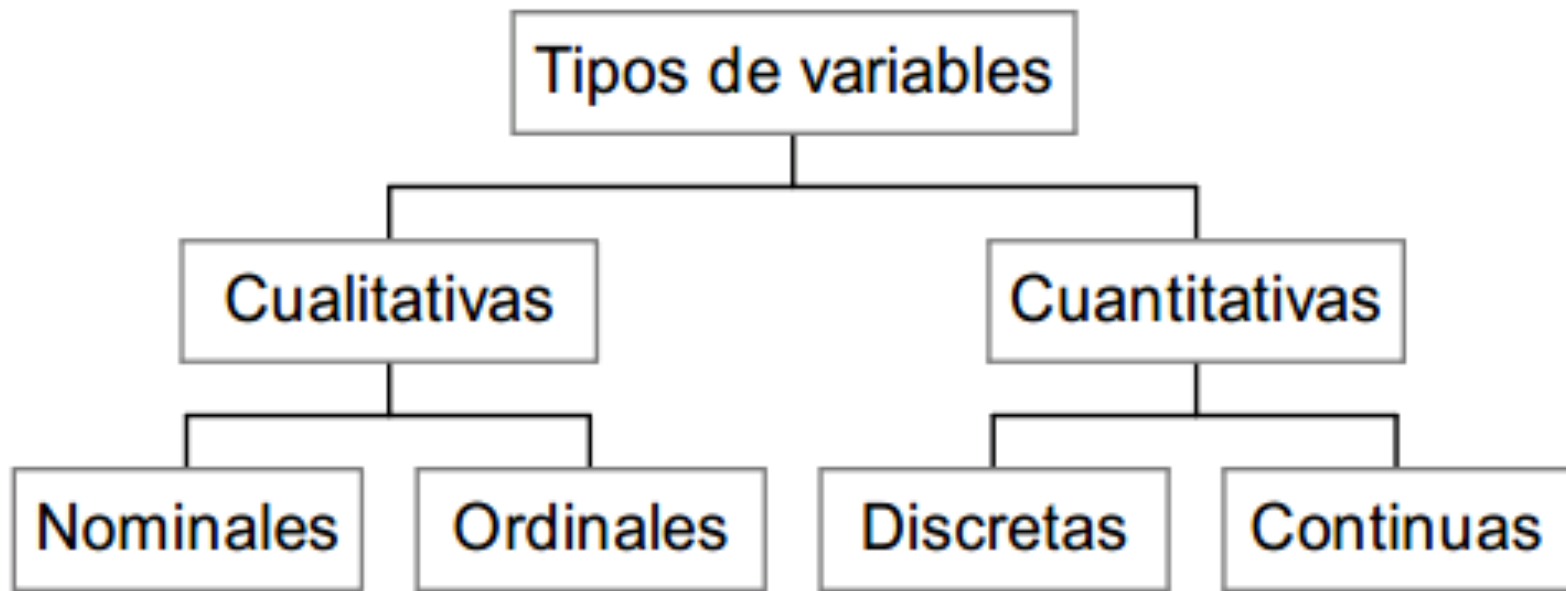


Máquinas de Soporte Vectorial

Tipos de Variables



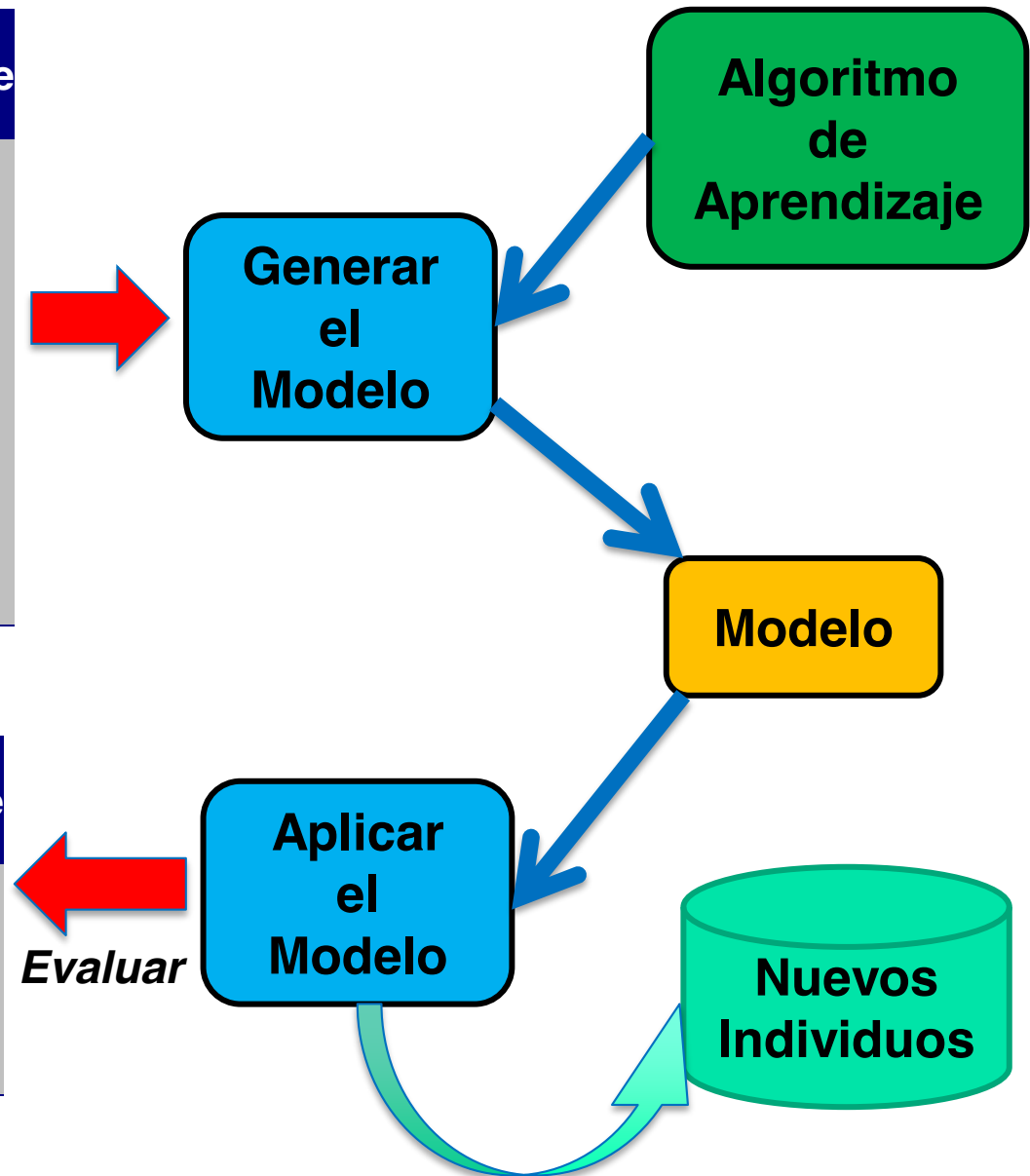
Modelo general de los métodos de Clasificación

<i>Id</i>	Reembolso	Estado Civil	Ingresos Anuales	Fraude
1	Sí	Soltero	125K	No
2	No	Casado	100K	No
3	No	Soltero	70K	No
4	Sí	Casado	120K	No
5	No	Divorciado	95K	Sí
6	No	Casado	60K	No

Tabla de Aprendizaje

<i>Id</i>	Reembolso	Estado Civil	Ingresos Anuales	Fraude
7	No	Soltero	80K	No
8	Si	Casado	100K	No
9	No	Soltero	70K	No

Tabla de Testing



Clasificación: Definición

- Dada una colección de registros (conjunto de entrenamiento) cada registro contiene un conjunto de variables (atributos) denominado x , con un variable (atributo) adicional que es la clase denominada y .
- El objetivo de la ***clasificación*** es encontrar un modelo (una función o algortimo) para predecir la clase a la que pertenecería cada registro, esta asignación una clase se debe hacer con la mayor precisión posible.
- Un conjunto de prueba (tabla de testing) se utiliza para determinar la precisión del modelo. Por lo general, el conjunto de datos dado se divide en dos conjuntos al azar de el de entrenamiento y el de prueba.

Definición de Clasificación

- Dada una base de datos $D = \{t_1, t_2, \dots, t_n\}$ de tuplas o registros (individuos) y un conjunto de clases $C = \{C_1, C_2, \dots, C_m\}$, el **problema de la clasificación** es encontrar una función $f: D \rightarrow C$ tal que cada t_i es asignada una clase C_j .
- $f: D \rightarrow C$ podría ser una Red Neuronal, un Árbol de Decisión, un modelo basado en Análisis Discriminante, o una Red Bayesiana.

Ejemplo: Créditos en un Banco

Tabla de Aprendizaje

Variable
Discriminante

OLDEMARRR.DMEx...ditoViviendaPeq							
	Id	MontoCredito	IngresoNeto	CoeficienteCre...	MontoCuota	GradoAcademico	BuenPagador
►	1	2	4	3	1	4	1
	2	2	3	2	1	4	1
	3	4	1	1	4	2	2
	4	1	4	3	1	4	1
	5	3	3	1	3	2	2
	6	3	4	3	1	4	1
	7	4	2	1	3	2	2
	8	4	1	3	3	2	2
	9	3	4	3	1	3	1
	10	1	3	2	2	4	1
*	NULL	NULL	NULL	NULL	NULL	NULL	NULL

Con la Tabla de Aprendizaje se entrena (aprende) el modelo matemático de predicción, es decir, a partir de esta tabla se calcula la función f de la definición anterior.

Ejemplo: Créditos en un Banco

Tabla de Testing

Variable
Discriminante

OLDEMARRR.DME...iviendaPeqPRED		OLDEMARRR.DMEx...ditoViviendaPeq					
	Id	MontoCredito	IngresoNeto	CoficienteCre...	MontoCuota	GradoAcademico	BuenPagador
►	11	3	3	3	3	1	2
	12	2	2	2	2	1	1
	13	2	2	3	2	1	1
	14	1	3	4	3	2	2
	15	1	2	4	2	1	1
*	NULL	NULL	NULL	NULL	NULL	NULL	NULL

- Con la Tabla de Testing se valida el modelo matemático de predicción, es decir, se verifica que los resultados en individuos que no participaron en la construcción del modelo es bueno o aceptable.
- Algunas veces, sobre todo cuando hay pocos datos, se utiliza la Tabla de Aprendizaje también como de Tabla Testing.

