


Técnicas de Aprendizaje Supervisado



Temas del día

1. Métodos paramétricos (1)
 - a. Naive Bayes (3)
 - b. Support Vector Machines (4)
2. Métodos no paramétricos (2)
 - a. **Árboles de decisiones**
 - b. **Random Forests**
3. Métodos de ensamblaje

Métodos Paramétricos

- Regresión Lineal
- Regresión Logística
- Perceptron
- Redes Neuronales Simples
- SVMs lineares y polinomiales
- Naive Bayes

Beneficios

- Simples
- Rápidos
- No necesita tanta data

Limitaciones

- Limitados a una forma
- Complejidad limitada

Árboles de decisiones

Árboles de Decisiones

Gender	Age	App
F	15	
F	25	
M	32	
F	40	
M	12	
M	14	

¿Qué influye más para ver qué aplicación descargan?

¿Edad o género?

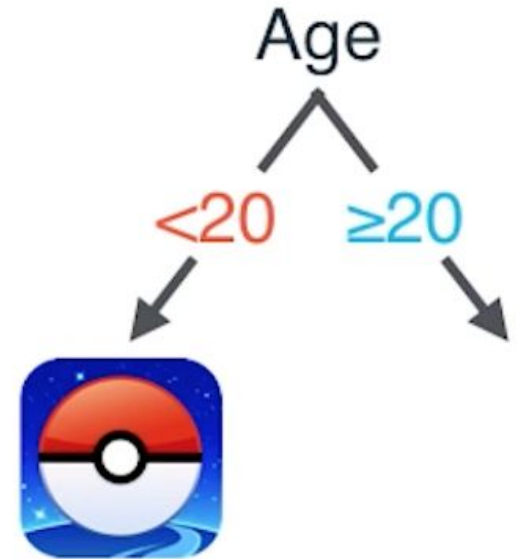
Árboles de Decisiones

Gender	Age	App
F	15	
F	25	
M	32	
F	40	
M	12	
M	14	

¿A partir de qué edad ya no descargan Pokemon Go?

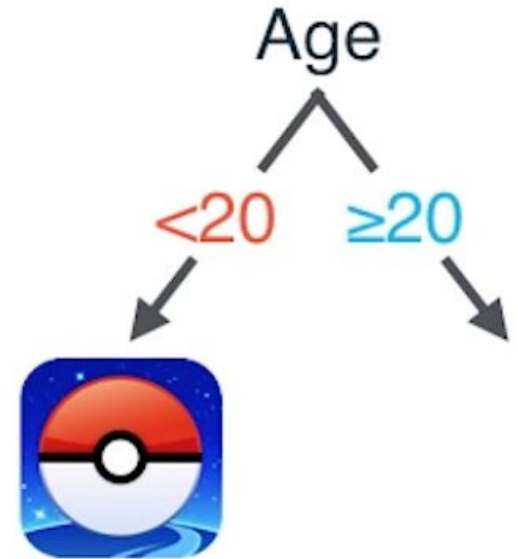
Árboles de Decisiones

Gender	Age	App
F	15	
F	25	
M	32	
F	40	
M	12	
M	14	



Árboles de Decisiones

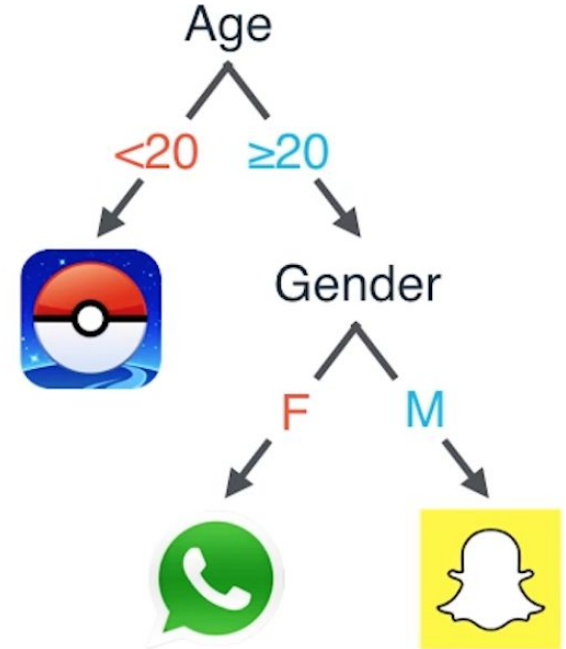
Gender	Age	App
F	15	
F	25	
M	32	
F	40	
M	12	
M	14	



De los restantes, ¿cómo influye el género?

Árboles de Decisiones







Gender	Age	App
F	15	
F	25	
M	32	
F	40	
M	12	
M	14	



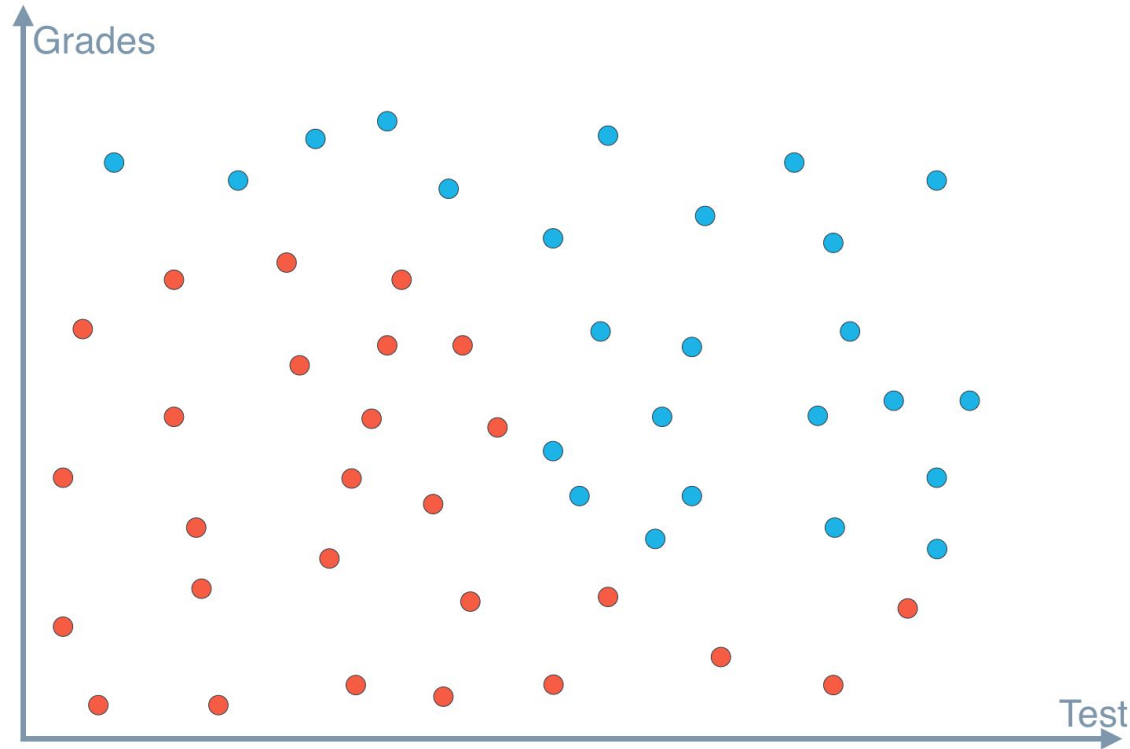
¿Qué recomiendas para una mujer en una fábrica?

¿Qué recomiendas para un hombre en una oficina?

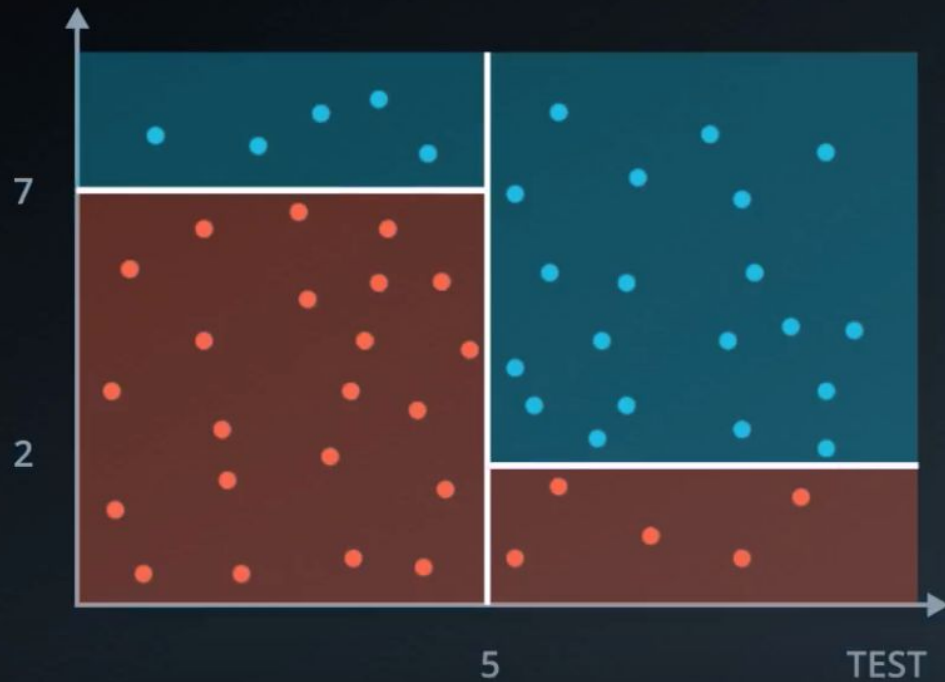
¿Qué recomiendas para una niña en una escuela?

