UNIVERSIDAD
NACIONAL
DE COLOMBIA

Analítica Predictiva

Carlos A. Madrigal

Profesor Ocasional

Departamento de Ciencias de la Computación y de la Decisión

Maestría en ingeniería - ingeniería de sistemas

Maestría en ingeniería - analítica

Especialización en sistemas

Nota: Este material se ha adaptado con base a diferentes fuentes de información académica

Contenido

Técnicas de Clasificación y Agrupamiento

- Naive Bayes
- KNN
- K-means
- Máquina de Soporte Vectorial

Naive Bayes

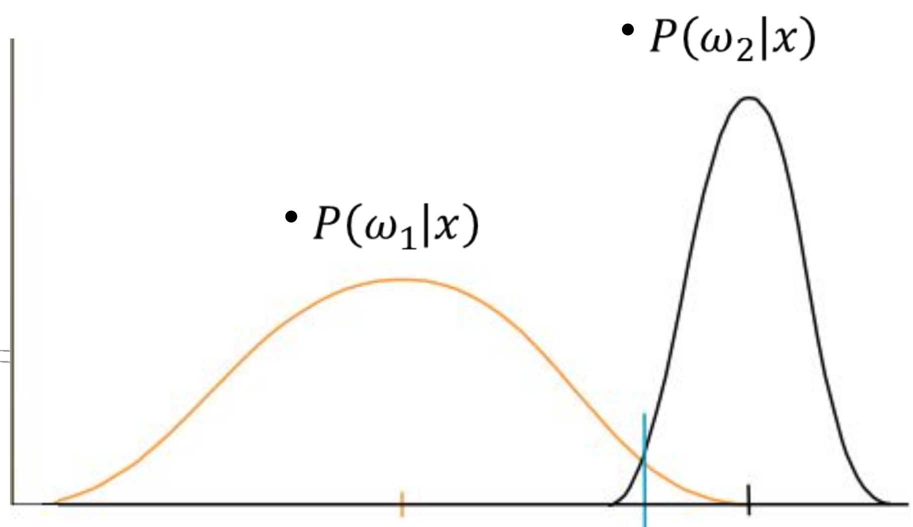
El teorema de Bayes expresa la probabilidad a posteriori de un evento Y dado X. Usar la teoría de la probabilidad para clasificar el objeto en la clase que tenga mayor probabilidad posteriori.

$$P(\omega_i|x) = \frac{p(x|\omega_i)P(\omega_i)}{p(x)}$$

- $P(\omega_i)$ = Probabilidad de que en la población haya un objeto de clase ω_i

  $p(x|\omega_i)$ = Probabilidad de que en la clase ω_i se de un vector de características x

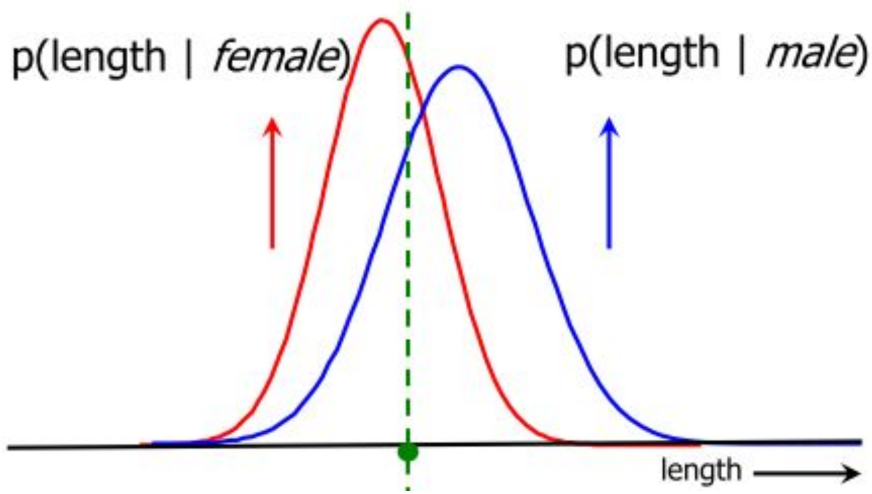
  $P(\omega_i|x)$ = Probabilidad de que el objeto de vector de características x pertenezca a la clase ω_i