Ejemplo: Créditos en un Banco

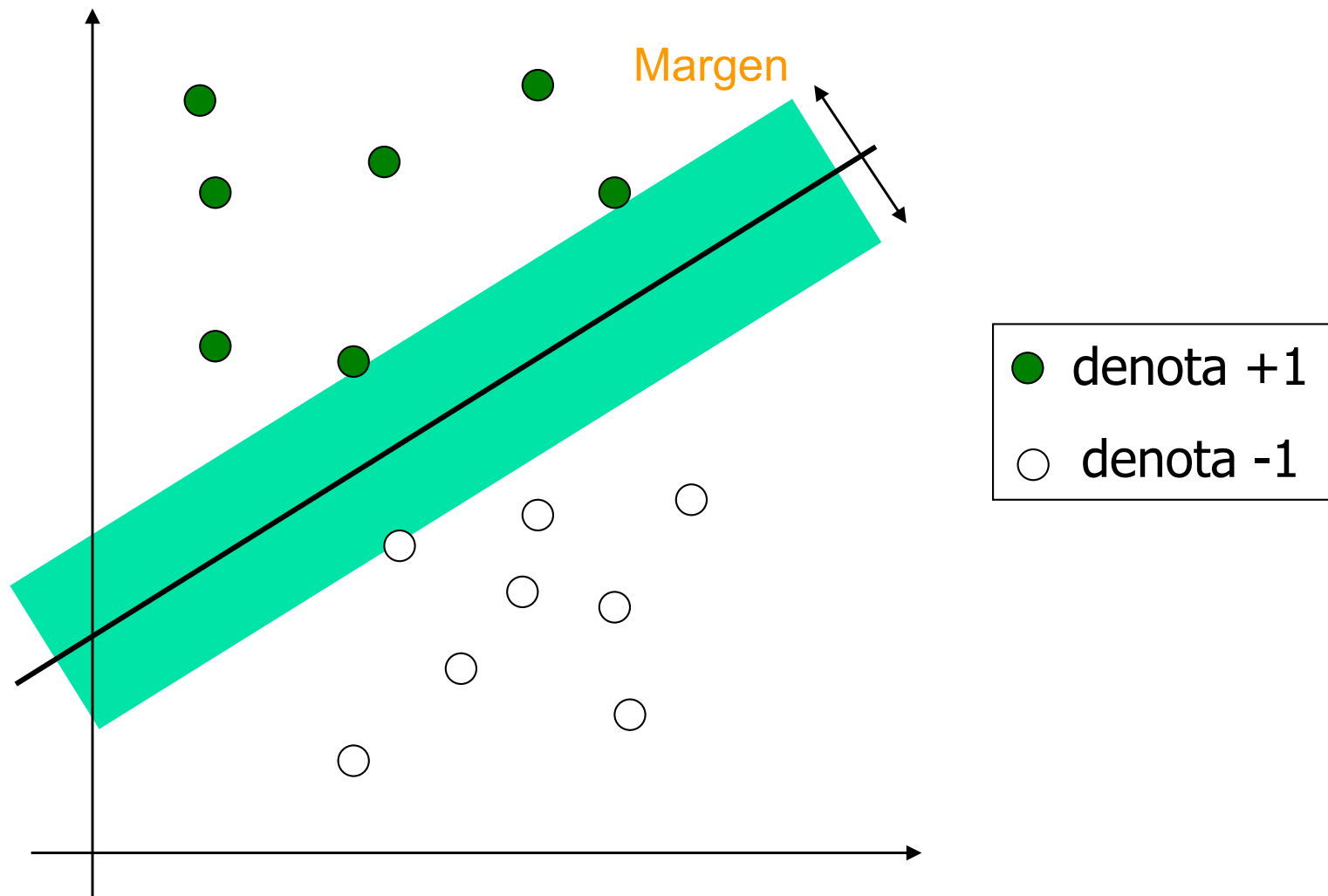
Nuevos Individuos

Variable
Discriminante

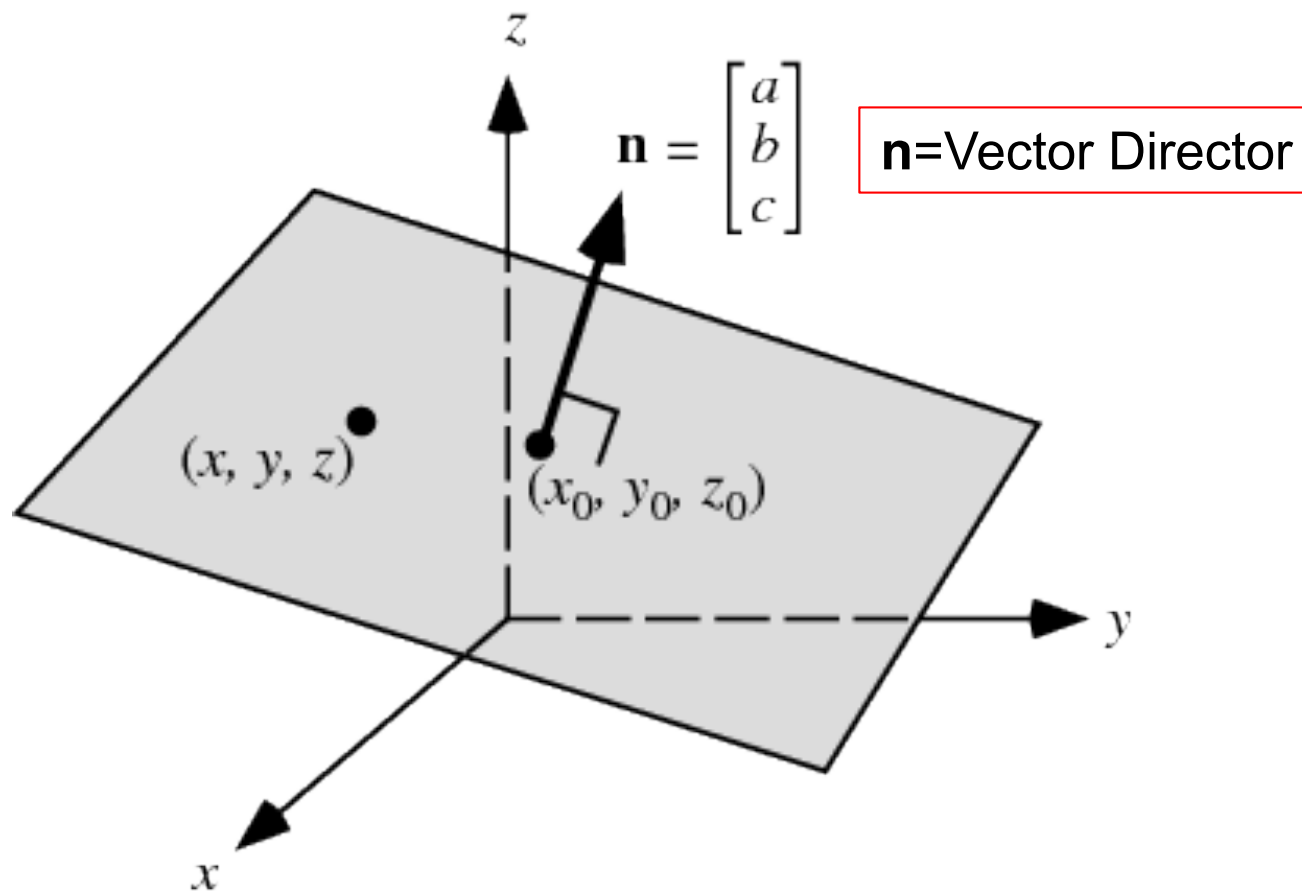
OLDEMARRR.DMEx ...editoViviendaNI							
	Id	MontoCredito	IngresoNeto	CoeficienteCre...	MontoCuota	GradoAcademico	BuenPagador
	100	4	4	2	2	3	?
	101	1	4	3	2	4	?
	102	3	2	3	4	2	?
►*	NULL	NULL	NULL	NULL	NULL	NULL	NULL

Con la Tabla de Nuevos Individuos se predice si estos serán o no buenos pagadores.

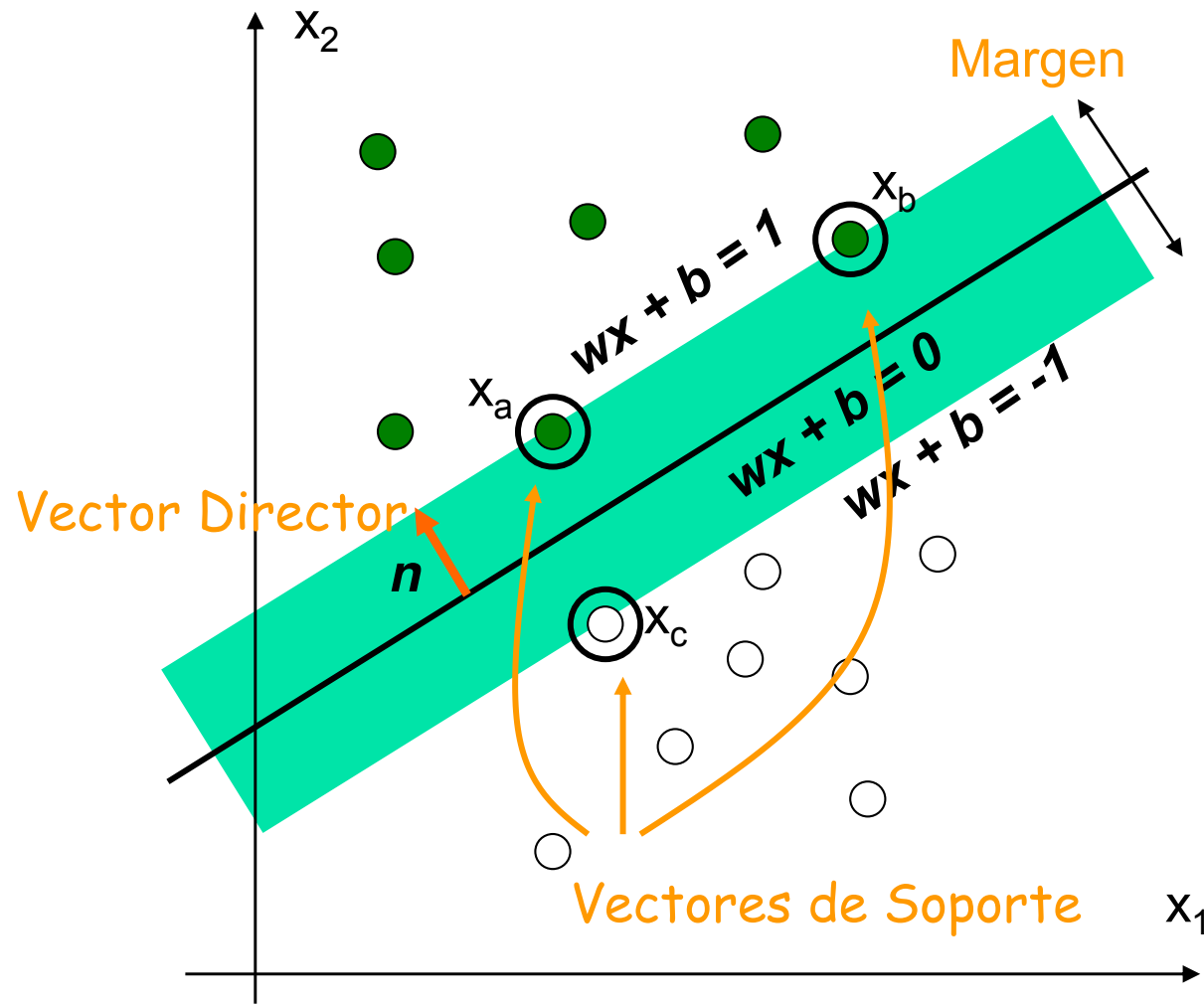
*Idea: Las Máquinas de Soporte Vectorial (Support Vector Machines) tratan de encontrar el **hiperplano** que separe a las clases con el mayor “margen” posible.*



¿Por qué se denominan *Máquinas de Soporte Vectorial* (Support Vector Machines)?



¿Por qué se denominan Máquinas de Soporte Vectorial (Support Vector Machines)?



Función discriminante lineal

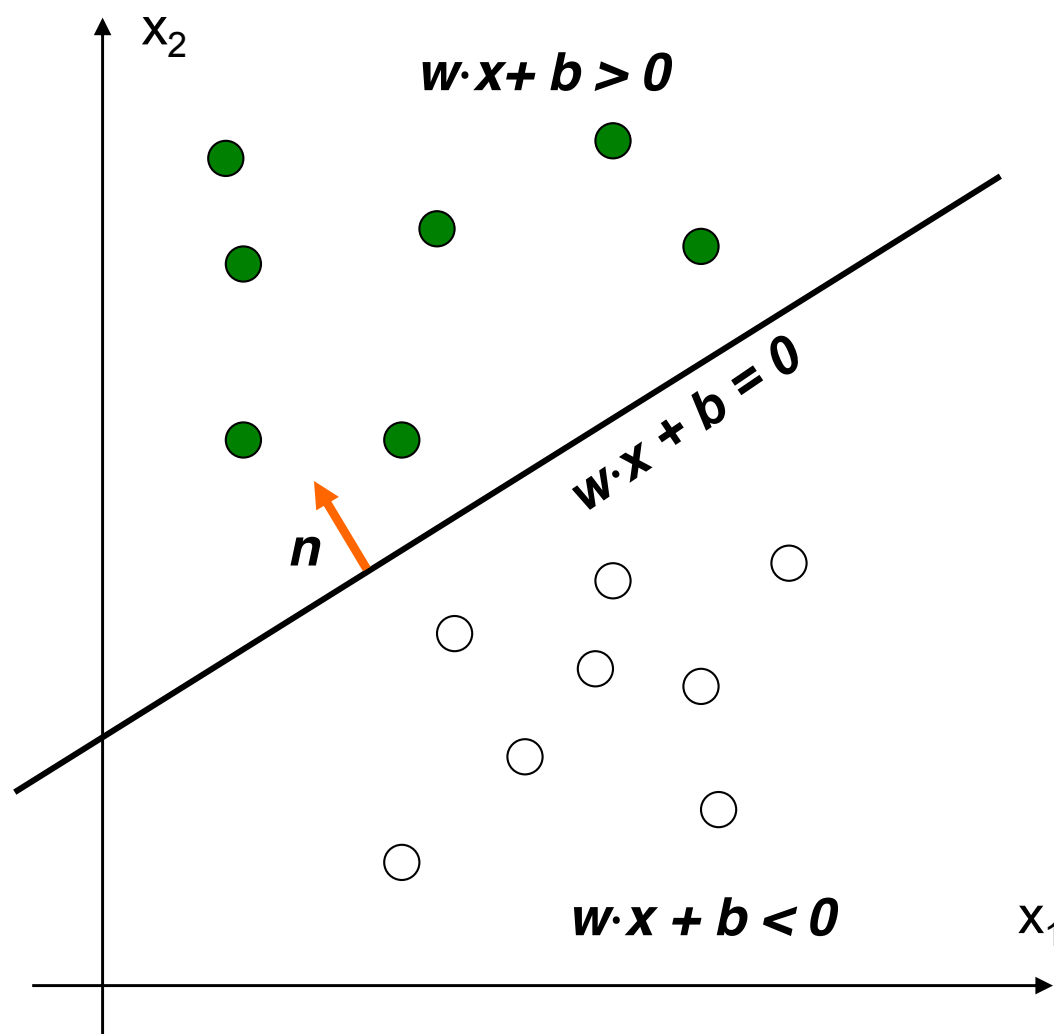
- $g(x)$ es una función lineal:

$$g(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$$

- Se busca un hiperplano en el espacio de las variables

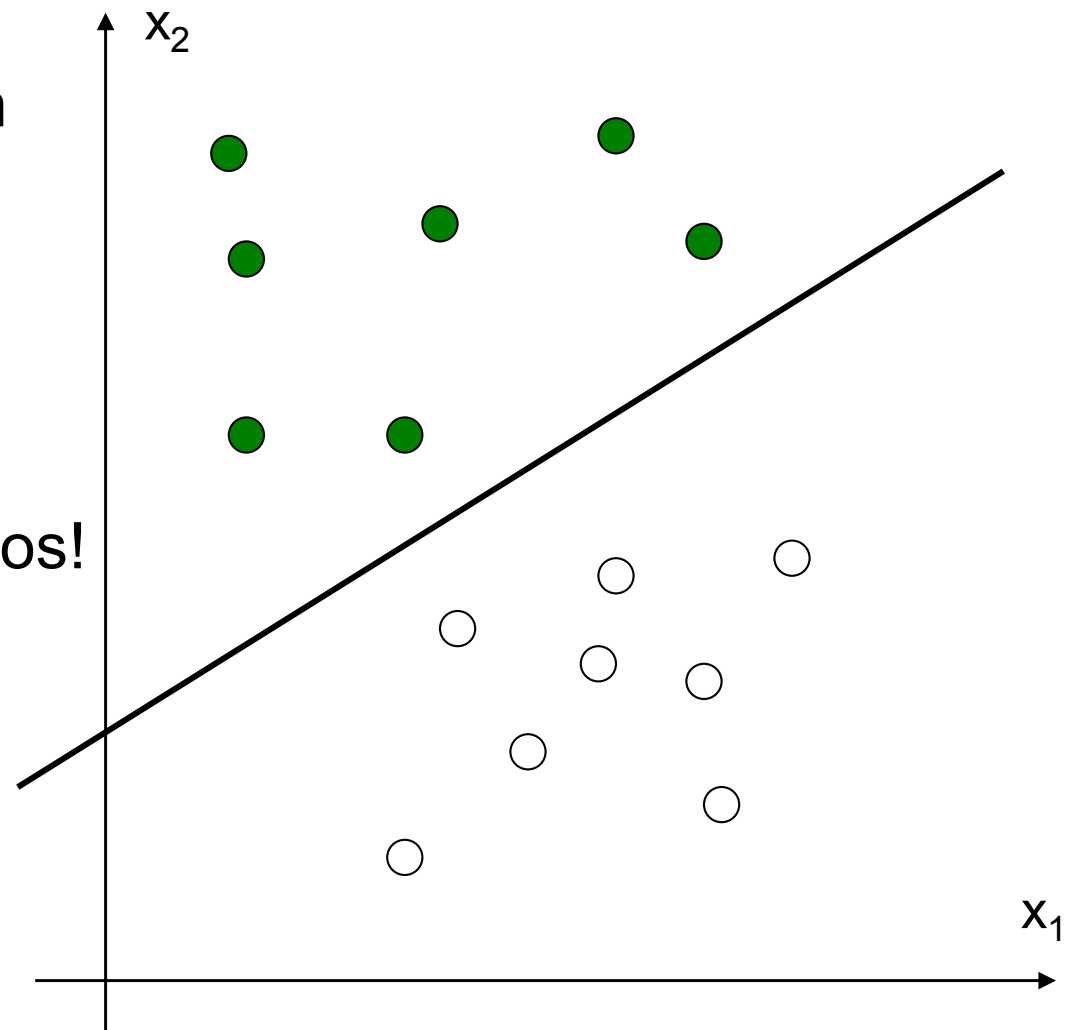
- \mathbf{n} es el vector normal del hiperplano

$$\mathbf{n} = \frac{\mathbf{w}}{\|\mathbf{w}\|}$$



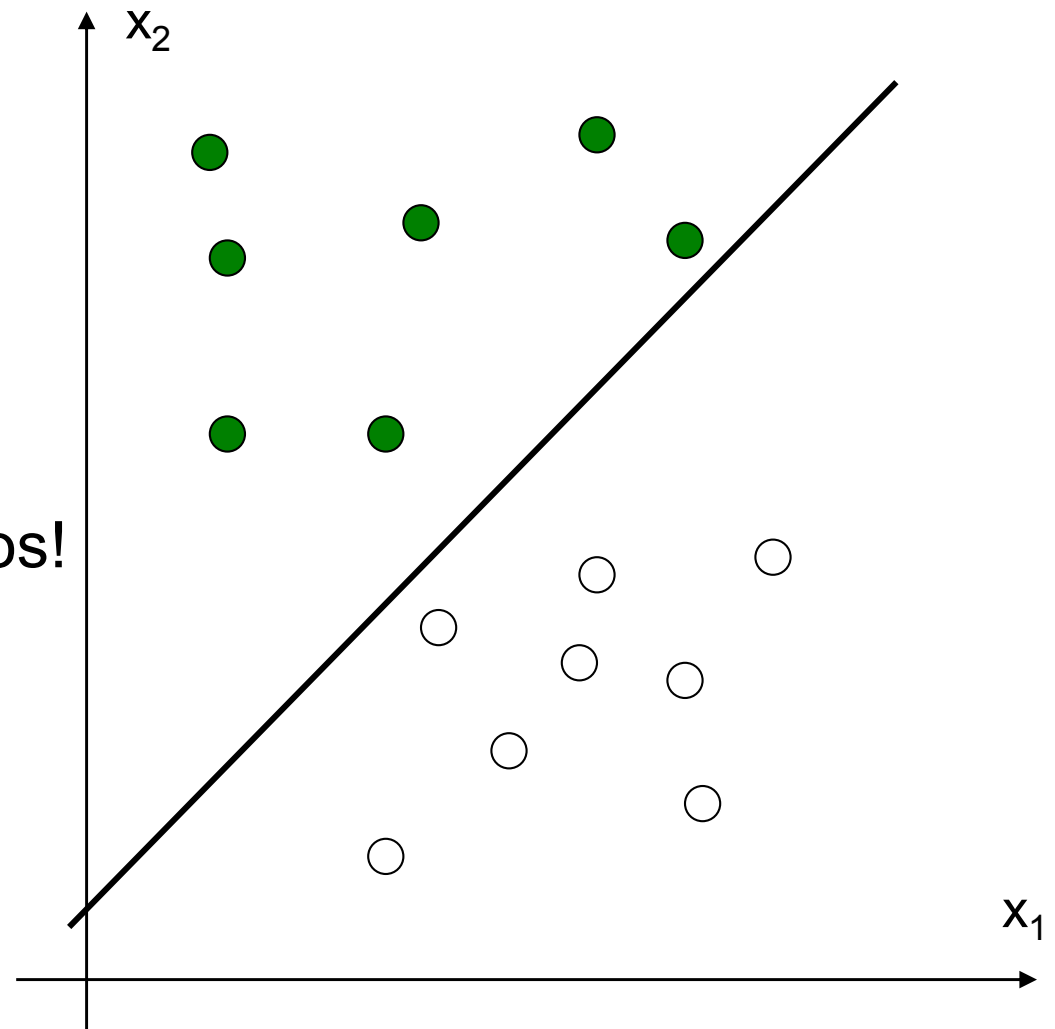
Función discriminante lineal

- ¿Cómo clasificar estos puntos mediante una función discriminante lineal reduciendo al mínimo el error?
- Podrían existir una cantidad infinita de posibles hiperplanos!



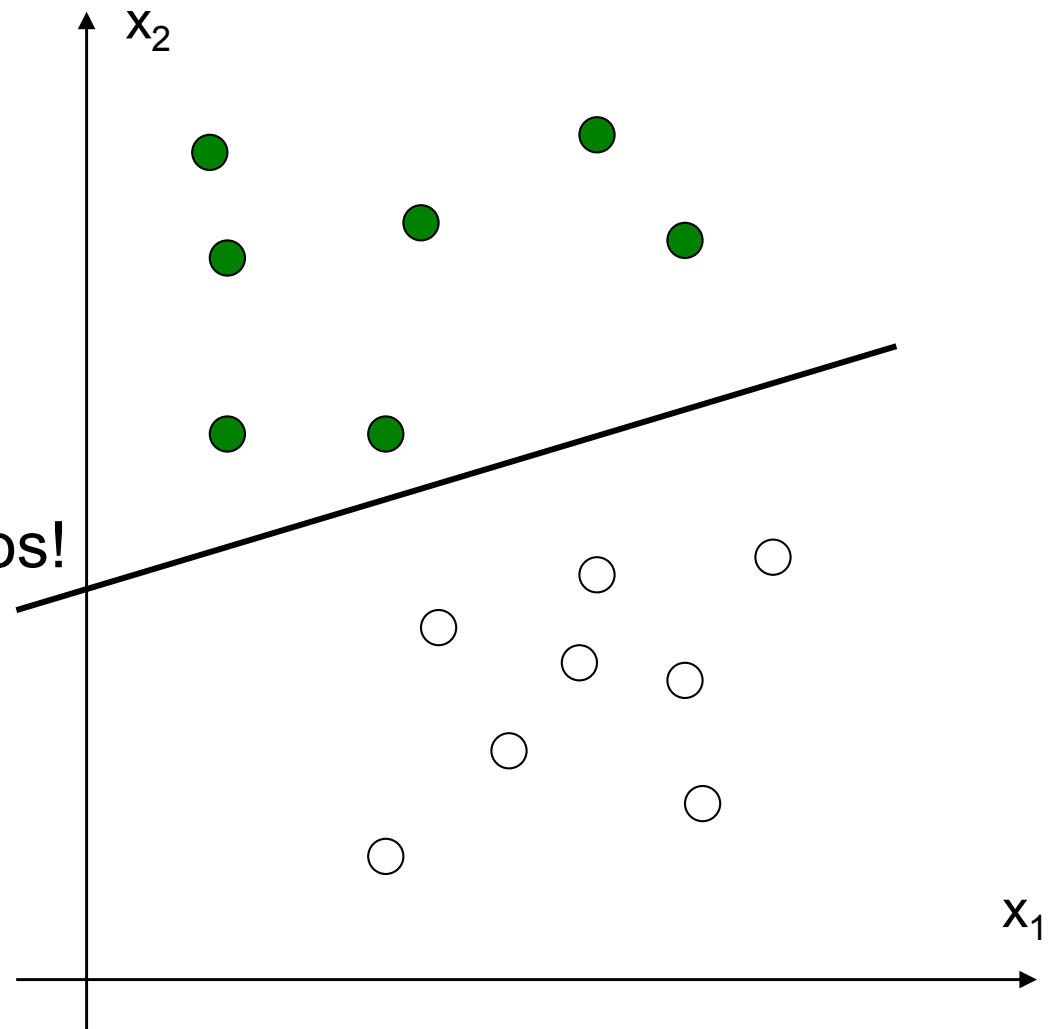
Función discriminante lineal

- ¿Cómo clasificar estos puntos mediante una función discriminante lineal reduciendo al mínimo el error?
- Podrían existir una cantidad infinita de posibles hiperplanos!