Gender	Occupation	App
F	Study	Pokemon Go 
F	Work	WhatsApp 
M	Work	Snapchat 
F	Work	WhatsApp 
M	Study	Pokemon Go 
M	Study	Pokemon Go 

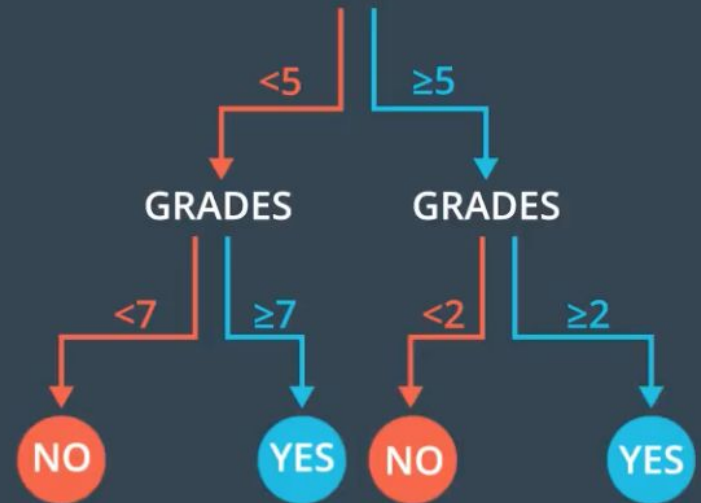
¿Nos conviene línea horizontal o vertical?



GRADES

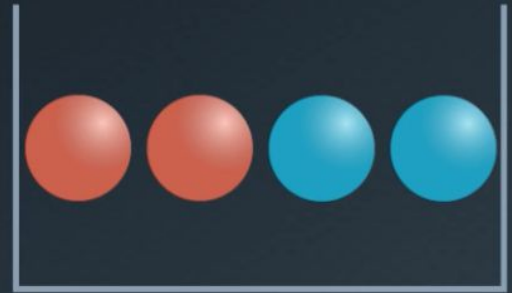
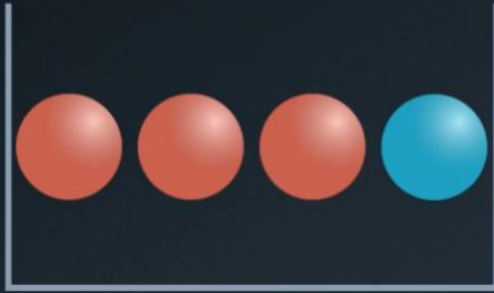
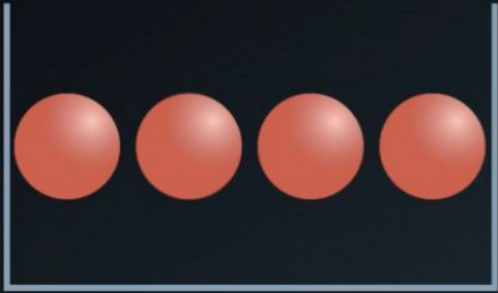


TEST



Entropía

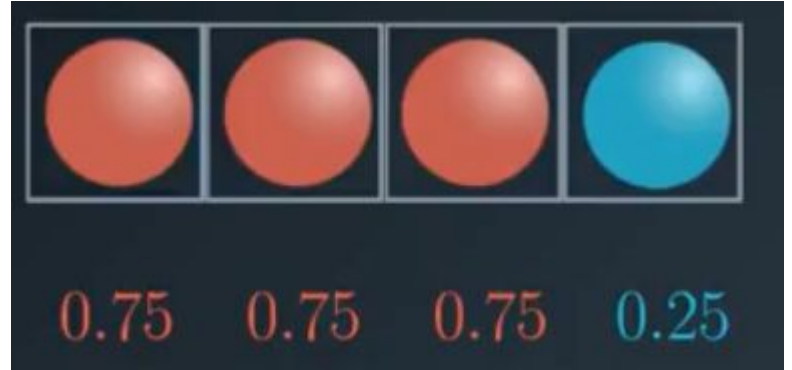
- Dado los estados, ¿qué tanto se puede mover?
- Entropía baja significa mucha ganancia de conocimiento



Baja entropía
Mucho conocimiento

Mucha entropía
Baja entropía

Entropía



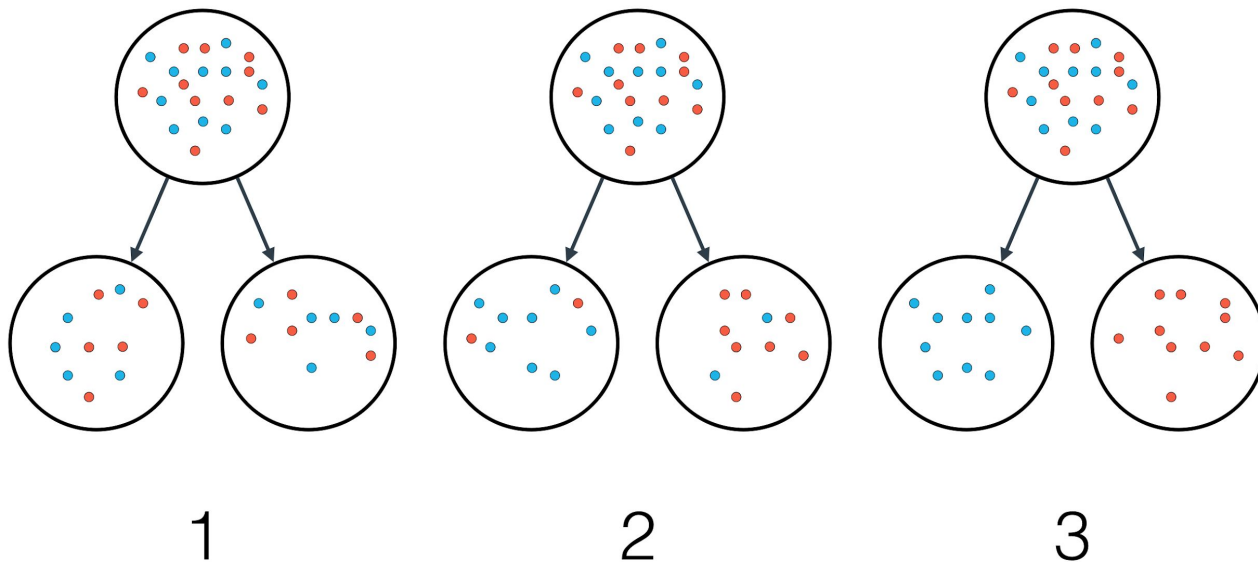
Entropía



$$Entropy = -\frac{m}{m+n} \log_2 \left(\frac{m}{m+n} \right) - \frac{n}{m+n} \log_2 \left(\frac{n}{m+n} \right)$$

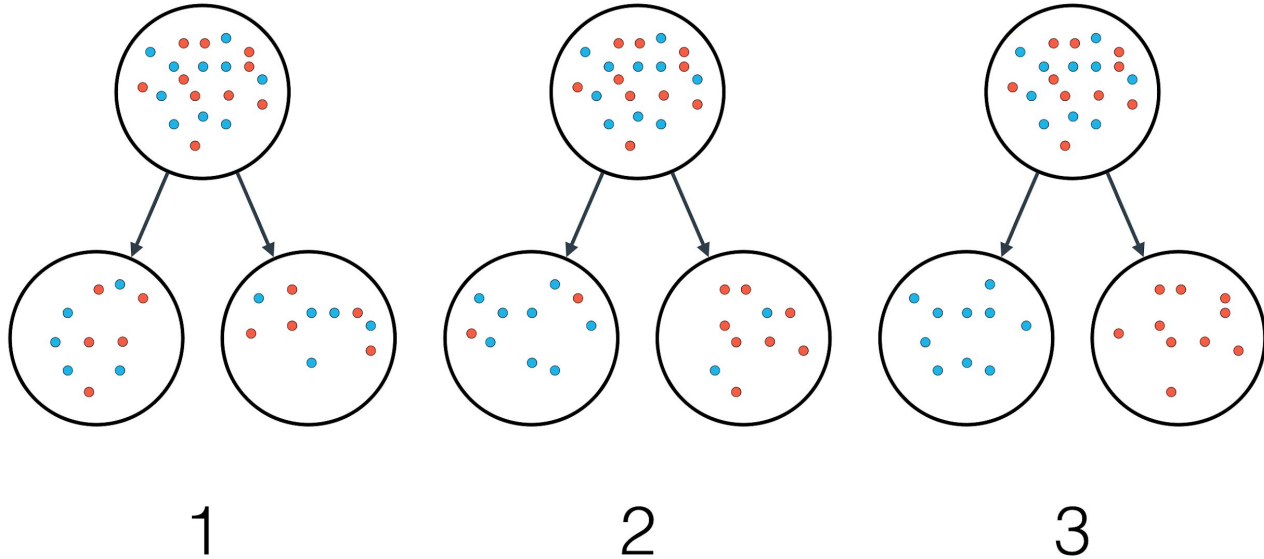
Ganancia de información

- ¿Dónde ganamos más información?