- $g(x) = \begin{cases} 1 & si \quad P(\omega_1|x) > P(\omega_2|x) \\ 2 & en\ otro\ caso \end{cases}$

Naive Bayes

- Cuál es el género de alguien con esta altura?



$$\text{Bayes:} \begin{cases} p(\textit{female} \mid \text{length}) = p(\text{length} \mid \textit{female})\, p(\textit{female}) \,/\, \cancel{p(\text{length})} \\ p(\textit{male} \mid \text{length}) = p(\text{length} \mid \textit{male})\, p(\textit{male}) \,/\, \cancel{p(\text{length})} \end{cases}$$

Naive Bayes

- Regla de Clasificación de Bayes:

$$p(\textit{female} \mid \text{length}) \quad > \quad p(\textit{male} \mid \text{length}) \quad \rightarrow \quad \textit{female} \text{ else } \textit{male}$$

<u>Bayes:</u>

$$\frac{p(\text{length} \mid \textit{female})\, p(\textit{female})}{p(\text{length})} \quad > \quad \frac{p(\text{length} \mid \textit{male})\, p(\textit{male})}{p(\text{length})}$$

$$p(\text{length} \mid \textit{female})\, p(\textit{female}) > p(\text{length} \mid \textit{male})\, p(\textit{male}) \rightarrow \textit{female} \text{ else } \textit{male}$$

pdf estimated from training set

class prior probabilities
known, guessed or estimated

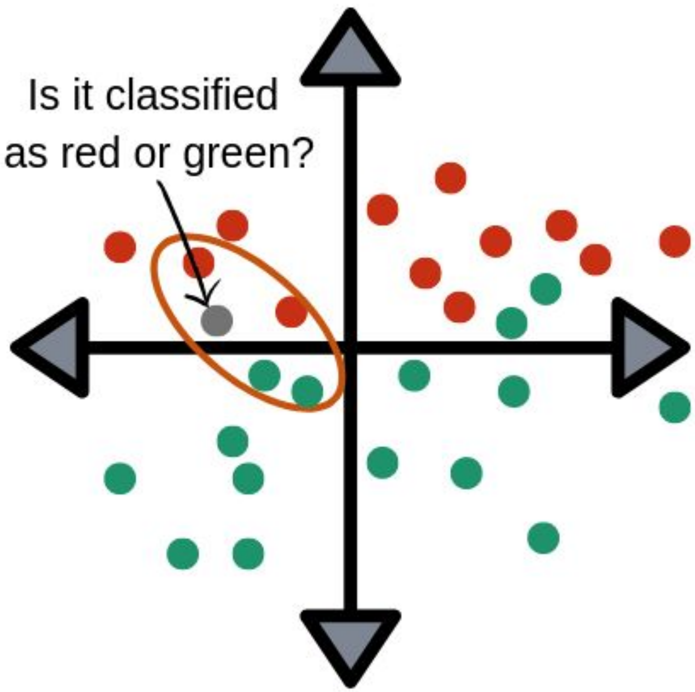
Gaussian Naive Bayes

Cuando las características X son no categóricas, la verosimilitud se calcula como la densidad de probabilidad de una distribución normal.

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Clasificador KNN

La idea básica sobre la que se fundamenta este paradigma es que un nuevo caso se va a clasificar en la clase más frecuente a la que pertenecen sus K vecinos más cercanos



