Support Vector Machine

Jorge Gallego

Facultad de Economía, Universidad del Rosario

Junio 21 de 2017

Introducción

- Estudiaremos otro modelo de aprendizaje supervizado
- Muy utilizado para tareas de clasificación
- Pero que también ha sido usado para predicción numérica
- Support Vector Machine es un método que ha ganado popularidad en los últimos tiempos
- A pesar de ser un modelo de "caja negra"

Introducción

- Las matemáticas del modelo son complejas
- La intuición es simple
- Se busca crear una superficie que divida el espacio que contiene los objetos a clasificar
- De tal forma que cada partición sea lo más homogénea posible
- Dicha superficie, que genera la frontera entre clases, es un hiperplano

Introducción

Algunas aplicaciones recientes de SVM incluyen:

1. Predicción del sentimiento de tweets sobre el proceso de paz en Colombia
2. Clasificación de información genética para predecir cáncer y otras enfermedades
3. Categorización de texto para tareas como identificación del lenguaje de un documento o clasificación según tema
4. Detección de eventos raros, como falla de un motor, brechas de seguridad o terremotos

Clasificación con hiperplanos

- Pensemos en objetos que queremos clasificar. Por ejemplo tweets, células o deudores
- Podemos representarlos como vectores en el espacio n-dimensional
- Cada característica es una dimensión
- Un hiperplano es un subespacio de dimensión $n - 1$, que usaremos para dividir el espacio al que pertenecen los vectores
- En una dimensión, un hiperplano es un punto; en dos, una recta; en tres, un plano. Y así sucesivamente

Clasificación con hiperplanos

- En general, un hiperplano de un espacio n-dimensional puede escribirse como:

$$a_1x_1 + a_2x_2 + \cdots a_nx_n = b$$

- De esta forma el espacio se divide en dos, según:

$$a_1x_1 + a_2x_2 + \cdots a_nx_n > b$$
$$a_1x_1 + a_2x_2 + \cdots a_nx_n < b$$

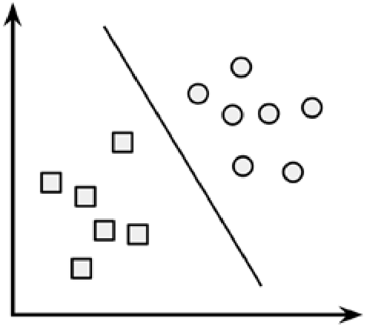
- Así un hiperplano crea fronteras entre los objetos en un espacio

Clasificación con hiperplanos

- Naturalmente, en un espacio euclídeo *n*-dimensional hay infinitos hiperplanos separadores
- Pero buscamos un hiperplano "óptimo"
- Aquel que mejor divida los objetos que queremos clasificar
- Veremos primero el caso en el que los objetos son *linealmente separables*
- Consideremos el siguiente ejemplo

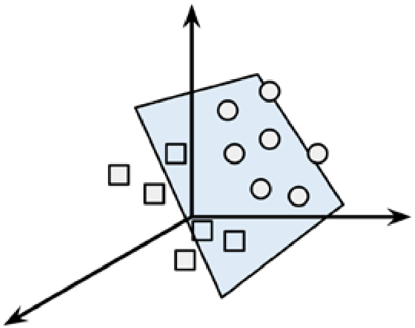
Clasificación con hiperplanos

Figure: Hiperplano en 2-Dimensiones



Clasificación con hiperplanos

Figure: Hiperplano en 3-Dimensiones

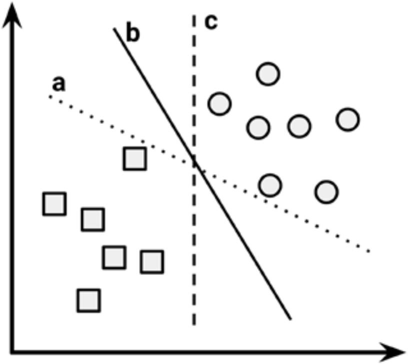


Clasificación con hiperplanos

- Estos son ejemplos de objetos linealmente separables
- Las categorías quedan perfectamente divididas con el hiperplano
- Sin embargo, existen varios hiperplanos que podrían funcionar
- ¿Cómo escoger el más adecuado?