Ganancia de información



- Cambio de entropía del padre al promedio de la entropía de los hijos






Algoritmo

- Mirar cada split posible para cada columna
- Calcular la ganancia de información
- Maximizarla


¿Separamos por género u ocupación?




Gender	Occupation	App
F	Study	
F	Work	
M	Work	
F	Work	
M	Study	
M	Study	

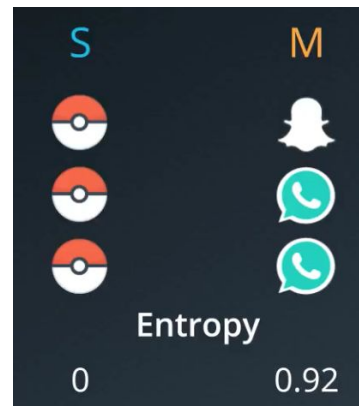
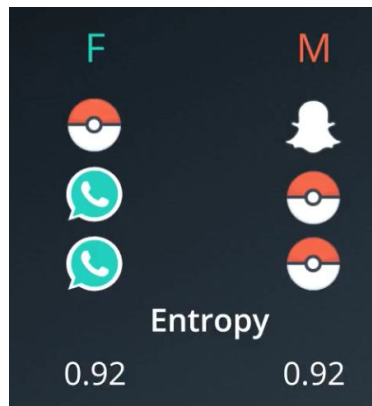
$$\text{Entropy} = -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right) - \frac{1}{6}\log_2\left(\frac{1}{6}\right)$$
$$= 1.46$$

¿Separamos por género u ocupación?




Gender	Occupation	App
F	Study	
F	Work	
M	Work	
F	Work	
M	Study	
M	Study	

$$\text{Entropy} = -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right) - \frac{1}{6}\log_2\left(\frac{1}{6}\right)$$
$$= 1.46$$



Gender	Occupation	App
F	Study	
F	Work	
M	Work	
F	Work	
M	Study	
M	Study	

Gender	Occupation	App
F	Study	
M	Study	
M	Study	

Gender	Occupation	App
F	Work	
M	Work	
F	Work	

Gender	Occupation	App
F	Work	
F	Work	

Gender	Occupation	App
M	Work	

Features continuos

- También se puede hacer
- Se prueban todas las líneas y se va iterando para construir árbol de decisiones

Pros

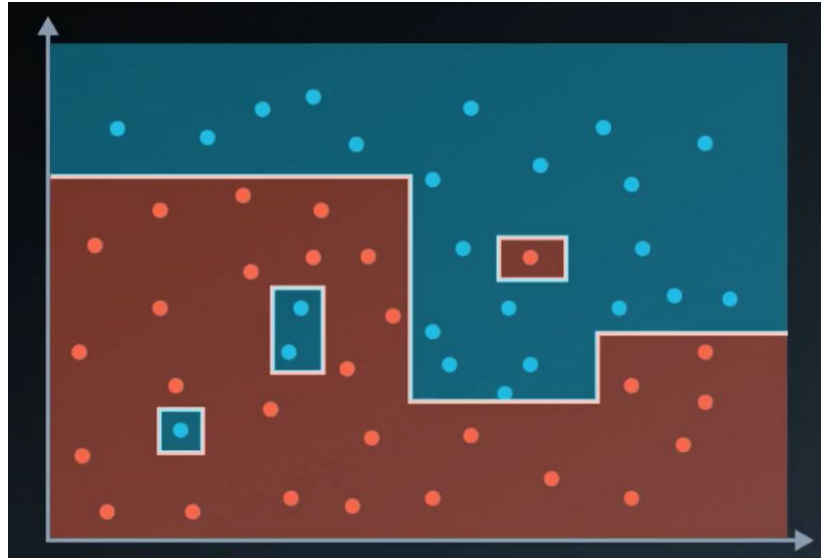
- Fácil de leer
- Poderoso con mala información
- Barato de lanzar
- Puede manejar data numérica y categórica
- Caja blanca

Cons

- Fácil de hacer overfitting
- Caro de entrenar
- Encuentra óptimos locales

Problema

- ¿Cómo prevenimos overfitting?
 - ¿Cómo prevenimos que memorice?



Random Forests

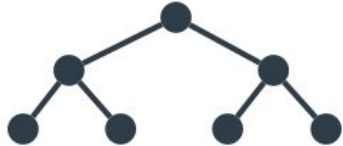
- Elige algunos features aleatoriamente y construye árboles de decisiones
- Ensamblarlos y revisa el promedio
 - Votación
- Puedes especificar el máximo número de features por árbol
- Útil para detectar qué features importan más.

Hiperparámetros

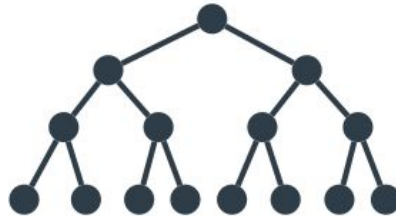
- Maximum Depth



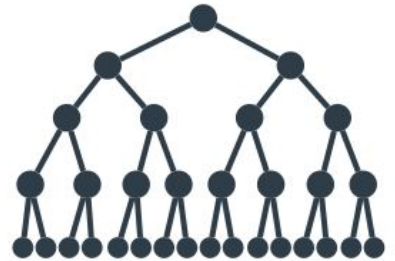
Depth = 1



Depth = 2



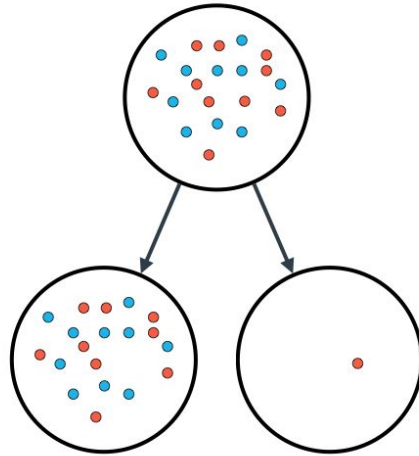
Depth = 3



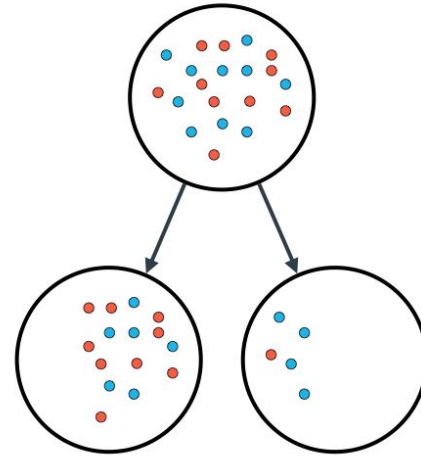
Depth = 4

Hiperparámetros

- Minimum samples per leaf



Minimum samples per leaf = 1



Minimum samples per leaf = 5

Naive Bayes

Naive Bayes

Spam



Non-spam



¿Qué palabra encontrarías en un correo de Spam?

Naive Bayes



“Cheap”

Spam



Non-spam



Si en el correo dice la palabra “cheap”, ¿cuál es la probabilidad de que sea spam?

Naive Bayes



“Cheap” → 80%



Spelling mistake → 70%



Missing title → 95%

Naive Bayes

- Se basa en probabilidades a priori
- Ejemplo:
 - Clasificador de spam

Ejemplo



Alex

Brenda

$$P(\text{Alex}) = 0.5 \quad P(\text{Brenda}) = 0.5$$

- $P(\text{rojo} | A)$ Probabilidad que Alex use suéter rojo
- $P(\text{rojo} | B)$ Probabilidad que Brenda use suéter rojo.

Probabilidad que la persona con suéter rojo es Alex.

Conocemos

Inferimos

Ejemplo



Alex

Brenda

$$P(\text{Alex}) = 0.5 \quad P(\text{Brenda}) = 0.5$$

- $P(\text{rojo} | A)$ Probabilidad que Alex use suéter rojo
- $P(\text{rojo} | B)$ Probabilidad que Brenda use suéter rojo.