COMIENZO

Entrada: $D = \{(\mathbf{x}_1, c_1), \ldots, (\mathbf{x}_N, c_N)\}$

$\mathbf{x} = (x_1, \ldots, x_n)$ nuevo caso a clasificar

PARA todo objeto ya clasificado (x_i, c_i)

calcular $d_i = d(\mathbf{x}_i, \mathbf{x})$

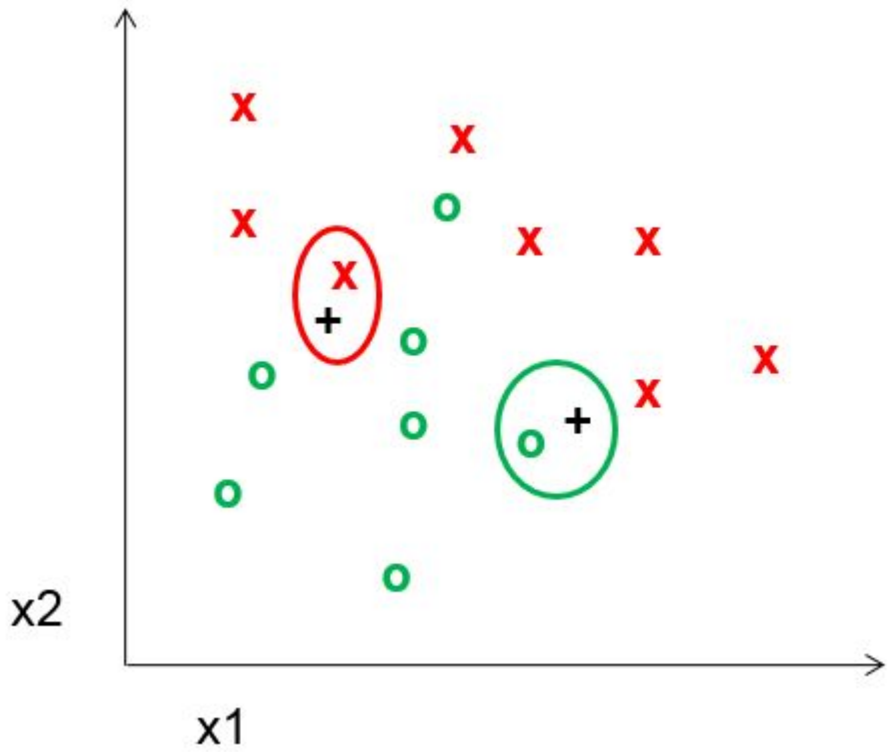
Ordenar $d_i (i = 1, \ldots, N)$ en orden ascendente

Quedarnos con los K casos $D_{\mathbf{x}}^K$ ya clasificados más cercanos a $\mathbf{x}$

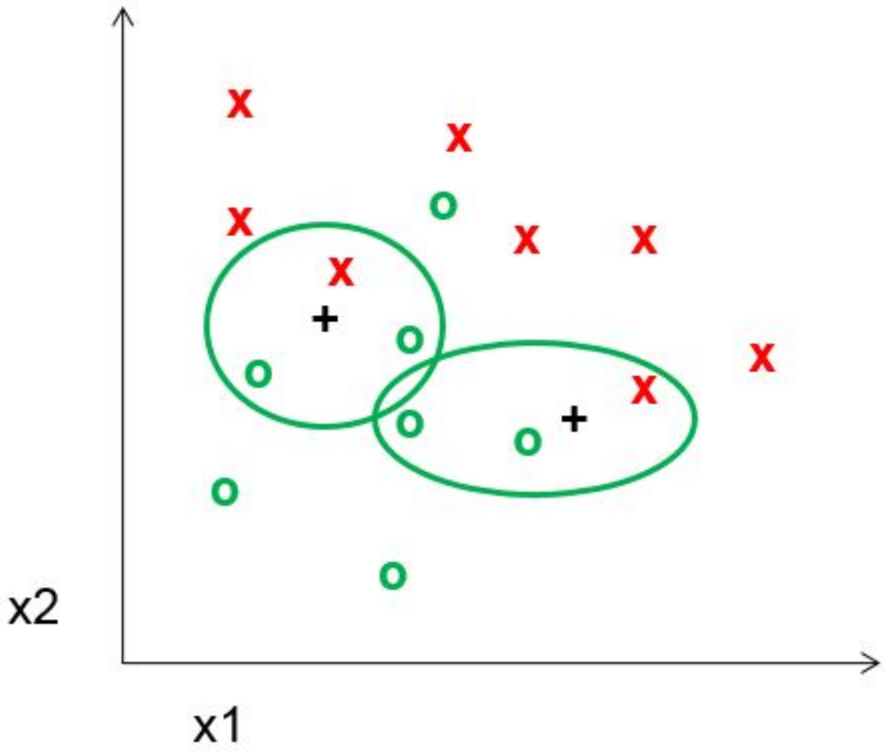
Asignar a $\mathbf{x}$ la clase más frecuente en $D_{\mathbf{x}}^K$