Clasificación con hiperplanos

Figure: Posibles Hiperplanos



Hiperplano de Máximo Margen

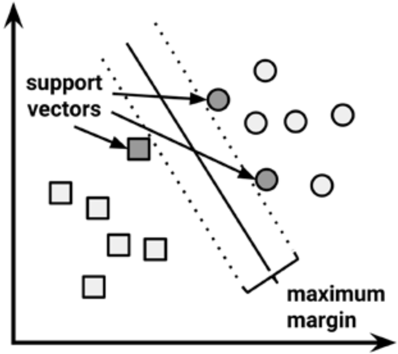
- ¿Cuál elegimos?
- Buscamos el hiperplano que genere el máximo margen (*Maximum Margin Hyperplane*–MMH)
- Es decir, el que genere la mayor separación entre las clases
- Esto es deseable porque así es mayor la probabilidad de que objetos futuros sean clasificados correctamente
- En el ejemplo anterior, esto lo logramos con la recta B

Vectores de Soporte

- Los vectores de soporte (*support vectors*) son los puntos de cada clase que están más cerca del MMH
- Cada clase tiene al menos un *support vector*
- Pero puede haber más de uno
- Así, los *support vectors* definen el MMH y por ende la forma de clasificar

Vectores de Soporte

Figure: *Support Vectors*



Vectores de Soporte

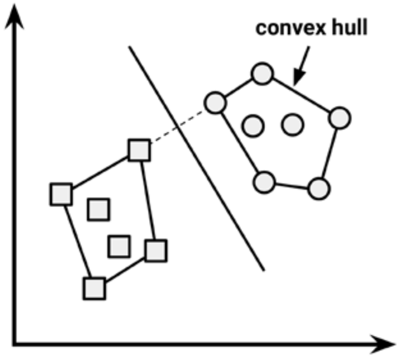
- ¿Cómo encontramos matemáticamente los vectores de soporte y por ende los MMH?
- Con herramientas de geometría vectorial
- Todo depende de si los datos son linealmente separables o no
- Veamos inicialmente el primer caso, que es más sencillo

Datos Linealmente Separables

- Si los datos son linealmente separables el concepto de envolvente convexa es clave
- La envolvente convexa son los puntos que representan la frontera exterior de un conjunto
- En este caso nos interesan las envolventes convexas de cada categoría
- Veámoslo gráficamente con el ejemplo anterior

Envolvente Convexa

Figure: *Envolvente Convexa*



Datos Linealmente Separables

- El MMH es el bisector perpendicular de la línea más corta que une las dos envolventes
- De esta forma definimos al hiperplano separador óptimo
- No haremos esto manualmente ni a ojo. Algoritmos computacionales identifican el máximo margen de esta forma
- Se basan en una técnica llamada optimización cuadrática

Datos Linealmente Separables

- Existe un método alternativo
- Consiste en buscar el par de hiperplanos paralelos que dividan en grupos homogéneos los puntos
- De tal forma que los dos planos estén lo más separados que sea posible
- Formalicemos este proceso

Datos Linealmente Separables

- Podemos definir el hiperplano como:

$$\mathbf{w} \cdot \mathbf{x} + b = 0$$

- donde $\mathbf{w} = \{w_1, w_2, \ldots, w_n\}$
- $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ es el vector de características
- b es un escalar que llamamos el sesgo. Es el intercepto

Datos Linealmente Separables

- Un ejemplo de un hiperplano es una recta en dos dimensiones
- La ecuación de una recta es $y = mx + b$
- Ajustando el intercepto, b, este es un ejemplo de la especificación que tenemos en la diapositiva anterior
- Así, SVM tiene elementos similares a los de los modelos de regresión pues la idea es buscar los parámetros óptimos

Datos Linealmente Separables

- De esta manera, buscamos los pesos que especifiquen dos hiperplanos tales que:

$$\mathbf{w} \cdot \mathbf{x} + b > +1$$
$$\mathbf{w} \cdot \mathbf{x} + b < -1$$

- Estos son los planos paralelos
- Pero además queremos que estén lo más separados posible

Datos Linealmente Separables

- Usando geometría vectorial, la distancia entre los dos planos está dada por:

$$\frac{2}{||\mathbf{w}||}$$

- donde $||\mathbf{w}||$ es la norma euclídea del vector $\mathbf{w}$
- Luego para maximizar la distancia, debemos minimizar $\mathbf{w}$

Datos Linealmente Separables

- El proceso de optimización para encontrar el hiperplano separador es

$$\begin{aligned} \min \quad & \frac{1}{2}||\mathbf{w}||^2 \\ \text{s.t} \quad & y_i(\mathbf{w}\mathbf{x}_i - b) \geq 1, \forall \mathbf{x}_i \end{aligned}$$