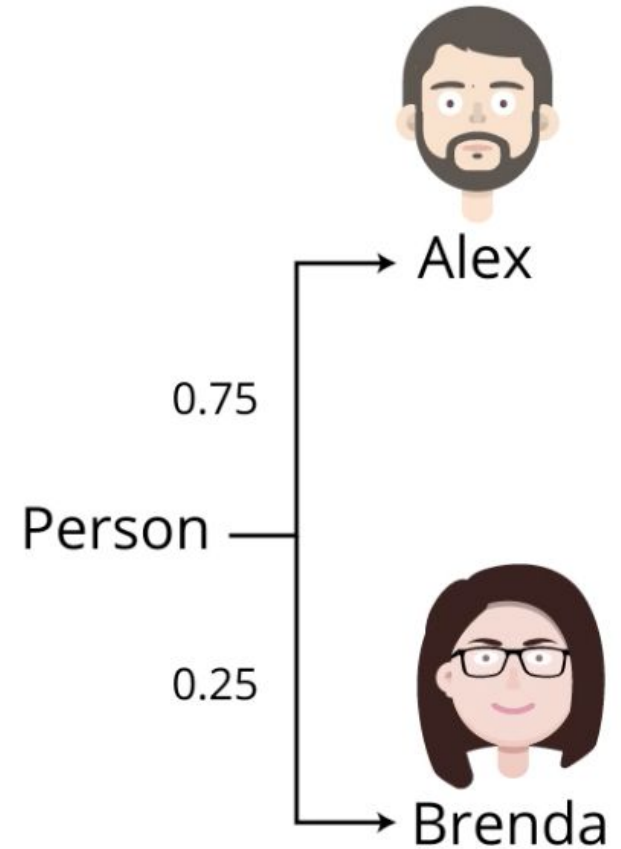
Probabilidad que la persona con suéter rojo es Alex.

Conocemos

Inferimos

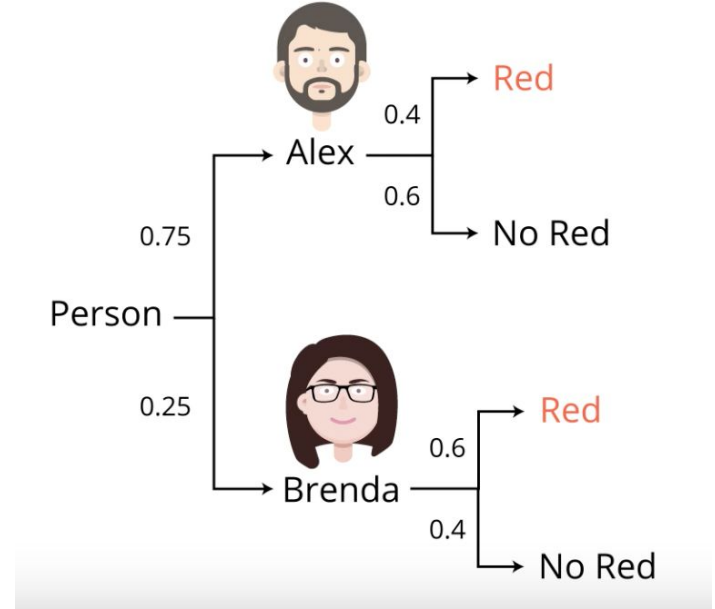
Ejemplo 2

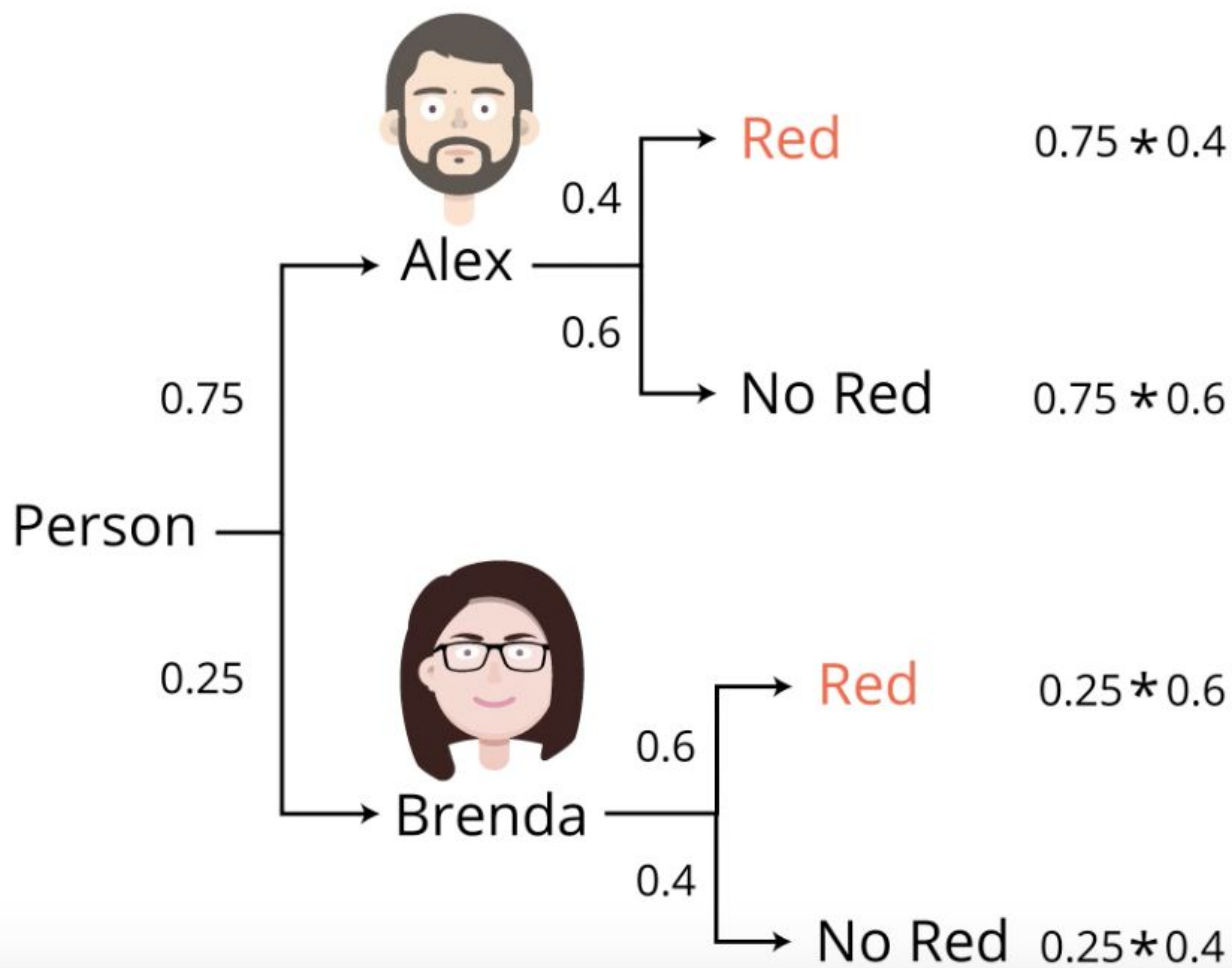
- Alex viene a la oficina 3 veces
- Brenda viene a la oficina 1 vez



Ejemplo 2

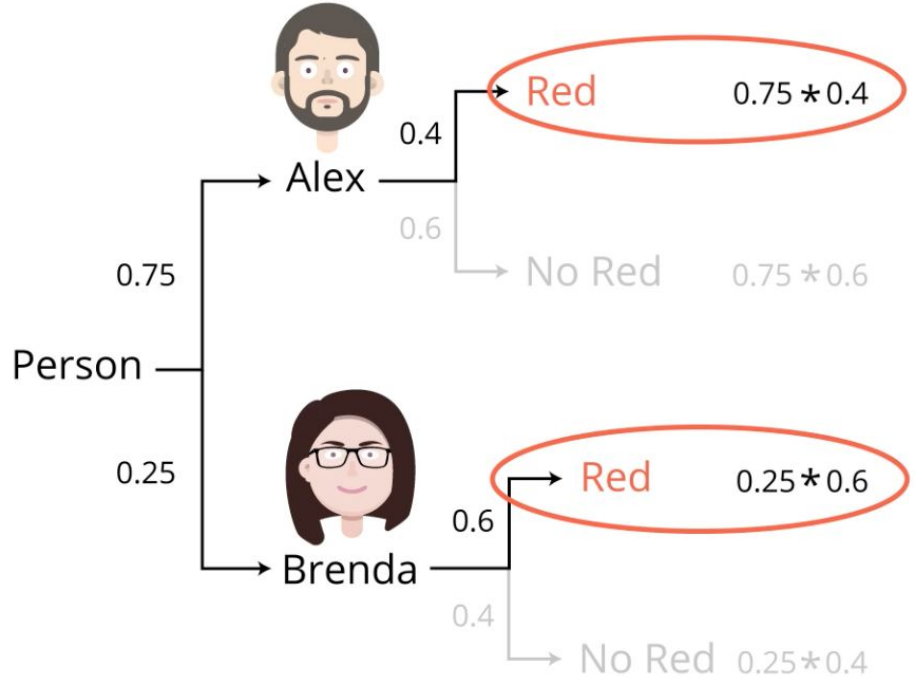
- Alex viene a la oficina 3 veces
- Brenda viene a la oficina 1 vez
- **La persona tiene suéter rojo**
- Alex usa rojo 2 veces a la semana
- Brenda usa rojo 3 veces a la semana