FIN

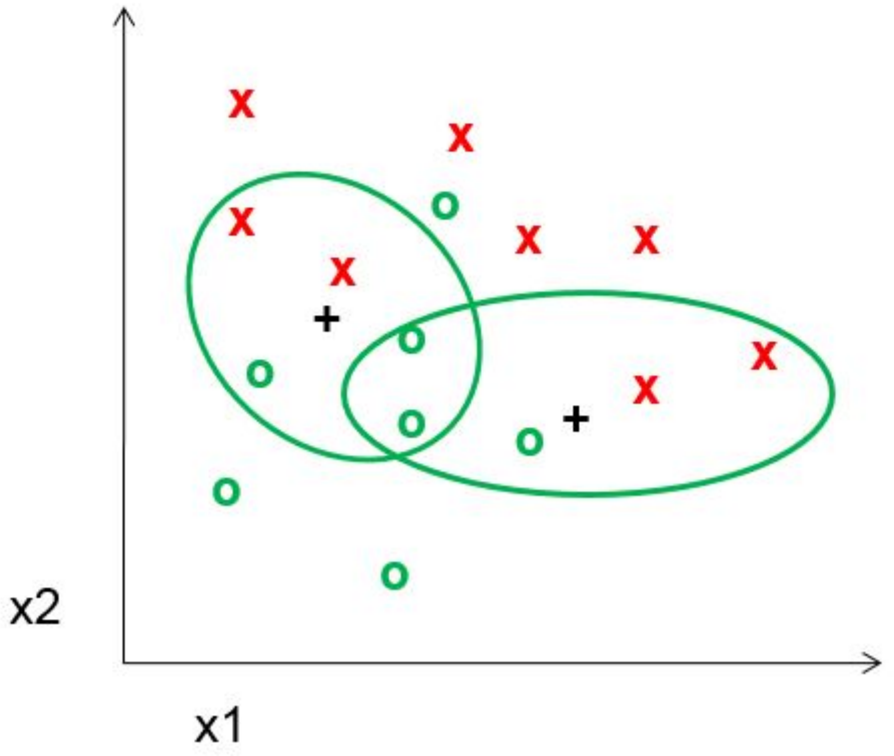
Clasificador KNN – 1 NN



Clasificador KNN – 3 NN



Clasificador KNN – 5 NN



Variaciones KNN

KNN- con Rechazo

Esta variación hace énfasis en las garantías que se deben dar para asignar una clase a un conjunto de características.

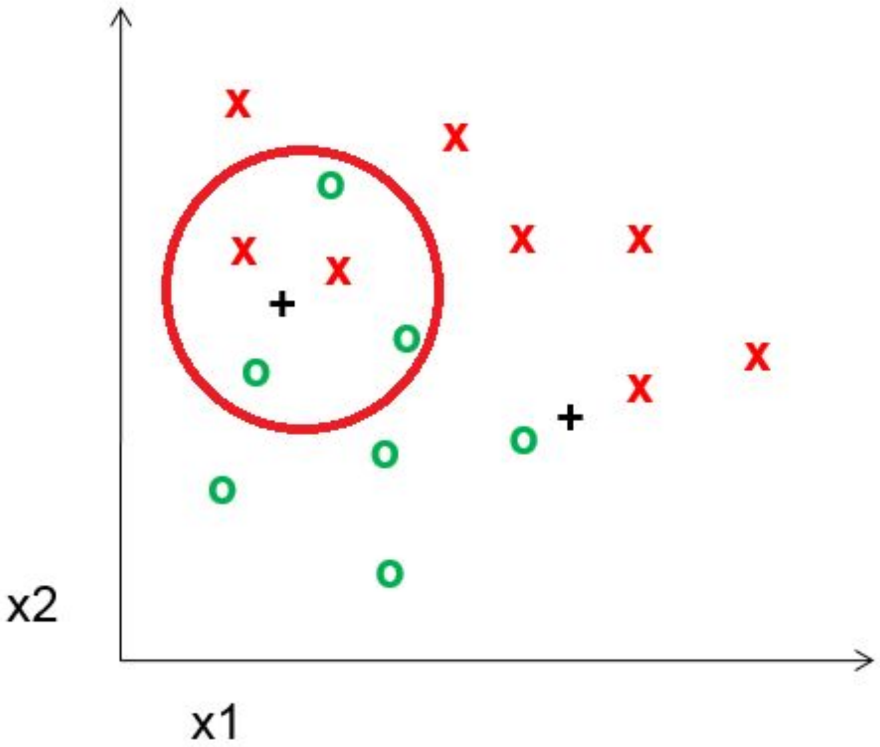
Umbral: Se refiere a que el número de votos para la clase a asignar sea superior a este valor. Ejemplo: si K=8, m=2, el umbral podría establecerse en 5 o 6.