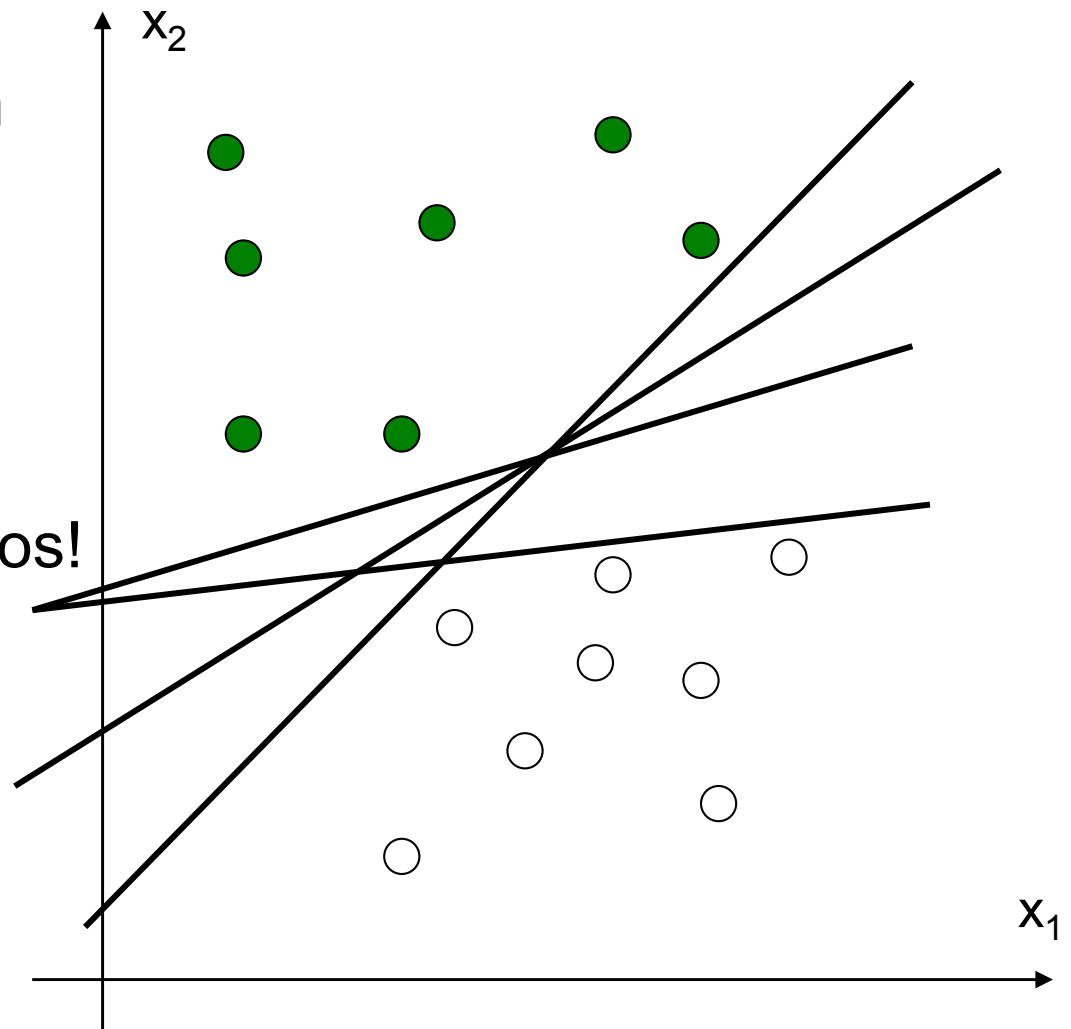
Función discriminante lineal

- ¿Cómo clasificar estos puntos mediante una función discriminante lineal reduciendo al mínimo el error?
- Podrían existir una cantidad infinita de posibles hiperplanos!



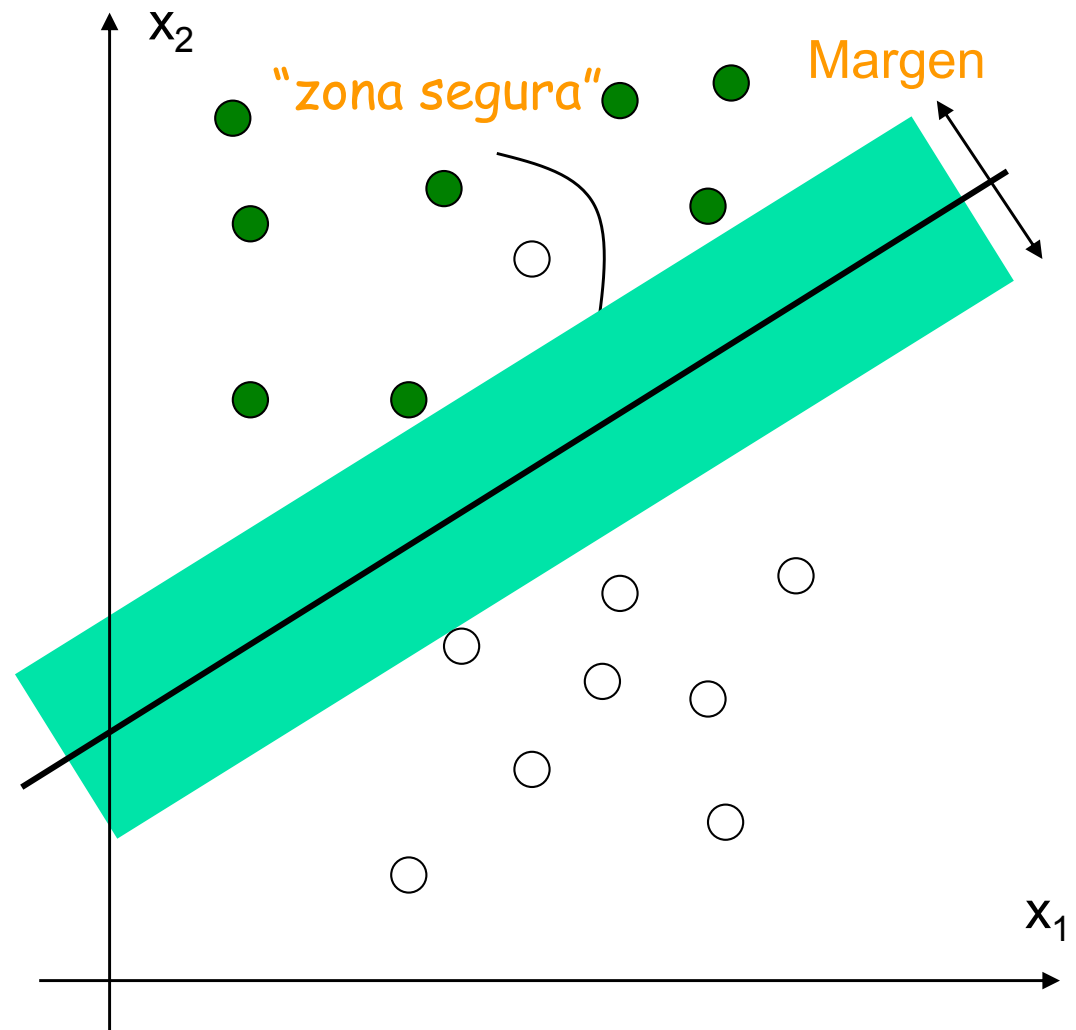
Función discriminante lineal

- ¿Cómo clasificar estos puntos mediante una función discriminante lineal reduciendo al mínimo el error?
- Podrían existir una cantidad infinita de posibles hiperplanos!
- ¿Cuál es el mejor?



Clasificador lineal con el margen más amplio

- La función discriminante lineal con el máximo **margen** es la mejor
- El margen se define como la ancho que limita los datos (podría no existir)
- ¿Por qué es la mejor?
 - Generalización robusta y resistente a los valores atípicos

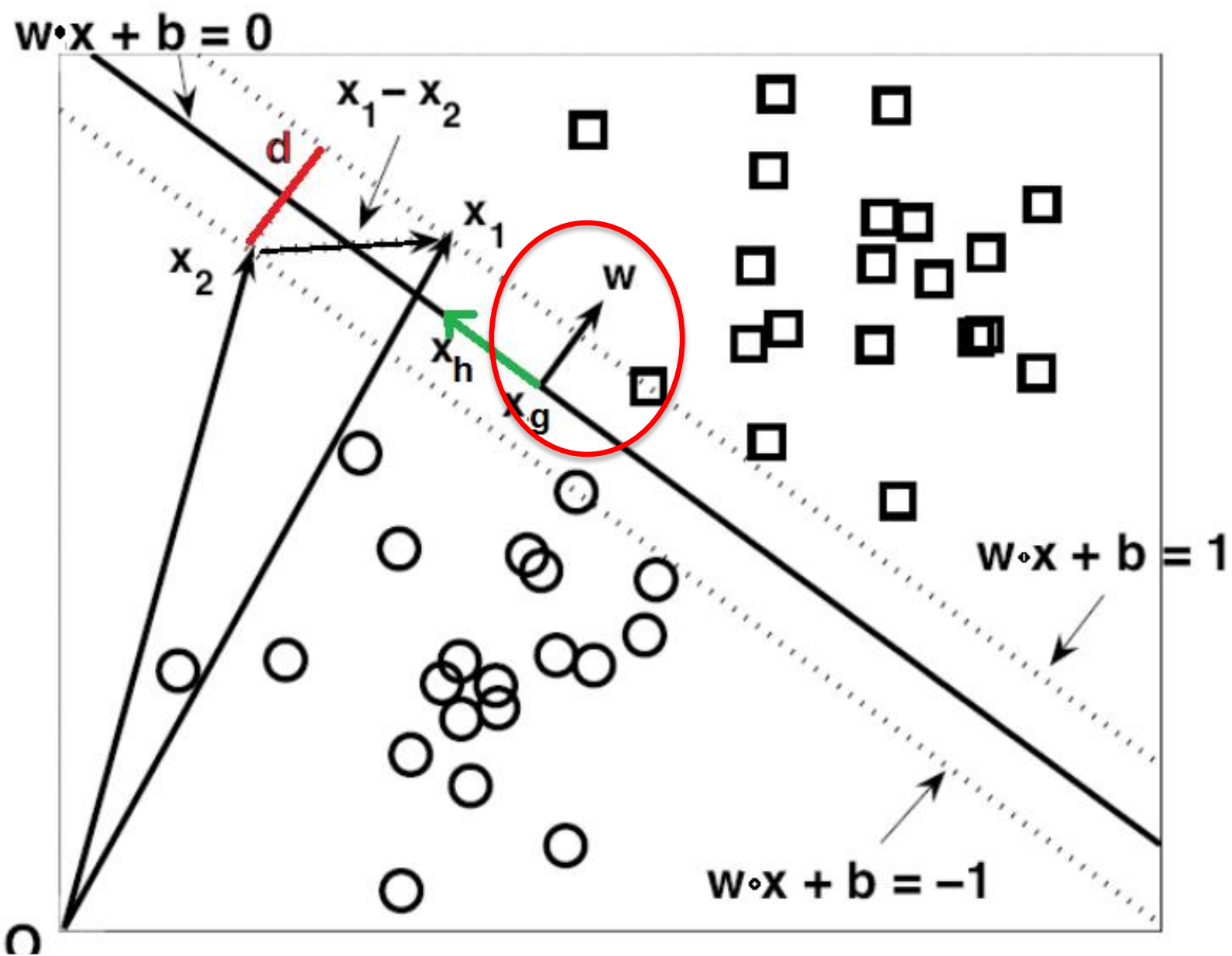


Formulación del Problema: Máquinas de Soporte Vectorial

- Supongamos que tenemos un problema de clasificación donde la variable a predecir es binaria y que tenemos n casos de entrenamiento (x_i, y_i) para $i = 1, 2, \dots, n$ donde $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$, es decir, los x_i son los predictores y y_i es la variable a predecir.
- Asumimos que $y_i \in \{-1, 1\}$ denota la etiqueta de clase.
- La frontera de decisión se puede escribir como:

$$w \cdot x + b = 0$$

- Donde w y b son los parámetros del modelo.



Formulación del Problema: Máquinas de Soporte Vectorial

Teorema: w es perpendicular a la frontera de decisión.

Teorema: El margen d se puede calcular como:

$$d = \frac{2}{\| w \|}$$

Formulación del Problema: Máquinas de Soporte Vectorial

Además maximizar:

$$d = \frac{2}{\|w\|}$$

Es equivalente a minimizar:

$$f(w) = \frac{\|w\|^2}{2}$$

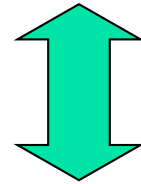
Resolver un Problema Optimización

Un problema de
programación
cuadrática con
restricciones
lineales

$$\min_w \frac{\|w\|^2}{2}$$

Sujeto a: $y_i(w \cdot x_i + b) \geq 1$ para $i = 1, 2, \dots, n$

Minimización de
Lagrange



$$L_P(w, b, \lambda_i) = \frac{\|w\|^2}{2} - \sum_{i=1}^n \lambda_i (y_i(w \cdot x_i + b) - 1)$$

con $\lambda_i \geq 0$ (los λ se llaman multiplicadores de Lagrange)

Ejemplo

Dada la siguiente tabla de datos calcule los parámetros del modelo w y b .

x_1	x_2	y	Lagrange Multiplier
0.3858	0.4687	1	65.5261
0.4871	0.611	-1	65.5261
0.9218	0.4103	-1	0
0.7382	0.8936	-1	0
0.1763	0.0579	1	0
0.4057	0.3529	1	0
0.9355	0.8132	-1	0
0.2146	0.0099	1	0

$$w_1 = \sum_i \lambda_i y_i x_{i1} = 65.5621 \times 1 \times 0.3858 + 65.5621 \times -1 \times 0.4871 = -6.64.$$