Ejemplo 2

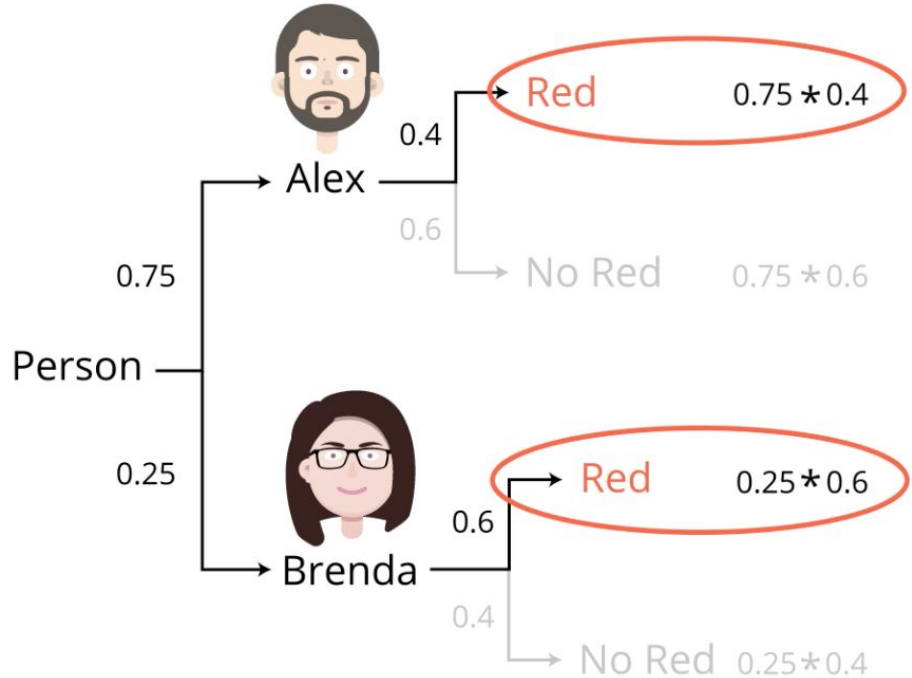
- **La persona tiene suéter rojo**
- Por lo tanto, debemos normalizar para mantener proporciones pero que sumen 1



Ejemplo 2

$$P(A|R) = \frac{0.75 * 0.4}{0.75 * 0.4 + 0.25 * 0.6} = 0.67$$

$$P(B|R) = \frac{0.25 * 0.6}{0.75 * 0.4 + 0.25 * 0.6} = 0.33$$



$$p(B \mid A) = \frac{p(A \mid B) p(B)}{p(A)}$$

Naive assumption

- Asume que todos los features son independientes

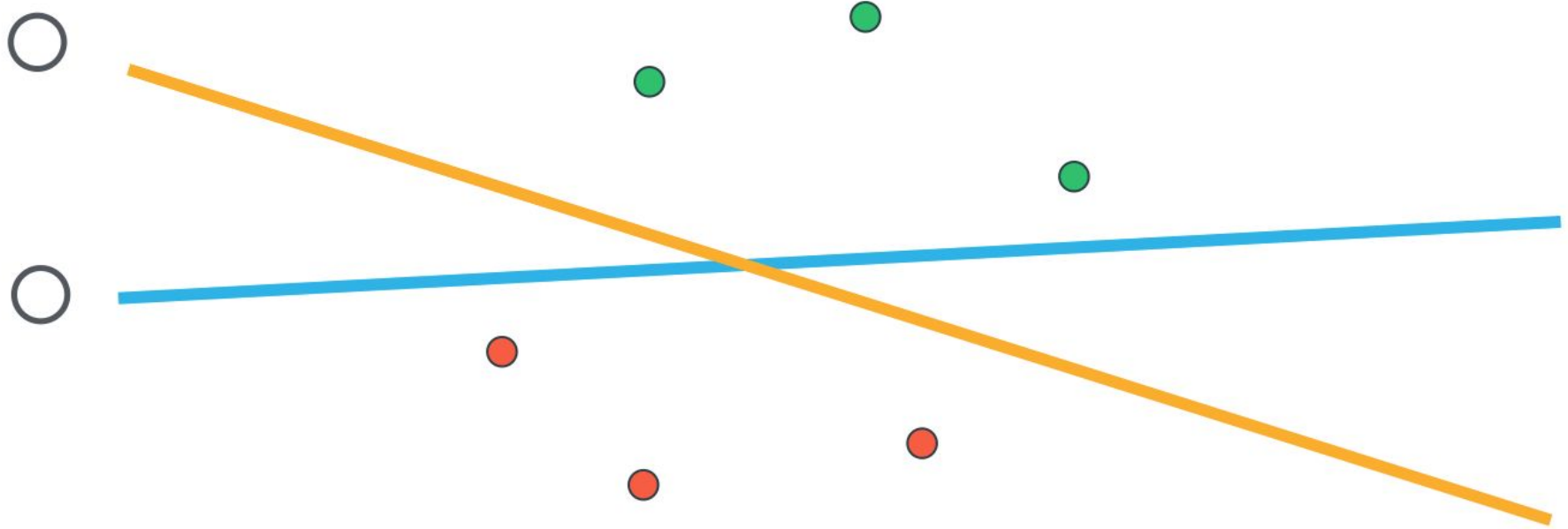
$$P(A \text{ inter } B) = P(A)P(B)$$

- Probabilidad condicional
 - $P(A|B)P(B) = P(B|A)P(A)$
 - $P(A|B)$ es proporcional a $P(B|A)P(A)$
- $P(\text{spam} | \text{easy, money})$ proporcional a
 - $P(\text{easy, money} | \text{spam})P(\text{spam})$ que es
 - $P(\text{easy} | \text{spam}) P(\text{money} | \text{spam}) P(\text{spam})$

SVM - Support Vector Machine

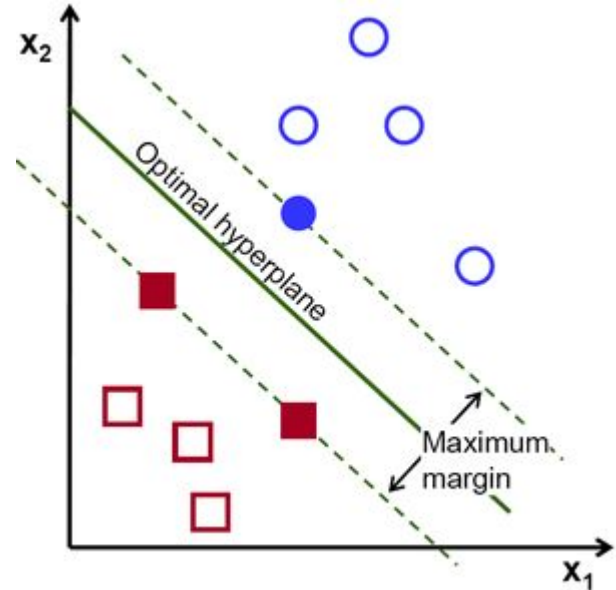
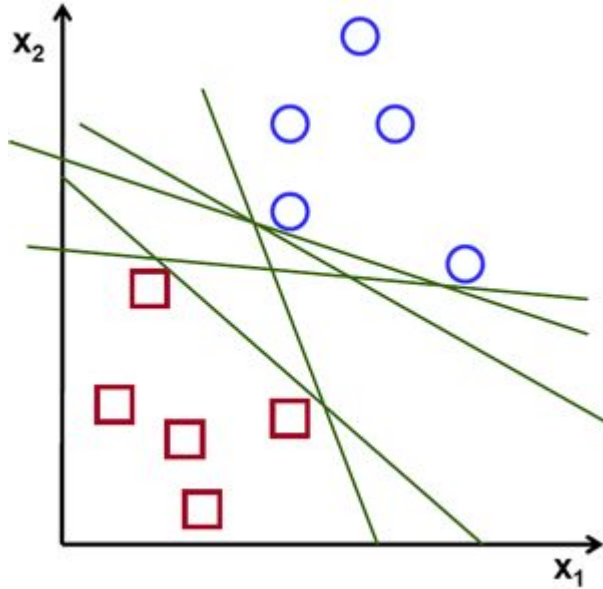
Support Vector Machines

- ¿Cómo podemos separar información?
- De las dos líneas propuestas, ¿cuál es mejor?