Mayoría Absoluta: Diferencias entre la frecuencia mayor y segunda mayor supere un valor. Ejemplo: Siendo K=15, m=3, Diferencia= 3;

Clasificador KNN

KNN – con distancia Media

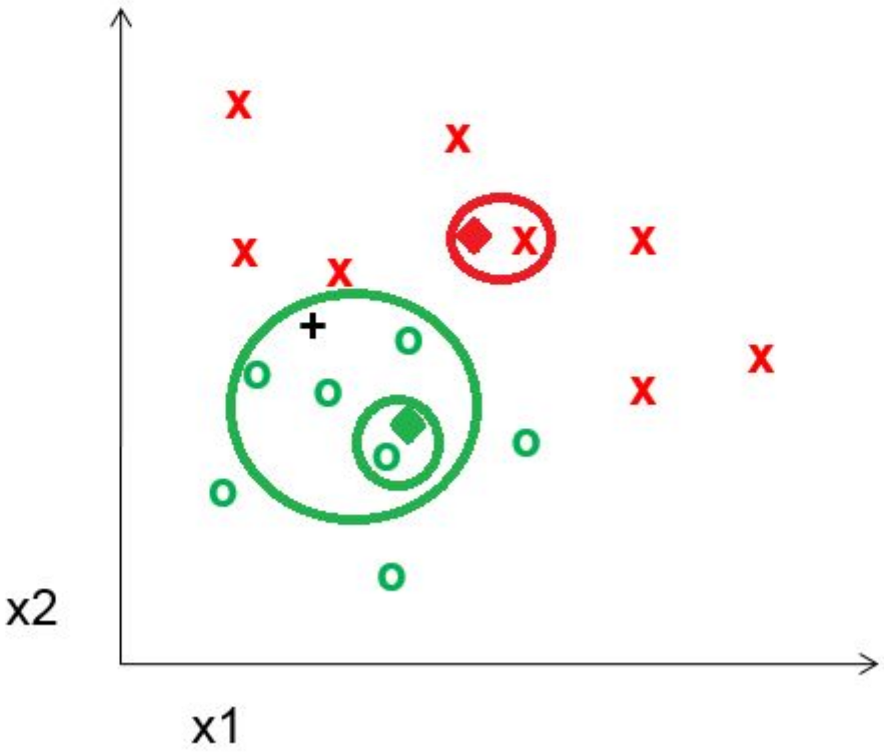
Hace referencia a la asignación de la clase con menor distancia media



Clasificador KNN

KNN – con distancia Mínima.

Se reduce el número de casos a uno por clase (baricentro) y luego se le asigna al conjunto de características la clase del baricentro más cercano.