- Minimizamos la norma (elevada al cuadrado y multiplicada por $1/2$ por simple conveniencia)
- Sujeto a que cada punto sea correctamente clasificado de acuerdo a su y_i. Nótese que y_i es 1 o -1

Datos Linealmente Separables

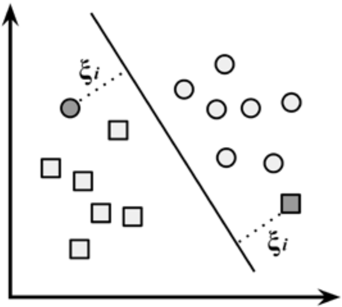
- Luego, si el objeto es clase $y_1 = 1$, debe estar en la región $\mathbf{wx}_i - b \geq 1$
- Si es $y_1 = -1$, debe estar en $\mathbf{wx}_i - b \leq -1$
- Nuevamente, este problema de optimización se resuelve con optimización cuadrática
- Se lo dejamos al computador

Datos que no son Linealmente Separables

- ¿Qué hacer cuando los datos no son linealmente separables?
- En este caso introducimos una variable de holgura (*slack variable*)
- Esta variable nos permite que algunos puntos caigan en el lado incorrecto de la frontera
- La llamamos ξ_i para observación

Datos que no son Linealmente Separables

Figure: *Slack Variable*



Datos que no son Linealmente Separables

- Para la optimización introducimos un costo
- Que penaliza cada punto que viola la restricción
- El problema de optimización se vuelve

$$
\begin{aligned}
\min \quad & \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n}\xi_i \\
\text{s.t} \quad & y_i(\mathbf{w}\mathbf{x}_i - b) \geq 1 - \xi_i, \forall \mathbf{x}_i, \xi_i \geq 0
\end{aligned}
$$

Métodos Kernel

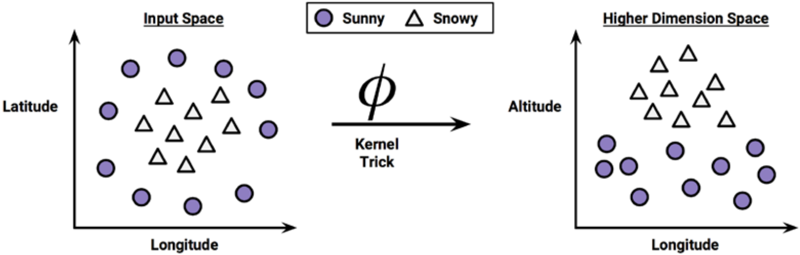
- Existe otro método para dar cuenta de estas no-linealidades
- Los métodos kernel
- La esencia es transformar el problema a una dimensión superior. En la que si puede haber separación lineal
- Este método se conoce como el truco kernel

Métodos Kernel

- Consideremos el siguiente ejemplo
- Queremos clasificar ciertos lugares como soleados o nevosos
- Y los caracterizamos en dos dimensiones: latitud y longitud
- En estas dos dimensiones no hay separabilidad lineal

Métodos Kernel

Figure: *Métodos Kernel*



Métodos Kernel

- Pero puede que haya una tercera dimensión que permita ver los objetos de otra forma
- Por ejemplo, la altura. Lugares nublados podría estar más arriba que los soleados
- Lo que el kernel hace es construir nuevas características a partir de las que podemos medir

Métodos Kernel

- En este caso, la altura podría ser la interacción entre la latitud y la longitud
- Cuánto más cerca esté un punto al centro de la escala latitud-longitud, mayor altura tendrá
- Así, el método consiste en generar características no incluidas originalmente en los datos
- A partir de ciertas transformaciones
- Alguna de ellas puede permitir que los datos sean linealmente separables

Métodos Kernel

- Una función kernel aplica una transformación a un conjunto de características
- En general, para dos vectores de características $\mathbf{x}_i$ y $\mathbf{x}_j$, una transformación tipo kernel es de la forma

$$K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$$

- donde ϕ es la función kernel que usemos

Métodos Kernel

- Existen diversas funciones utilizadas en la práctica
- Kernel lineal, polinomial, sigmoidal, gaussiana RBF
- ¿Cuál usar? Aquí la práctica es muy de ensayo y error
- La que mejor ajuste logre
- Por algo este método (SVM) es considerado de "caja negra"