SVM

- Calculamos las distancias a los puntos más cercanos
- Maximizamos la distancia (margen)



Error

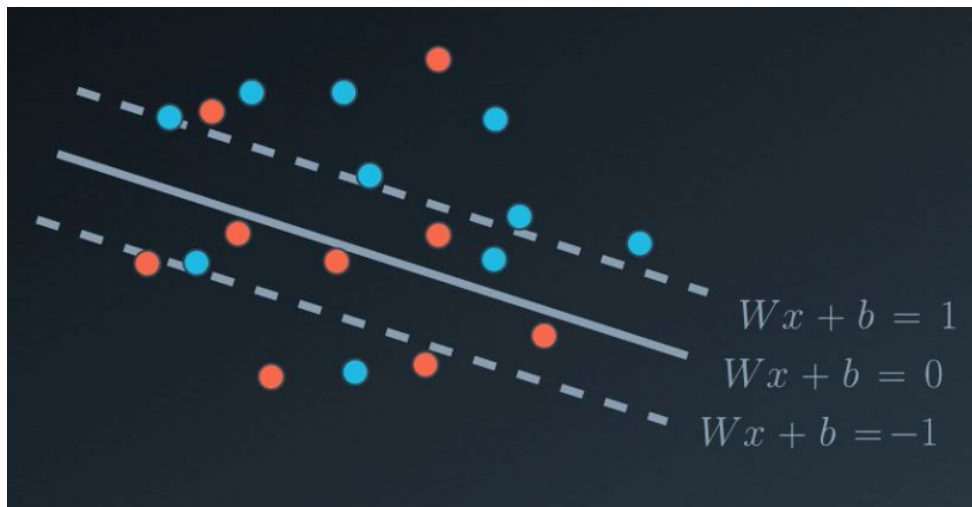
- Tenemos dos errores:
 - Error de clasificación
 - Error de margen

Error de clasificación

- Digamos que nuestra línea es $Wx + b = 0$
- El margen lo definimos con
 - $Wx + b = 1$
 - $Wx + b = -1$
- Calculamos el error a partir

de las líneas hacia arriba y

hacia abajo.

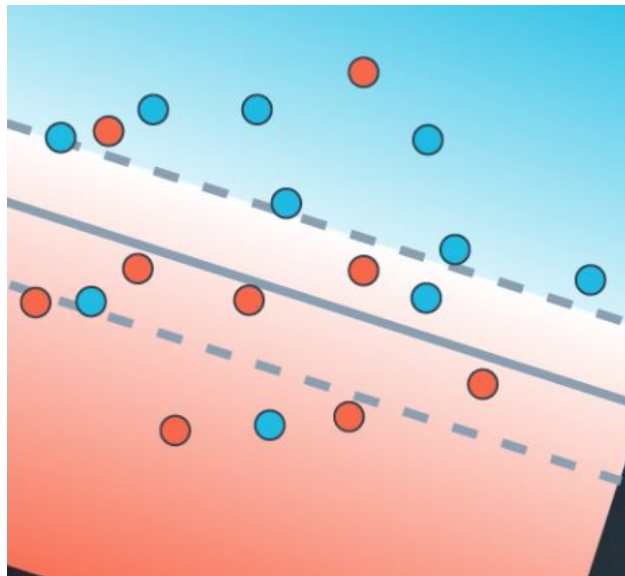


Error de clasificación

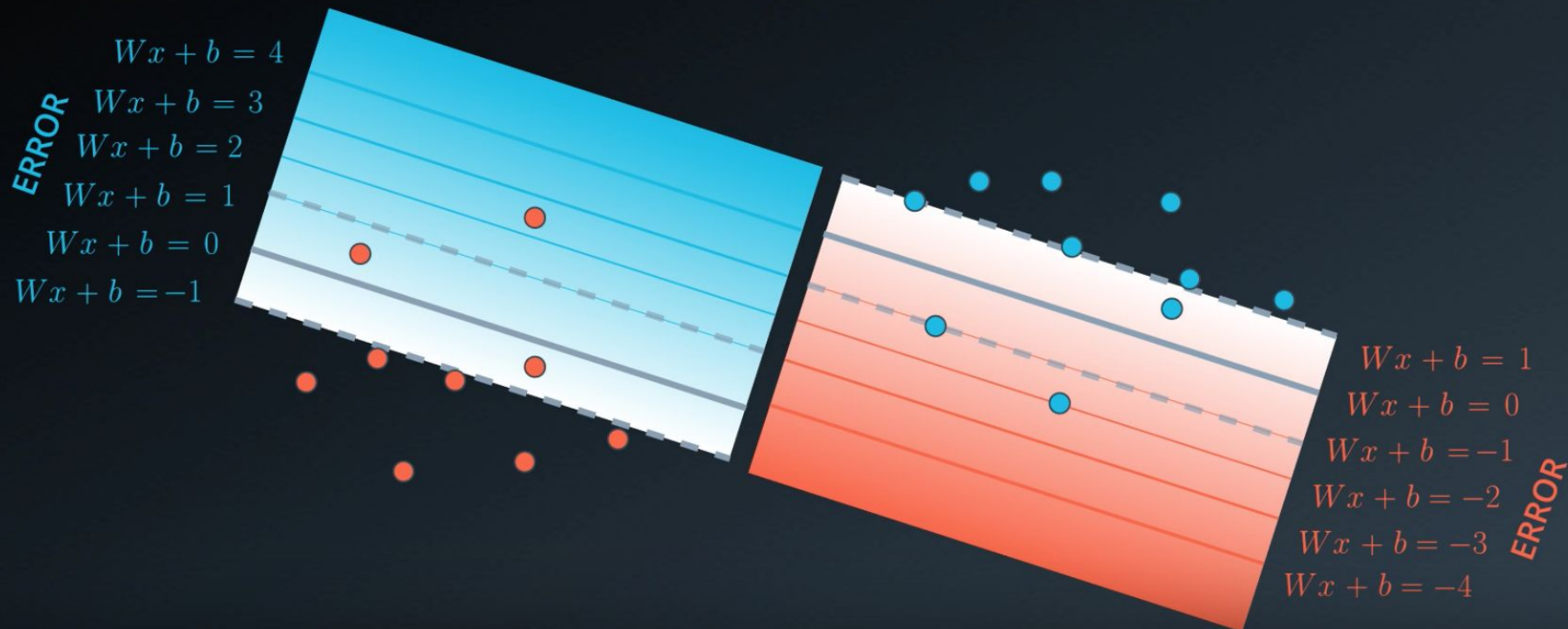
- Digamos que nuestra línea es $Wx + b = 0$
- El margen lo definimos con
 - $Wx + b = 1$
 - $Wx + b = -1$
- Calculamos el error a partir

de las líneas hacia arriba y

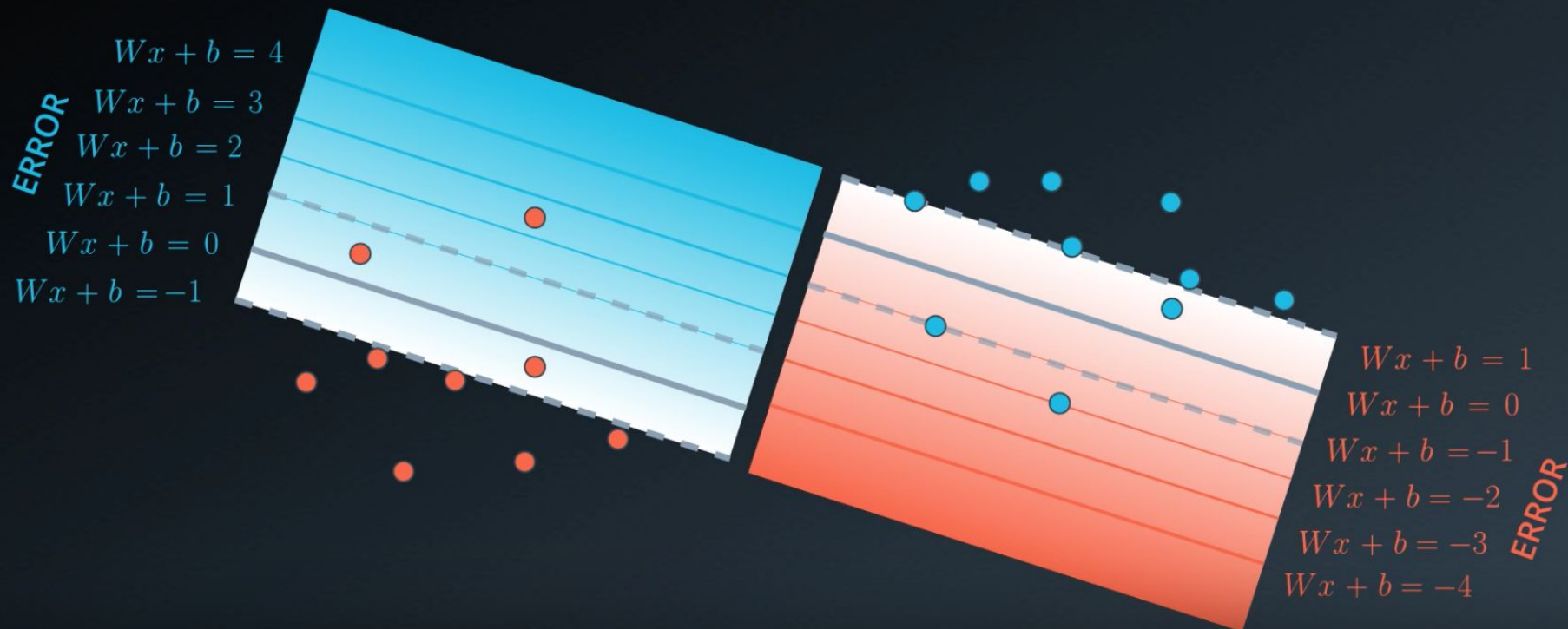
hacia abajo.