Clasificador KNN

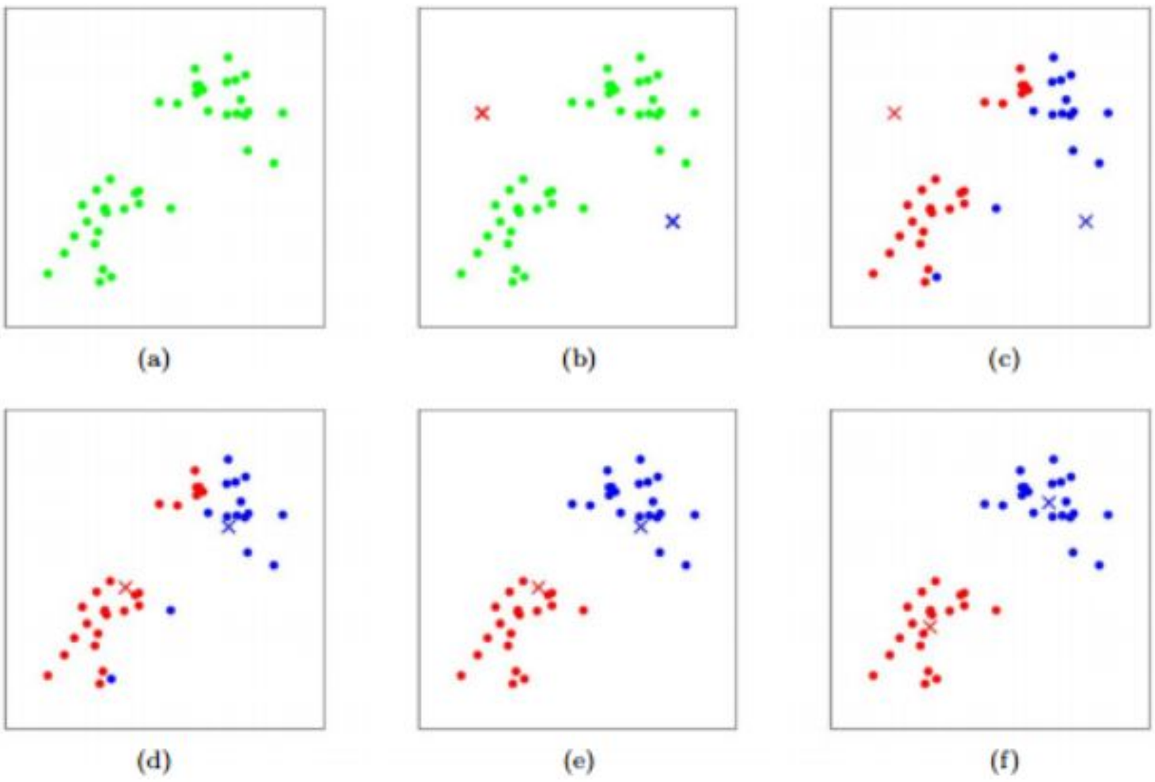
KNN – con pesado de Casos.

Hace referencia a no contabilización homogénea de los casos, sino que se genera un peso para contabilizar cada caso, por ejemplo, el inverso de la distancia entre los casos seleccionados y el nuevo caso.