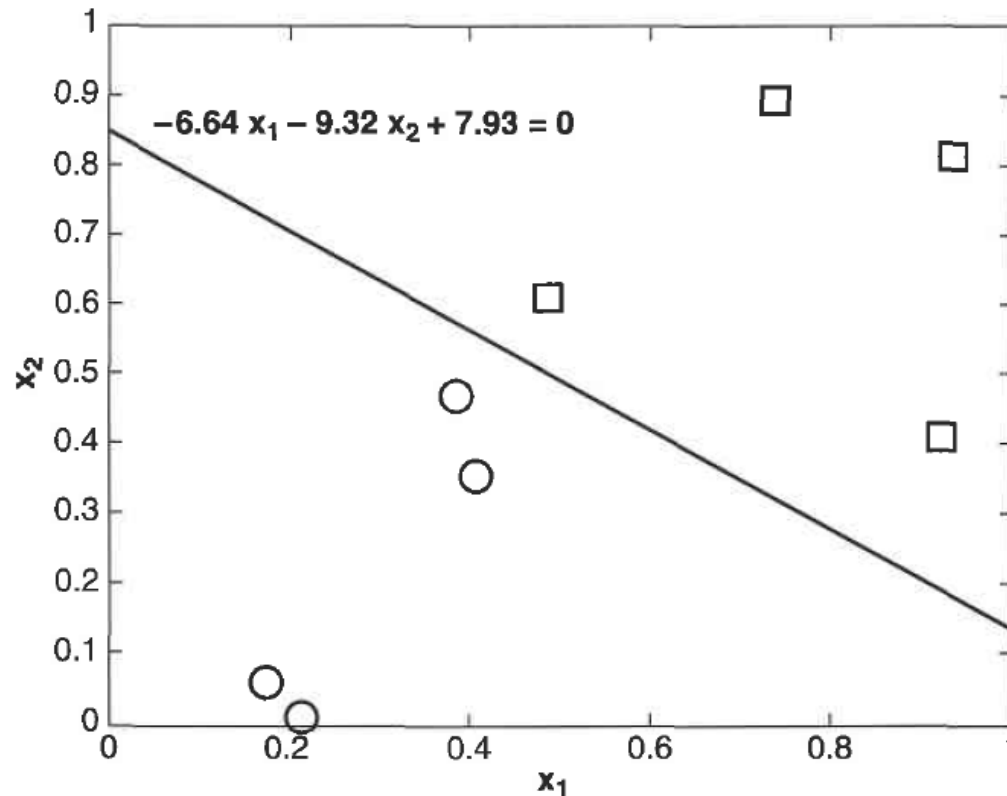
$$w_2 = \sum_i \lambda_i y_i x_{i2} = 65.5621 \times 1 \times 0.4687 + 65.5621 \times -1 \times 0.611 = -9.32.$$

Ejemplo

$$b^{(1)} = 1 - \mathbf{w} \cdot \mathbf{x}_1 = 1 - (-6.64)(0.3858) - (-9.32)(0.4687) = 7.9300.$$

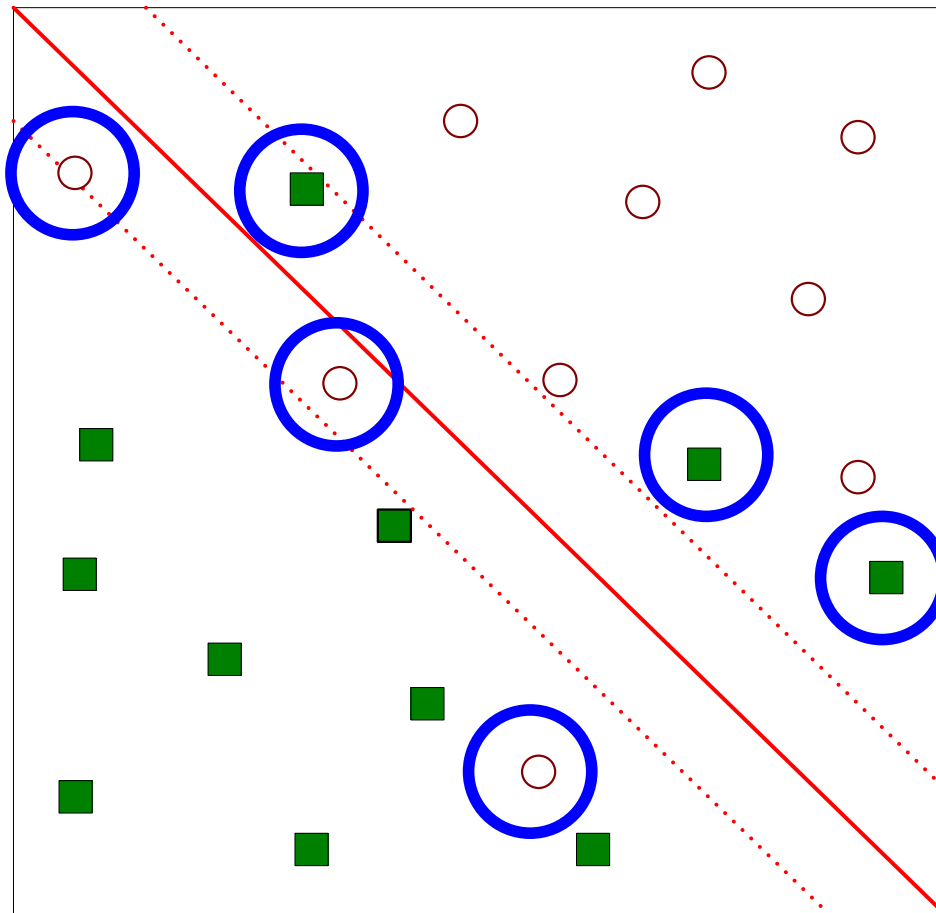
$$b^{(2)} = -1 - \mathbf{w} \cdot \mathbf{x}_2 = -1 - (-6.64)(0.4871) - (-9.32)(0.611) = 7.9289$$

Promediando los b 's se tiene que $b = 7.92945$ luego redondeando $b = 7.93$.

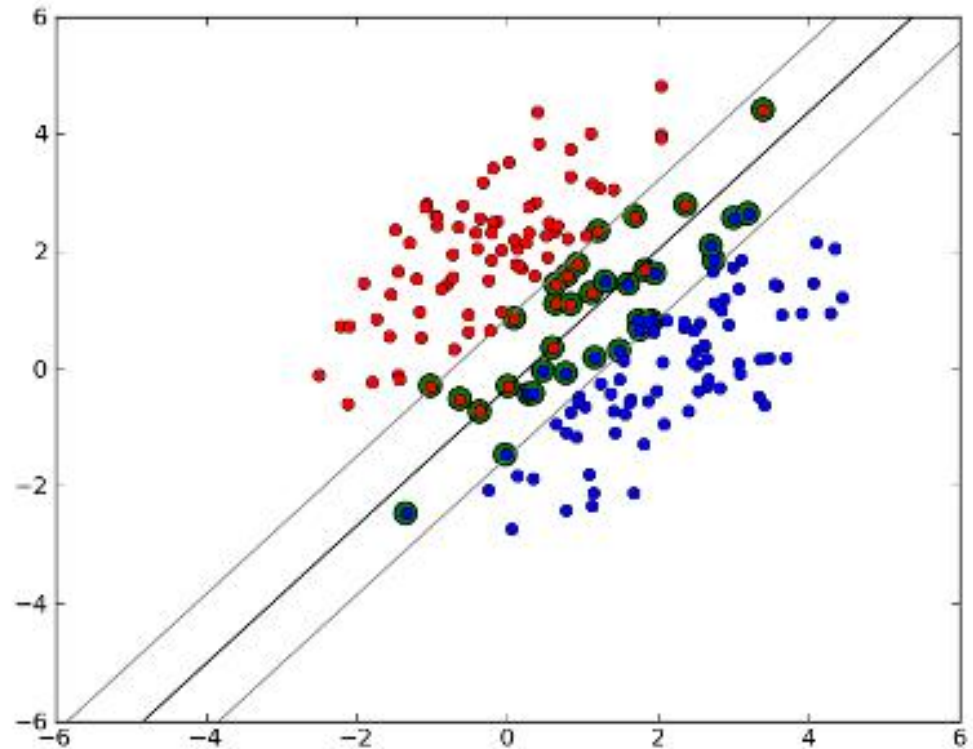


Máquinas de Soporte Vectorial

- ¿Qué pasa si el problema no es linealmente separable?



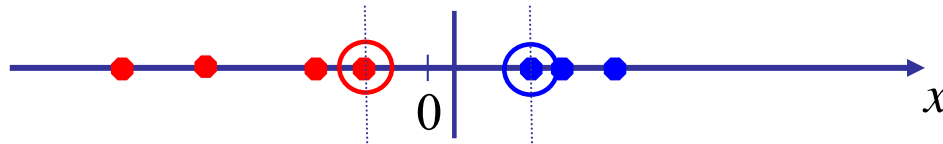
Margen de soporte débil (más vectores de soporte)



- El número de vectores de soporte entra a jugar cuando los datos no son linealmente separable, o sea, no existe un hiperplano que separe las 2 clases en los datos.
- En este caso el método trata entonces de encontrar un ***margen débil***, lo cual quiere decir que permite que algunos puntos de ambas clases queden dentro del margen, esto se logra permitiendo más vectores de soporte lo cual hace que el error aumente.
- Esto se hace porque de contrario entonces no existiría el plano de separación, claro el precio es un mayor error.

MVS no linealmente separables

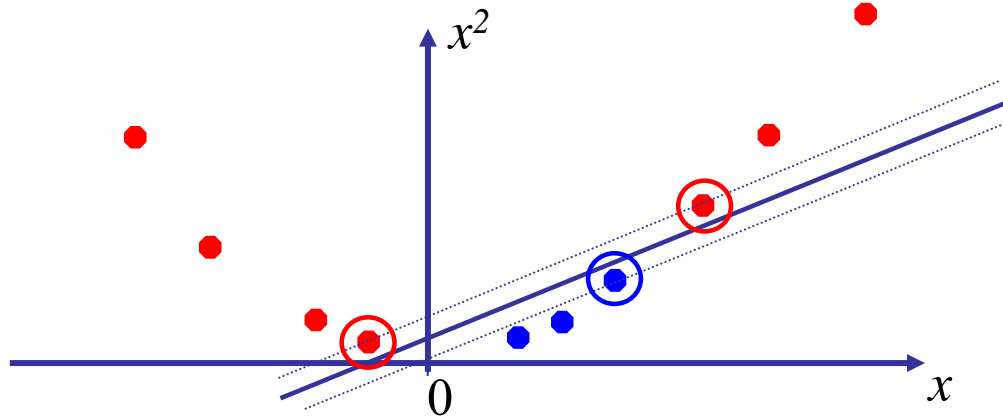
- Datos linealmente separables:



- Datos no linealmente separables:

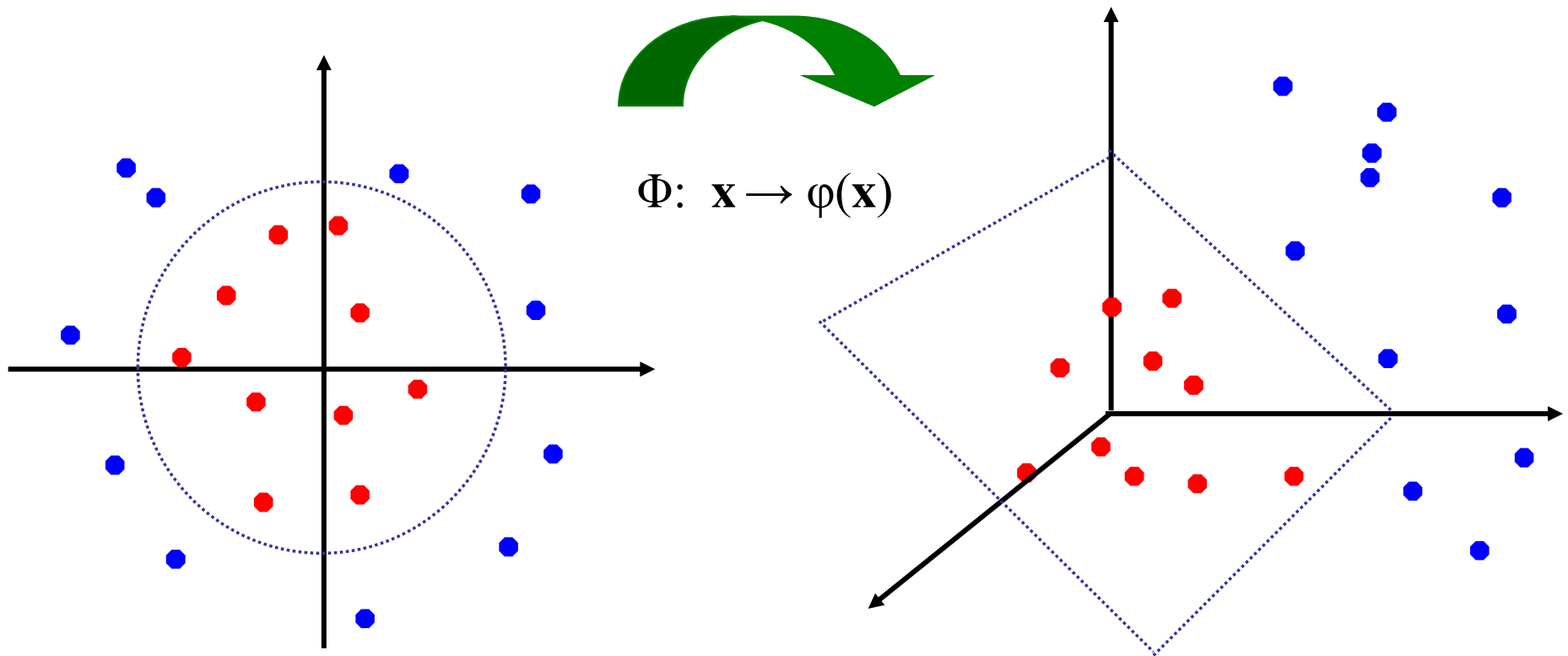


- La idea es... Encontrar una función para trasladar los datos a un espacio de mayor dimensión:

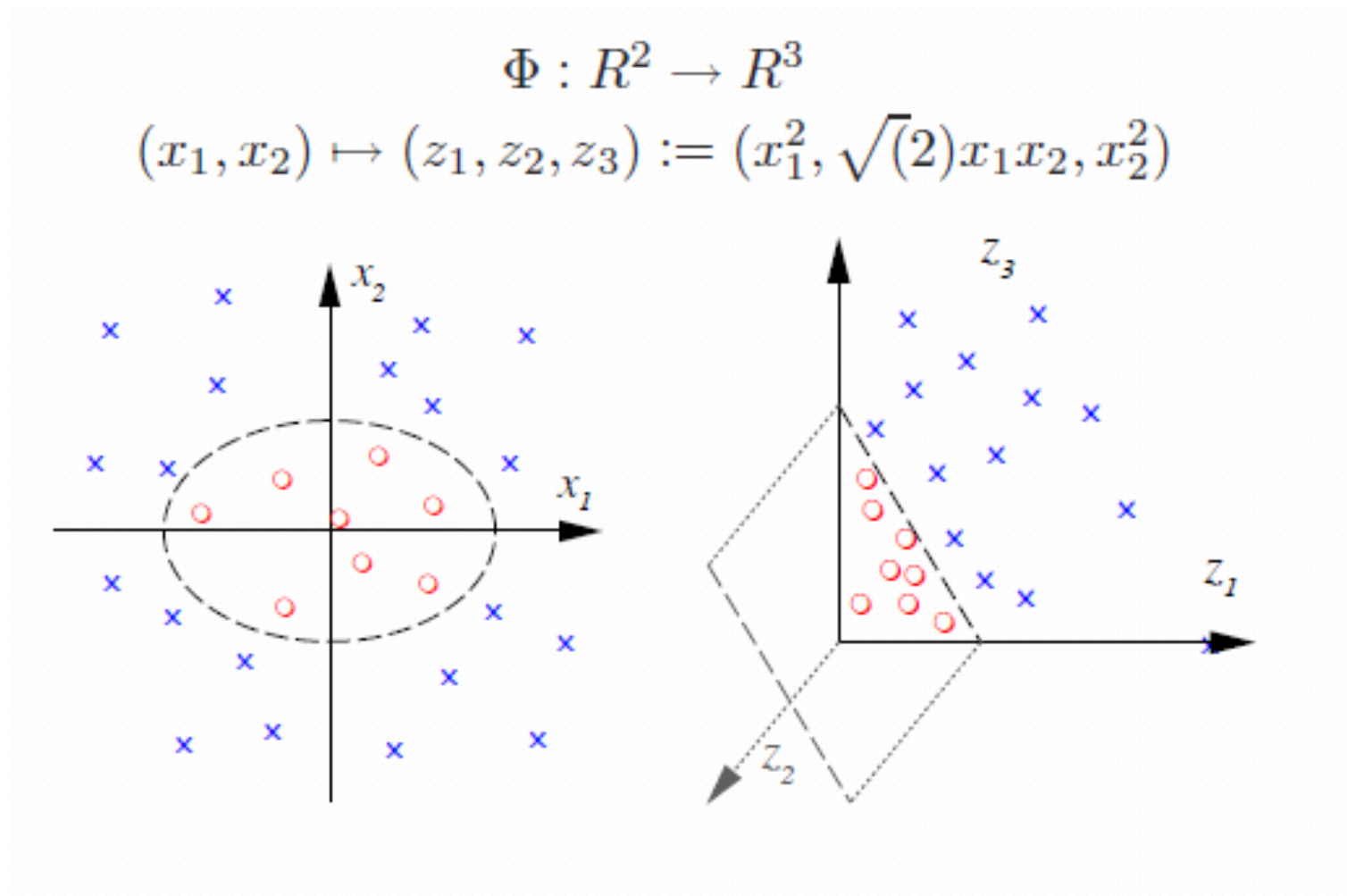


MVS no linealmente separables

- Idea general: Los datos de entrada se puede trasladar a algún espacio de mayor dimensión en el que la Tabla de Entrenamiento sí sea separable:

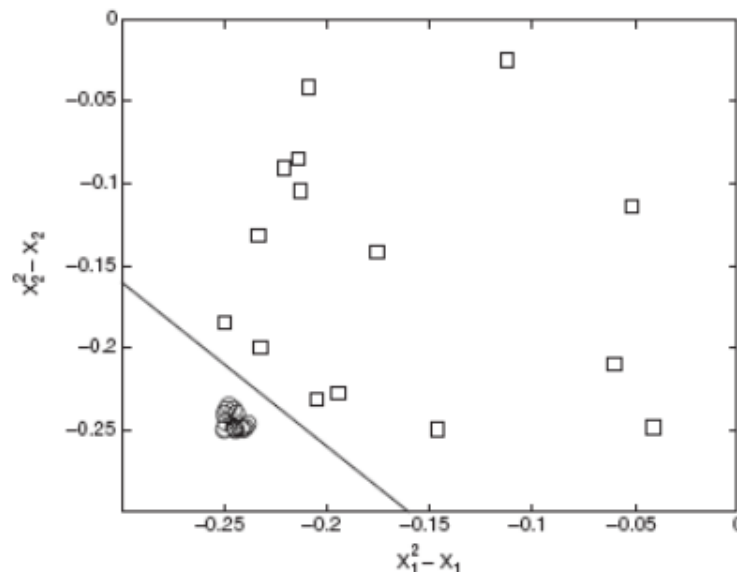
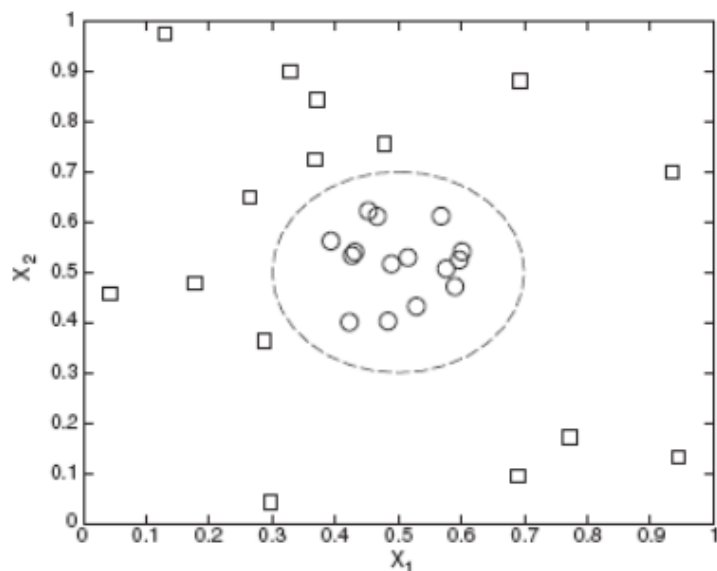


MVS no linealmente separables



El Truco del Núcleo (Kernel Trick)

- Las funciones de transformación del espacio vectorial pueden ser vistas como un producto punto.
- Ejemplo:



$$\Phi : (x_1, x_2) \longrightarrow (x_1^2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2, 1).$$

El Truco del Núcleo (Kernel Trick)

Ejemplos: Algunas funciones núcleo K usadas son:

$$K(x, y) = (x \cdot y + 1)^p$$

$$K(x, y) = e^{-\|x-y\|^2/(2\sigma^2)}$$

$$K(x, y) = \tanh(kx \cdot y - \delta)$$

SVM en Rattle

```
> library(rattle)
```

Rattle: A free graphical interface for data mining with R.
Versión 2.6.21 Copyright (c) 2006-2012 Togaware Pty Ltd.
Escriba 'rattle()' para agitar, sacudir y rotar sus datos.

```
> rattle()
```


Ejemplo 1: IRIS.CSV

Ejemplo con la tabla de datos IRIS

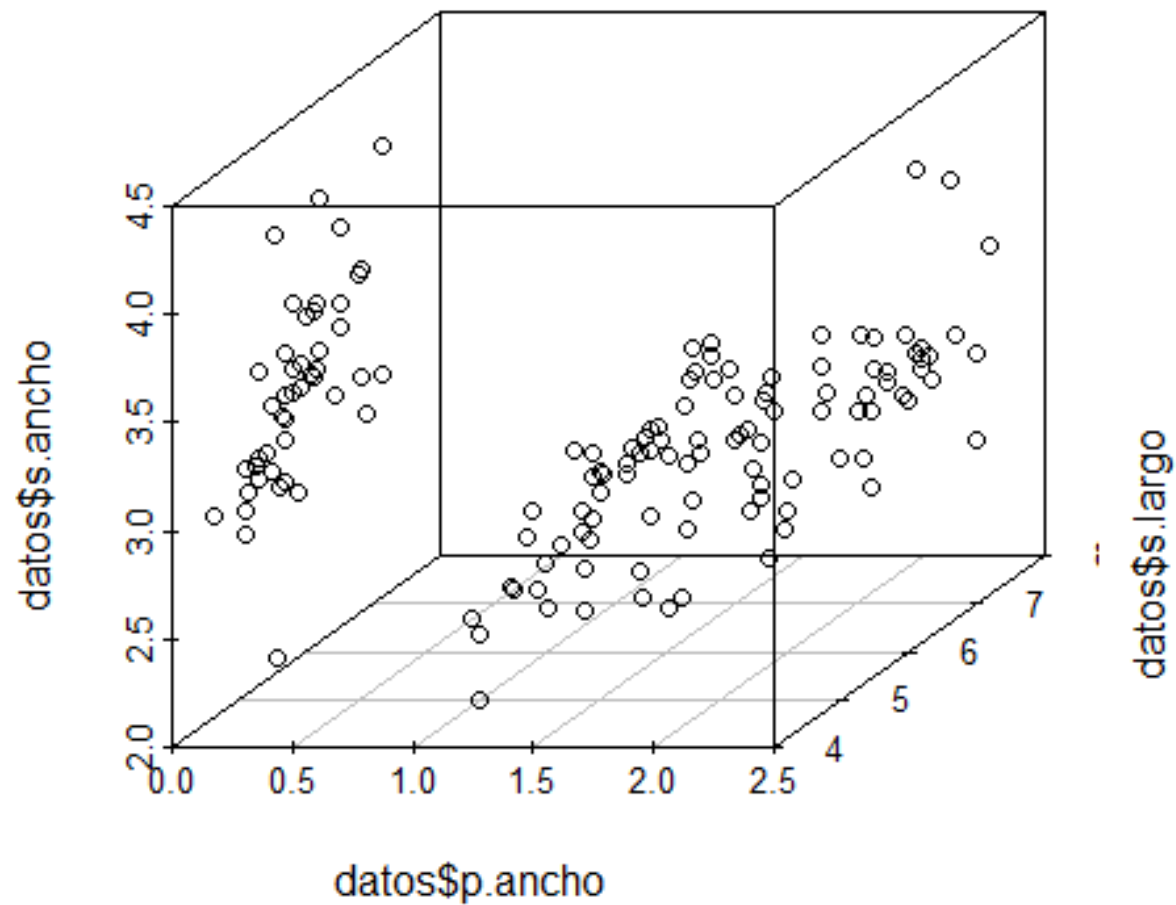
IRIS Información de variables:

- 1.sepal largo en cm
- 2.sepal ancho en cm
- 3.petal largo en cm
- 4.petal ancho en cm
- 5.clase:

- Iris Setosa
- Iris Versicolor
- Iris Virginica

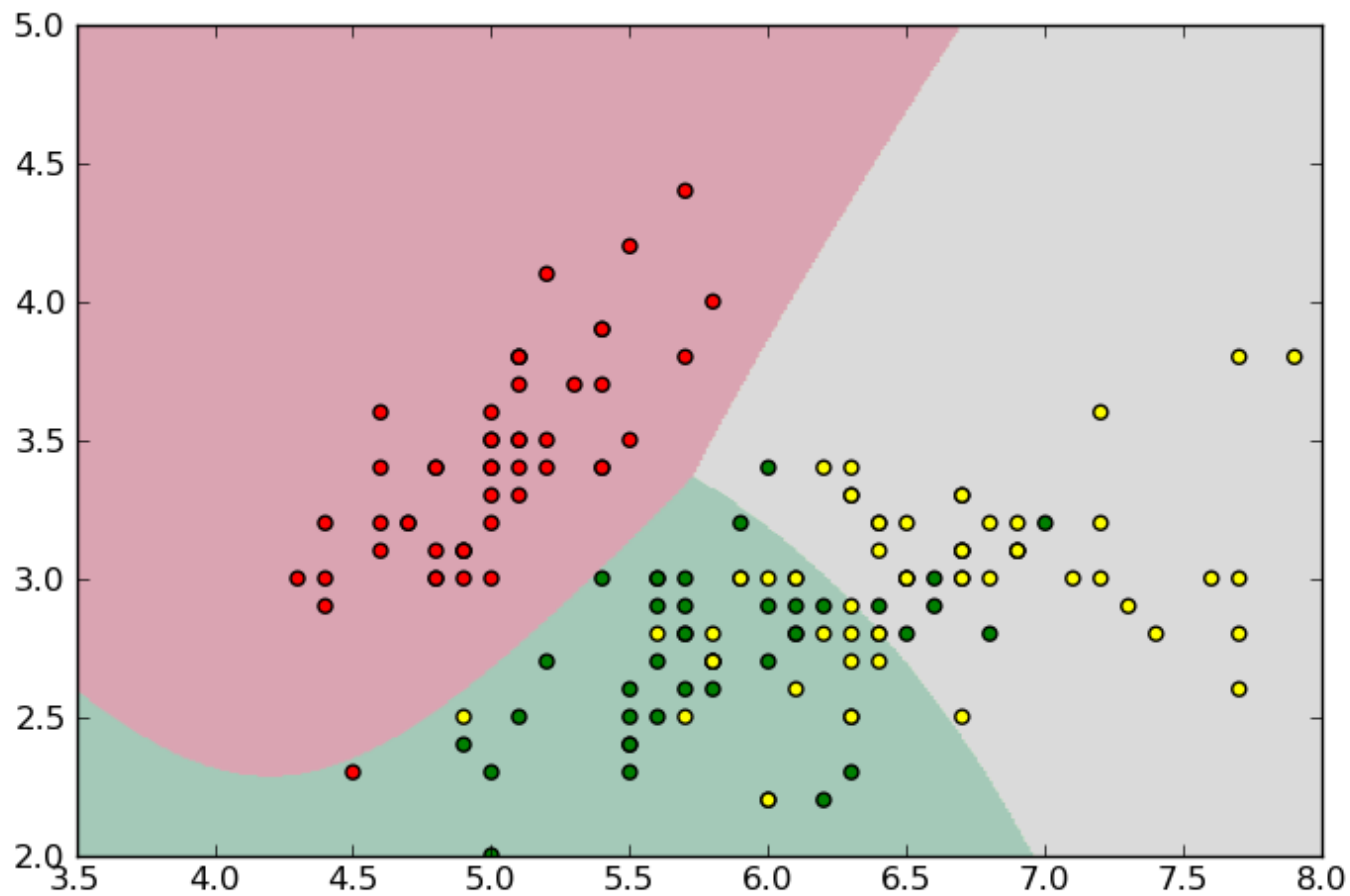


	A	B	C	D	E
1	s.largo	s.ancho	p.largo	p.ancho	tipo
2	5.1	3.5	1.4	0.2	setosa
3	4.9	3.0	1.4	0.2	setosa
4	4.7	3.2	1.3	0.2	setosa
5	4.6	3.1	1.5	0.2	setosa
6	5.0	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	setosa
8	4.6	3.4	1.4	0.3	setosa
9	5.0	3.4	1.5	0.2	setosa
10	4.4	2.9	1.4	0.2	setosa
11	4.9	3.1	1.5	0.1	setosa
12	5.4	3.7	1.5	0.2	setosa
13	4.8	3.4	1.6	0.2	setosa
14	4.8	3.0	1.4	0.1	setosa
15	4.3	3.0	1.1	0.1	setosa
16	5.8	4.0	1.2	0.2	setosa
17	5.7	4.4	1.5	0.4	setosa
18	5.4	3.9	1.3	0.4	setosa
19	5.1	3.5	1.4	0.3	setosa
20	5.7	3.8	1.7	0.3	setosa
21	5.1	3.8	1.5	0.3	setosa
22	5.4	3.4	1.7	0.2	setosa
23	5.1	3.7	1.5	0.4	setosa
24	4.6	3.6	1.0	0.2	setosa
25



```
> library(scatterplot3d)
> scatterplot3d(datos$p.ancho,datos$s.largo,datos$s.ancho)
```

Ejemplo 1: iris.csv



SVM en Rattle

Minero de datos R - [Rattle (iris.csv)]

Proyecto Herramientas Configuración Ayuda

Ejecutar Nuevo Abrir Guardar Informe Exportar Detener Salir

Datos Explorar Prueba Transformar Clúster Asociada Modelo Evaluar Registro

Origen: ☒ Hoja de cálculo ☐ ARFF ☐ ODBC ☐ Conjunto de datos R ☐ Archivo de datos R ☐ Librería ☐ Corpus ☐ Rutina

Archivo: Separador: Decimal: ☒ Encabezado

☒ Partición Semilla: Ver Editar

☒ Entrada ☐ Ignorar Calculadora de peso:

Tipo de datos de destino: ☒ Automática ☐ Categórica ☐ Numérica ☐ Supervivencia

No.	Variable	Tipo de datos	Entrada	Destino	Riesgo	Ident	Ignorar	Weight	Comentario
1	s.largo	Numérica	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 35
2	s.ancho	Numérica	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 23
3	p.largo	Numérica	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 43
4	p.ancho	Numérica	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 22
5	tipo	Categórica	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 3

Minero de datos R - [Rattle (iris.csv)]

Proyecto Herramientas Configuración Ayuda

Ejecutar Nuevo Abrir Guardar Informe Exportar Detener Salir

Datos Explorar Prueba Transformar Clúster Asociada Modelo Evaluar Registro

Tipo: ☐ Árbol ☐ Bosque ☐ Potencia ☒ SVM ☐ Lineal ☐ Red neural ☐ Supervivencia ☐ Todos

Destino: tipo

Núcleo: Radial Basis (rbfdot) Options:

Constructor de modelos: ksvm

Resumen del modelo SVM (construido con ksvm):

Support Vector Machine object of class "ksvm"

SV type: C-svc (classification)
parameter : cost C = 1

Gaussian Radial Basis kernel function.
Hyperparameter : sigma = 0.595104071822472

Number of Support Vectors : 48

Objective Function Value : -3.7399 -4.0892 -17.2768
Training error : 0.009524
Probability model included.

Tiempo transcurrido: 0.36 segs

¿Cómo evaluar la calidad del Modelo Predictivo?

Matriz de confusión (Matriz de Error)

- La **Matriz de Confusión** contiene información acerca de las predicciones realizadas por un **Método o Sistema de Clasificación**, comparando para el conjunto de individuos en de la tabla de aprendizaje o de testing, la predicción dada versus la clase a la que estos realmente pertenecen.
- La siguiente tabla muestra la matriz de confusión para un clasificador de dos clases:

		Predicción	
		Negativo	Positivo
Valor Real	Negativo	a	b
	Positivo	c	d

Ejemplo: Matriz de confusión

		Predicción	
		Mal Pagador	Buen Pagador
Valor Real	Mal Pagador	800	200
	Buen Pagador	500	1500

- 800 predicciones de Mal Pagador fueron realizadas correctamente, para un 80%, mientras que 200 no, para un 20%.
- 1500 predicciones de Buen Pagador fueron realizadas correctamente, para un 75%, mientras que 500 no (para un 25%).
- En general 2300 de 3000 predicciones fueron correctas para un 76,6% de efectividad en las predicciones. **Cuidado**, este dato es a veces engañoso y debe ser siempre analizado en la relación a la dimensión de las clases.

Matriz de confusión

		Predicción	
		Negativo	Positivo
Valor Real	Negativo	a	b
	Positivo	c	d

- La Precisión ***P*** de un modelo de predicción es la proporción del número total de predicciones que son correctas respecto al total. Se determina utilizando la ecuación: **$P = (a+d)/(a+b+c+d)$**
- ***Cuidado***, este índice es a veces engañoso y debe ser siempre analizado en la relación a la dimensión de las clases.

Ejemplo: Matriz de confusión

		Predicción	
		Fraude	No Fraude
Valor Real	Fraude	0	8
	No Fraude	3	989

- **Cuidado**, este índice es a veces engañoso y debe ser siempre analizado en la relación a la dimensión de las clases.
- En la Matriz de Confusión anterior la Precisión ***P*** es del 98,9%, sin embargo, el modelo no detectó ningún fraude.

Matriz de confusión

		Predicción	
		Negativo	Positivo
Valor Real	Negativo	a	b
	Positivo	c	d

- La Precisión Positiva (**PP**) es la proporción de casos positivos que fueron identificados correctamente, tal como se calcula usando la ecuación: **$PP = d/(c+d)$**
- En el ejemplo anterior Precisión Positiva **PP** es del 99,6% .

Matriz de confusión

		Predicción	
		Negativo	Positivo
Valor Real	Negativo	a	b
	Positivo	c	d

- La Precisión Negativa (***PN***) es la proporción de casos negativos que fueron identificados correctamente, tal como se calcula usando la ecuación: ***PN = a/(a+b)***
- En el ejemplo anterior Precisión Negativa ***PN*** es del 0% .

Matriz de confusión

		Predicción	
		Negativo	Positivo
Valor Real	Negativo	a	b
	Positivo	c	d

- Falsos Positivos (**FP**) es la proporción de casos negativos que fueron clasificados incorrectamente como positivos, tal como se calcula utilizando la ecuación: **$FP = b/(a+b)$**
- Falsos Negativos (**FN**) es la proporción de casos positivos que fueron clasificados incorrectamente como negativos, tal como se calcula utilizando la ecuación: **$FN = c/(c+d)$**

Matriz de confusión

		Predicción	
		Negativo	Positivo
Valor Real	Negativo	a	b
	Positivo	c	d

- Asertividad Positiva (**AP**) indica la proporción de buena predicción para los positivos, tal como se calcula utilizando la ecuación: **$AP = d/(b+d)$**
- Asertividad Negativa (**AN**) indica la proporción de buena predicción para los negativos, tal como se calcula utilizando la ecuación: **$AN = a/(a+c)$**

Curva ROC

- Especificidad (SPC) Representa la proporción de Verdaderos negativos.

$$SPC = a / (a + d).$$

- Curva ROC (**Receiver Operating Characteristic**) es una representación gráfica de la sensibilidad frente a la especificidad para un sistema clasificador binario según se varía el umbral de discriminación.

Matriz de confusión para más de 2 clases

- La Matriz de Confusión puede calcularse en general para un problema con p clases.
- En la matriz ejemplo que aparece a continuación, de 8 alajuelenses reales, el sistema predijo que 3 eran heredianos y de 6 heredianos predijo que 1 era un limonense y 2 eran alajuelenses. A partir de la matriz se puede ver que el sistema tiene problemas distinguiendo entre alajuelenses y heredianos, pero que puede distinguir razonablemente bien entre limonenses y las otras provincias.