¿Cuál es el error?

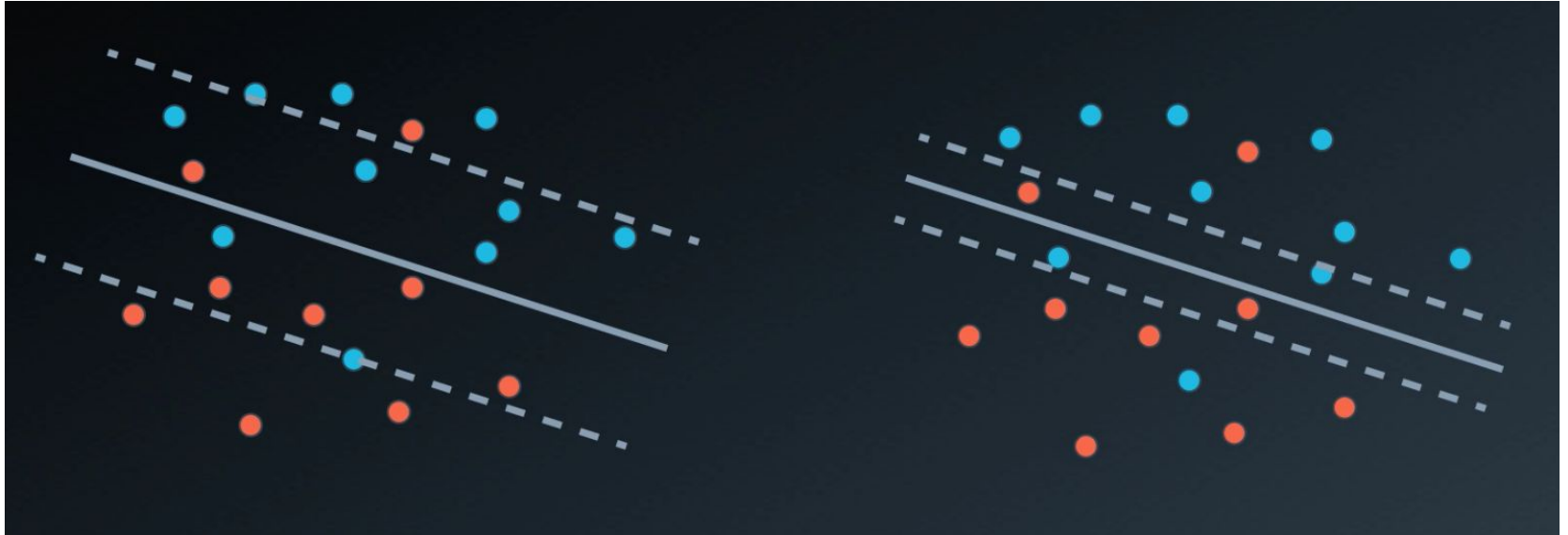


¿Cuál es el error? aproximadamente 10



Error de margen

- Error que podemos optimizar con Gradient Descent.
- Mientras más grande es el margen, más pequeño es el error

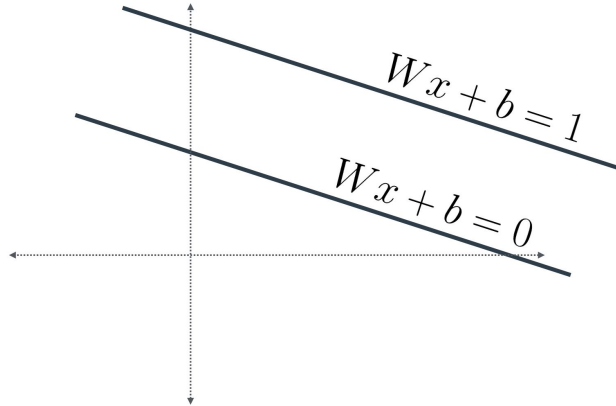


Cálculo de error de margen

- Tenemos tres líneas
 - $Wx + b = 1$
 - $Wx + b = 0$
 - $Wx + b = -1$
- El objetivo es calcular la distancia de la primera a la tercera.
 - Podemos calcular de la primera a la segunda y duplicarla

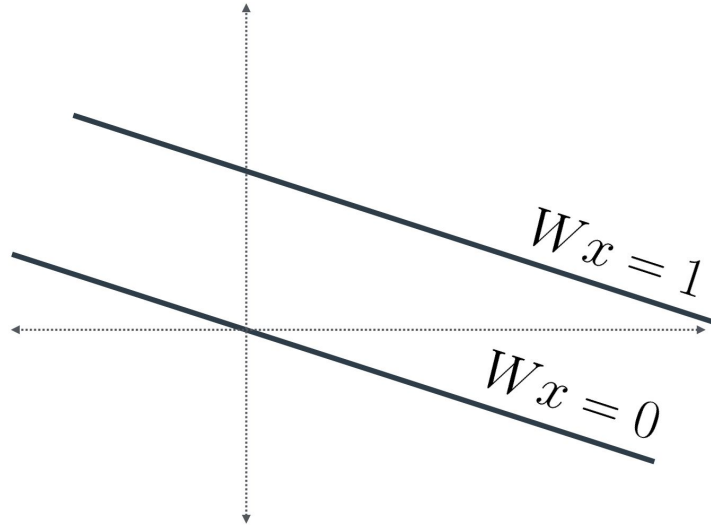
Cálculo de error de margen

- Tenemos tres líneas
 - $Wx + b = 1$
 - $Wx + b = 0$
 - $Wx + b = -1$
- El objetivo es calcular la distancia de la primera a la tercera.
 - Podemos calcular de la primera a la segunda y duplicarla

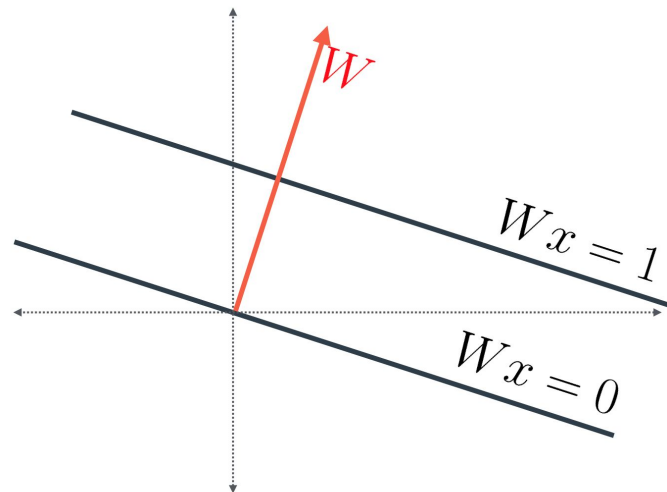
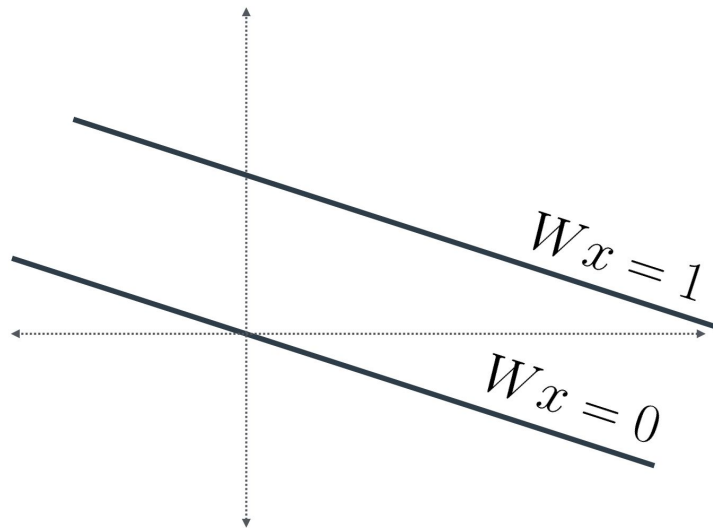