$$d(\mathbf{x}, \mathbf{x}_r) = \sum_{j=1}^{n} w_j (x_j, x_{rj})^2$$

KMEANS

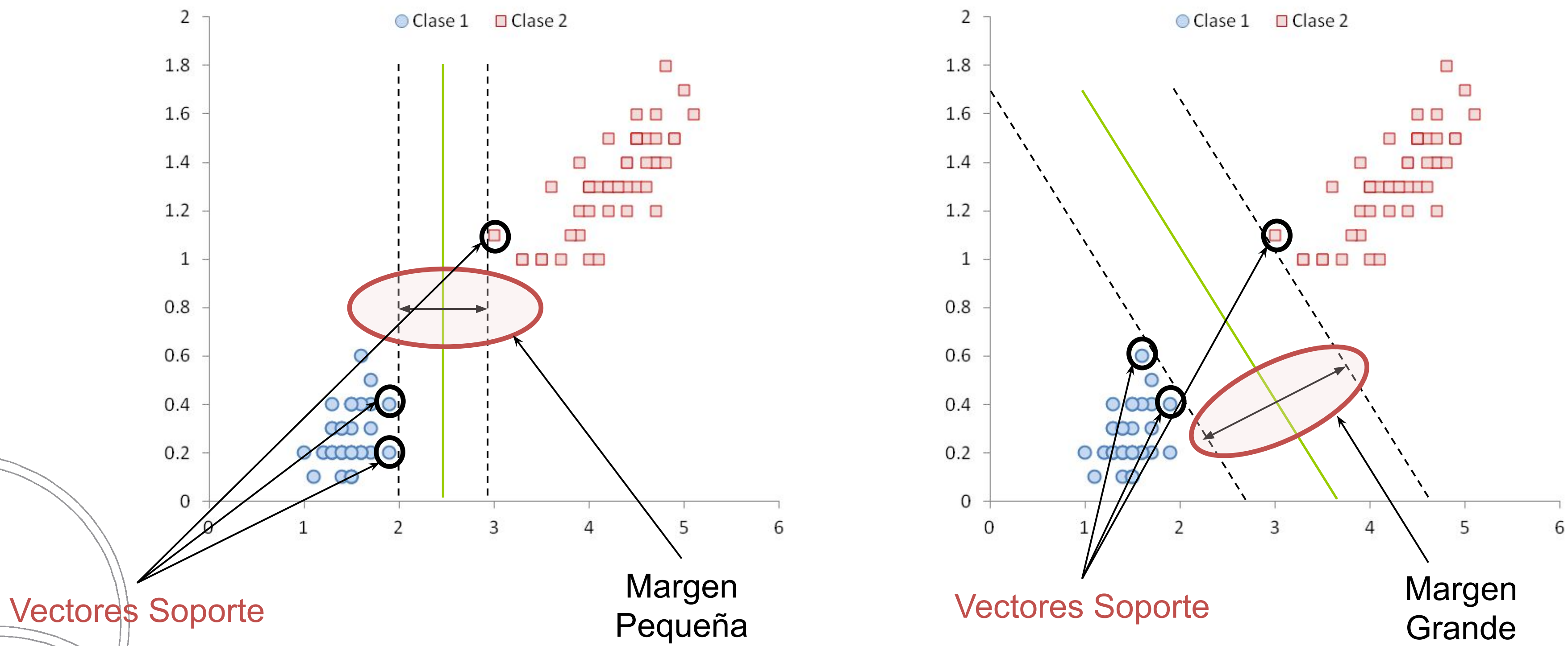
Permiten hacer agrupaciones entre los datos de tal manera que los casos de un cluster tengan una alta similaridad entre ellos y baja con respecto a casos de otro cluster.



