresando los ejemplares en un espacio de mayor dimensionalidad, hay una mayor probabilidad de que los ejemplares de las clases se vuelvan separables por medio (de nuevo) de un hiperplano, es decir, se vuelvan separables linealmente (figura 12). No obstante, hay que mapear los ejemplares de un espacio de dimensionalidad menor a otro mayor (aunque en rigor nunca se haga), pero sin que los ejemplares queden expresados en un subespacio con la dimensionalidad del menor pero dentro del de mayor dimensionalidad. Si el espacio de menor dimensionalidad es de dos dimensiones, al mapear las coordenadas a uno de tres dimensiones, las coordenadas no quedarán de nuevo en un plano de dos dimensiones (un subespacio de dos dentro del de tres). Al contrario. Las coordenadas han de quedar repartidas convenientemente en uno de tres, es decir, tendrán ya altura. Por tanto, al introducir otra dimensión en el espacio original se debe también introducir en el proceso de optimización un núcleo (a modo de filtro o mapeo de una operación entre ambos espacios) que distribuya convenientemente las

coordenadas en él. Este filtro colocará las coordenadas una ubicación en que se maximice el margen entre los ejemplares de una y otra clase. Esta es la clave. Aquí es dónde se introducen distintos núcleos en el proceso de optimización. Es parecido a cuando a una imagen se le aplica un filtro de “ojo de pez” para poder ver mejor los detalles de alguna zona. El lector intuirá que de lo dicho se sigue que el empleo de un núcleo y la obtención del mejor hiperplano están estrechamente relacionados en el proceso de optimización. De entre los núcleos más populares destacan el lineal, polinomial y gaussiano.

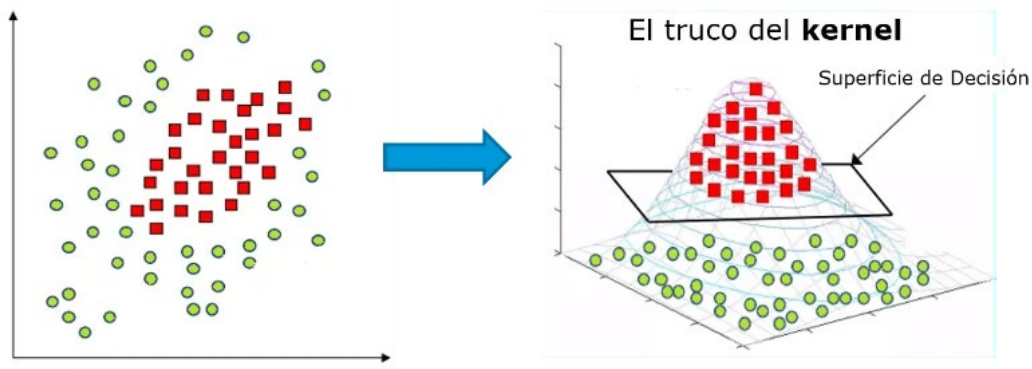

Figura 12. Tomada de https://www.iartificial.net/maquinas-de-vectores-de-soporte-svm/

5. Cuestiones adicionales

En este texto hemos explicado los fundamentos básicos de las Máquinas de Vector Soporte. Entre los aspectos fundamentales hemos estudiado la formalización del hiperplano separador y de la distancia entre vectores soporte. Teniendo dichos formalismos, todo se reducía a una tarea de optimización: encontrar los argumentos que hagan máxima la distancia entre vectores soporte, pero preservando la capacidad discriminativa del hiperplano. Incluso, decíamos, esto podía ser aplicado a conjuntos no linealmente separables en el inicio con el truco del núcleo.

No obstante, hemos ejemplificado el proceso con conjuntos de datos en los que la variable dependiente es binaria. De ahí la homologación con la regresión logística. Sin embargo, si la variable no es binaria, sino multicategorial, también puede aplicarse la técnica de SVM. En general, los paquetes siguen estrategia de "uno contra uno" para trazar distintos hiperplanos de separación entre los ejemplares de cada par de categorías. También se emplea la estrategia "uno frente a los demás", que como se intuye, traza separaciones entre los ejemplares de una categoría y de las demás.

En otro orden de cosas, las redes neuronales profundas actuales tienen una mayor capacidad de aprendizaje y generalización que los SVM, aunque los SVM siguen formando parte del arsenal de técnicas dentro del aprendizaje automático supervisado. Esto es porque hay tareas coyunturales dentro del cauce de datos que pueden ser afrontados por un modelo más sencillo y de propósito específico.

6. Código de ejemplo

https://datagy.io/python-support-vector-machines/

```
# Importing required libraries
from seaborn import load_dataset, pairplot
import matplotlib.pyplot as plt
import pandas as pd
from sklearn.svm import SVC
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score


# Rerunning the algorithm with a binary classifier
```

```
df = load_dataset('penguins')
df = df.dropna()
df = df[df['species'] != 'Adelie']    # This limits us to two classes
print(df)

# X = df.select_dtypes('number')
X = df[['bill_length_mm', 'bill_depth_mm']]
y = df['species']

X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=100)

clf = SVC(kernel='linear')
clf.fit(X_train, y_train)

# Visualizing the linear function for our SVM classifier
import numpy as np
from seaborn import scatterplot
w = clf.coef_[0]
b = clf.intercept_[0]
x_visual = np.linspace(32,57)
y_visual = -(w[0] / w[1]) * x_visual - b / w[1]

scatterplot(data = X_train, x='bill_length_mm', y='bill_depth_mm', hue=y_train)
plt.plot(x_visual, y_visual)
plt.show()

predictions = clf.predict(X_test)
print(predictions[:5])
print(accuracy_score(y_test, predictions)) plt.show()
```

URLs:

https://www.analyticsvidhya.com/blog/2020/10/the-mathematics-behind-svm/

https://www.quora.com/What-are-kernels-in-machine-learning-and-SVM-and-why-do-we-need-them