		Predicción		
		alajuelense	herediano	limonense
Valor Real	alajuelense	5	3	0
	herediano	2	3	1
	limonense	0	2	11

Índices de Calidad

- Precisión Global
 - **$P=0,7$ (o sea 70%)**
- Precisión en cada variable:
 - **$P(\text{Alajuelense})=0.625$ (o sea 63%)**
 - **$P(\text{Heradiano})=0.5$ (o sea 50%)**
 - **$P(\text{Limonense})=0.846$ (o sea 85%)**

Matriz de confusión en Rattle (Matriz de Error)

Minero de datos R - [Rattle (iris.csv)]

Proyecto Herramientas Configuración Ayuda

Ejecutar Nuevo Abrir Guardar Informe Exportar Detener Salir

Datos Explorar Prueba Transformar Clúster Asociada Modelo Evaluar Registro

Tipo: ☒ Matriz de error ☐ Riesgo ☐ Curva de costo ☐ Hand ☐ Elevación ☐ ROC ☐ Precisión ☐ Sensibilidad

Modelo: ☐ Arbol ☐ Potenciar ☐ Bosque ☒ SVM ☐ Lineal ☐ Red neural ☐ Supervivencia ☐ KMeans ☐ HClust

Datos: ☐ Entrenamiento ☐ Convalidación ☒ Prueba ☐ Completo ☐ Ingresar ☐ Archivo CSV

Variable de riesgo: Informe: ☒ Clase ☐ Probabilidad Incluir:

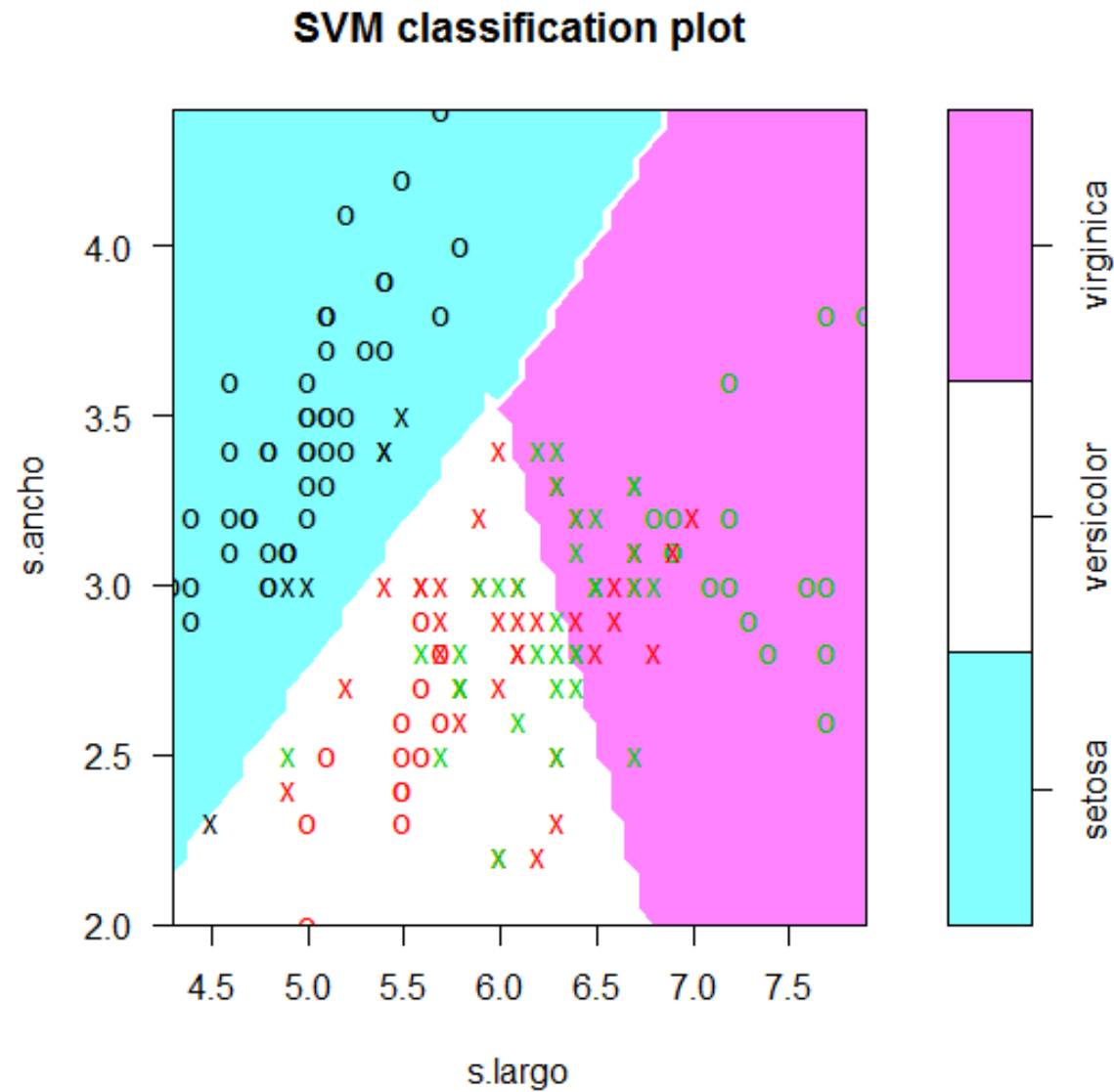
Matriz de error para el modelo SVM en iris.csv [prueba] (cuentas):

	Predicho		
Real	setosa	versicolor	virginica
setosa	11	0	0
versicolor	0	7	1
virginica	0	0	4

Matriz de error para el modelo SVM en iris.csv [prueba] (%):

	Predicho		
Real	setosa	versicolor	virginica
setosa	48	0	0
versicolor	0	30	4
virginica	0	0	17

Usando solo 2 variables para poder graficar



Ejemplo 2:

Credit-Scoring

MuestraAprendizajeCredito2500.csv
MuestraTestCredito2500.csv

```
> setwd("C:/Users/Oldemar/Google Drive/Curso Minería Datos II - Optativo/Datos")
> taprendizaje<-read.csv("MuestraAprendizajeCredito2500.csv",sep = ";",header=T)
> taprendizaje
```

	MontoCredito	IngresoNeto	CoefCreditoAvaluo	MontoCuota	GradoAcademico	BuenPagador
1	1	1	1	1	1	Si
2	3	1	1	1	1	Si
3	2	1	1	1	1	Si
4	1	2	1	1	1	Si
5	1	1	1	1	1	Si
6	2	1	1	1	1	Si
7	4	1	1	1	1	Si
8	1	2	1	1	1	Si
9	1	2	1	1	1	Si
10	3	2	1	1	1	Si
11	1	1	1	1	1	Si
12	1	2	1	1	1	Si
13	3	1	1	1	1	Si
14	3	1	1	1	1	Si
15	2	1	1	1	1	Si
16	3	1	1	1	1	Si
17	3	1	1	1	1	Si

Descripción de Variables

MontoCredito

1=Muy Bajo
2=Bajo
3=Medio
4=Alto

MontoCuota

1=Muy Bajo
2=Bajo
3=Medio
4=Alto

IngresoNeto

1=Muy Bajo
2=Bajo
3=Medio
4=Alto

GradoAcademico

1=Bachiller
2=Licenciatura
3=Maestría
4=Doctorado

CoeficienteCreditoAvaluo

1=Muy Bajo
2=Bajo
3=Medio
4=Alto

BuenPagador

1=NO
2=Si



Minero de datos R - [Rattle (MuestraAprendizaje)]

Proyecto Herramientas Configuración Ayuda

Ejecutar Nuevo Abrir Guardar Informe Exportar Detener Salir

Datos Explorar Prueba Transformar Clúster Asociada Modelo Evaluar Registro

Origen: ☒ Hoja de cálculo ☐ ARFF ☐ ODBC ☐ Conjunto de datos R ☐ Archivo de datos R ☐ Librería ☐ Corpus ☐ Rutina

Archivo:  MuestraAprendiz...  Separador: ; Decimal: . ☒ Encabezado

☐ Partición 70/15/15 Semilla: 42 Ver Editar

☒ Entrada ☒ Ignorar Calculadora de peso:

Tipo de datos de destino: ☒ Automática ☐ Categórica ☐ Numérica ☐ Supervivencia

No.	Variable	Tipo de datos	Entrada	Destino	Riesgo	Ident	Ignorar	Weight	Comentario
1	MontoCredito	Numérica	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 4
2	IngresoNeto	Numérica	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 2
3	CoefCreditoAvaluo	Numérica	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 12
4	MontoCuota	Numérica	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 4
5	GradoAcademico	Numérica	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 2
6	BuenPagador	Categórica	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 2

Minero de datos R - [Rattle (MuestraAprendizajeCredito2500.csv)]

Proyecto Herramientas Configuración Ayuda

Ejecutar Nuevo Abrir Guardar Informe Exportar Detener Salir

Datos Explorar Prueba Transformar Clúster Asociada Modelo Evaluar Registro

Tipo: ☐ Árbol ☐ Bosque ☐ Potenciar ☒ SVM ☐ Lineal ☐ Red neural ☐ Supervivencia ☐ Todos

Destino: BuenPagador

Núcleo: Radial Basis (rbfdot) Options:

Resumen del modelo SVM (construido con ksvm):

Support Vector Machine object of class "ksvm"

SV type: C-svc (classification)
parameter : cost C = 1

Gaussian Radial Basis kernel function.
Hyperparameter : sigma = 0.195027075304086

Number of Support Vectors : 466

Objective Function Value : -352.8905
Training error : 0.073143
Probability model included.

Tiempo transcurrido: 1.60 segs

Rattle marca de tiempo: 2012-11-30 16:31:15 Oldemar

=====

Matriz de confusión en Rattle (Matriz de Error)

Minero de datos R - [Rattle (MuestraAprendizajeCredito2500.csv)]

Proyecto Herramientas Configuración Ayuda

Ejecutar Nuevo Abrir Guardar Informe Exportar Detener Salir

Datos Explorar Prueba Transformar Clúster Asociada Modelo Evaluar Registro

Tipo: ☒ Matriz de error ☐ Riesgo ☐ Curva de costo ☐ Hand ☐ Elevación ☐ ROC ☐ Precisión ☐ Sensibilidad ☐ O

Modelo: ☐ Árbol ☐ Potenciar ☐ Bosque ☒ SVM ☐ Lineal ☐ Red neural ☐ Supervivencia ☐ KMeans ☐ HClust

Datos: ☐ Entrenamiento ☐ Convalidación ☐ Prueba ☐ Completo ☐ Ingresar ☒ Archivo CSV ☐ C

Variable de riesgo: Informe: ☒ Clase ☐ Probabilidad ☐ Incluir ☐ Ide

Matriz de error para el modelo SVM en MuestraTestCredito2500.csv (cuentas):

	Predicho	
Real	No	Si
No	189	156
Si	21	2134

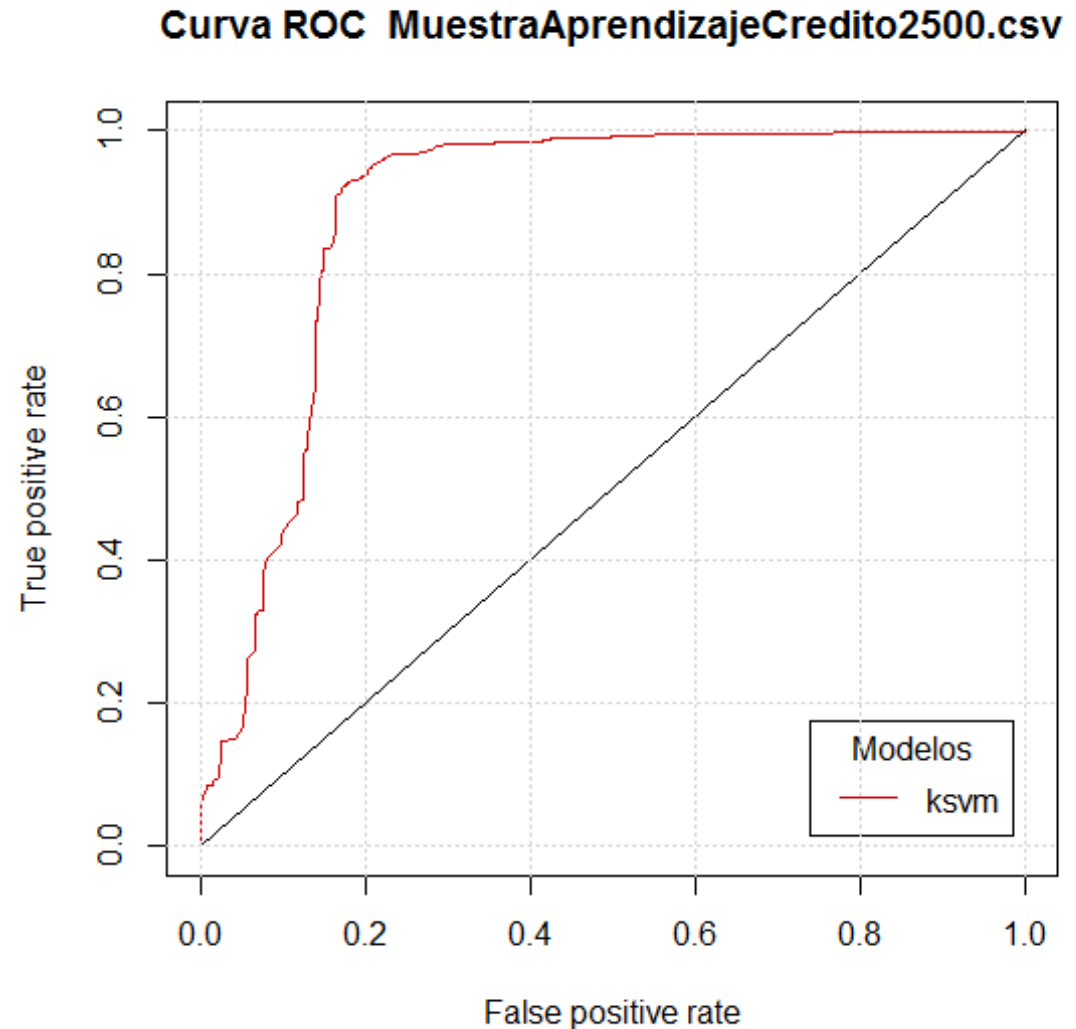
Matriz de error para el modelo SVM en MuestraTestCredito2500.csv (%):

	Predicho	
Real	No	Si
No	8	6
Si	1	85

Error general: 0.0708

Curva ROC

- Una curva ROC compara la tasa de falsos positivos con la de verdaderos positivos.
- El área bajo la curva ROC = 0.8880



Curvas ROC - Árboles y SVM

Curva ROC MuestraAprendizajeCredito2500.csv [validar]