https://stanford.edu/~cpiech/cs221/handouts/kmeans.html

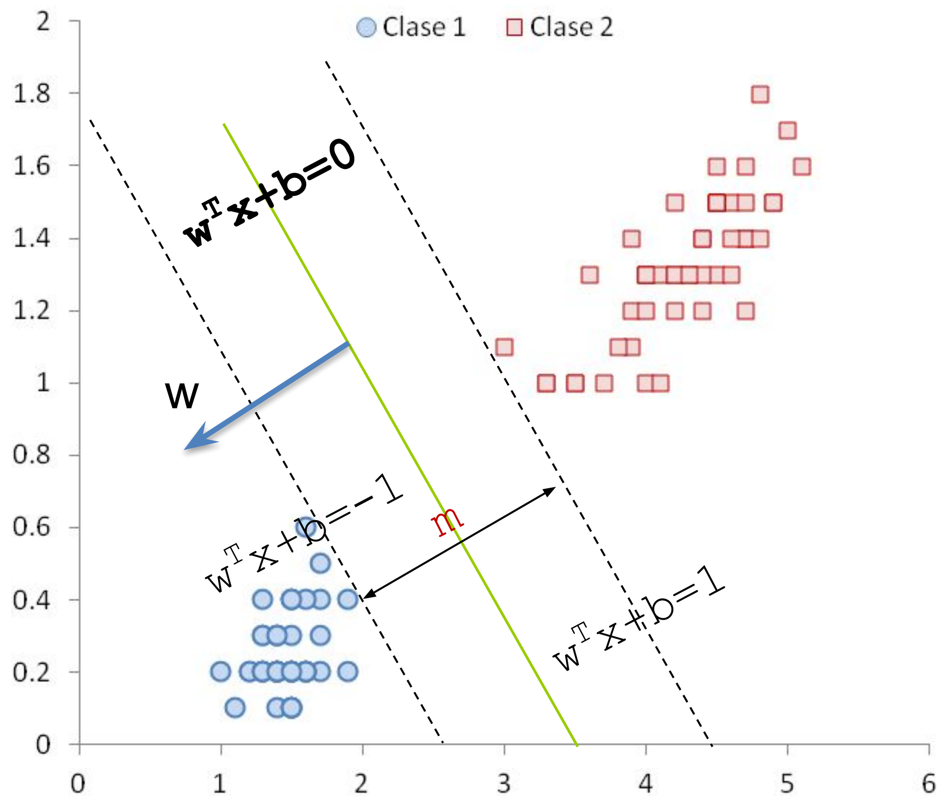
Máquinas de Soporte Vectorial

- Las **SVM** son un tipo de clasificadores de patrones basados en técnicas estadísticas de aprendizaje y están a la cabeza de los métodos de clasificación por permitir construir fronteras de decisión flexibles, y su buena capacidad de generalización.



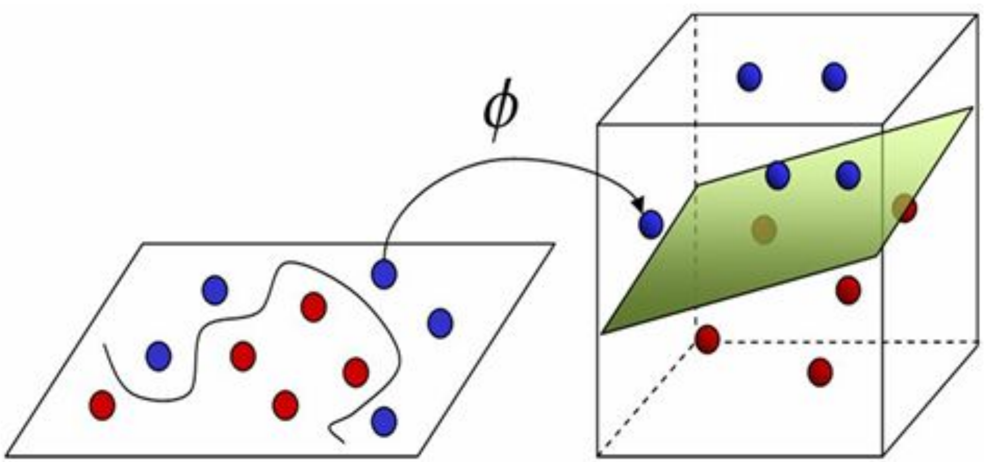
Máquinas de Soporte Vectorial

- Clasificación Lineal: Las SVM generan un hiperplano que separa el espacio en dos o más regiones, una para cada clase.



Máquinas de Soporte Vectorial

- La Clasificación NO Lineal con una SVM realiza una transformación del espacio de entrada a otro de dimensión más alta, en el que los datos son separables linealmente.



Lineal: $K(x_i, x_j) = x_i \cdot x_j$

Polinómico: $K(x_i, x_j) = (x_i \cdot x_j + 1)^d$

Gausiano: $K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$

$$\tilde{L}(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$$

- Al introducir un kernel, los parámetros α del vector w se calculan así:

Ejemplos

Naive Bayes https://www.kaggle.com/dilip990/spam-ham-detection-using-naive-bayes-classifier

KNN https://www.kaggle.com/jmataya/k-nearest-neighbors-classifier

SVM https://www.kaggle.com/migeruj/svm-sentiment-analysis-an-lisis-de-sentimientos

Kmeans https://www.kaggle.com/karthickaravindan/k-means-clustering-project
https://www.kaggle.com/gabrielrs3/clustering-customers-k-means-algorithm/data

Preguntas

UNIVERSIDAD
NACIONAL
DE COLOMBIA