Cálculo de error de margen

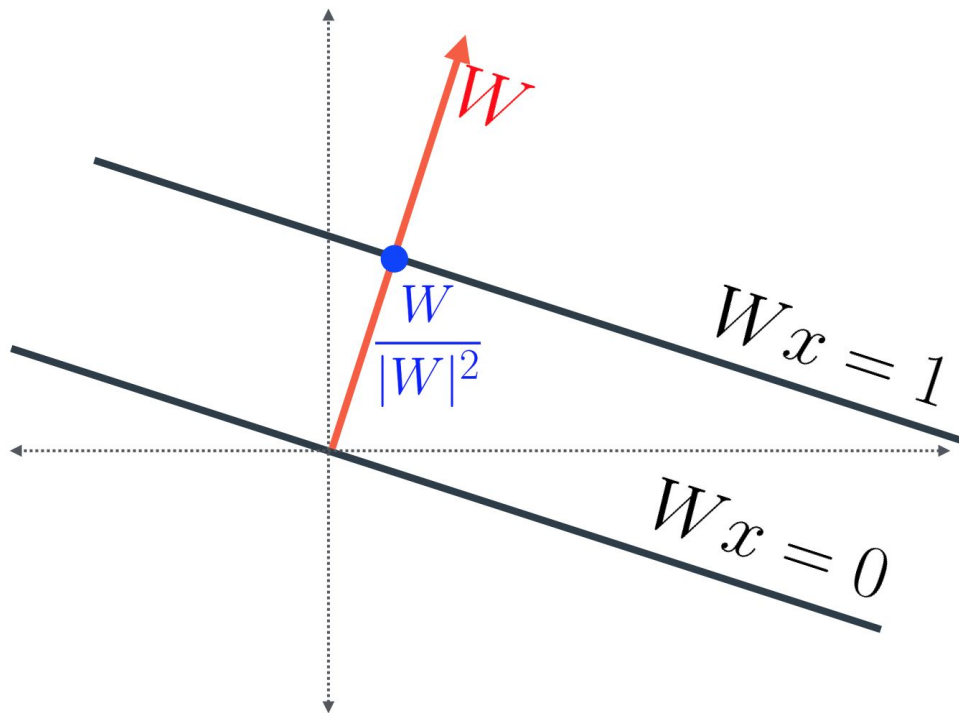
- Podemos simplificar haciendo que $b=0$



Cálculo de error de margen

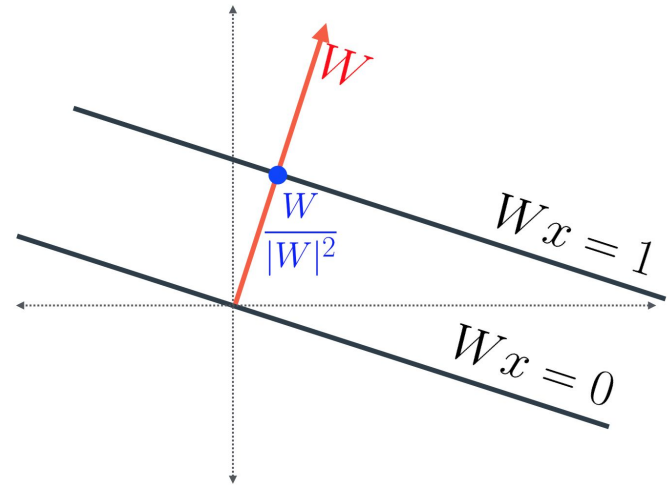


Cálculo de error de margen

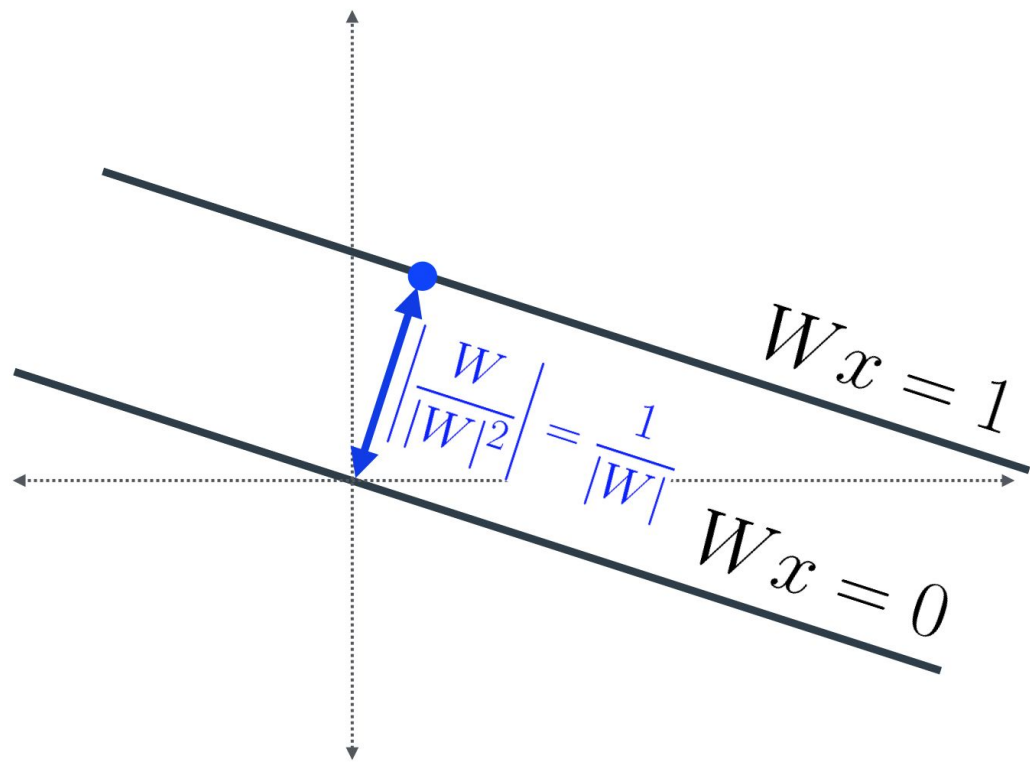


Cálculo de error de margen

- Asumamos que la intersección de W es en (p, q) , el cuál debe ser un múltiple de (w_1, w_2) .

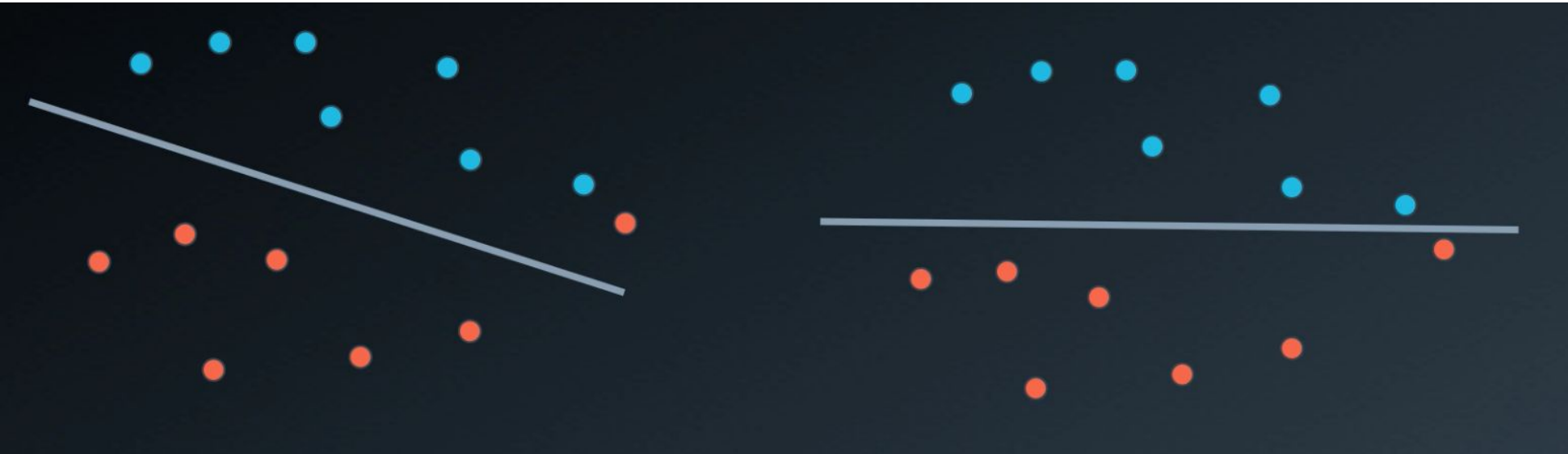


Cálculo de error de margen



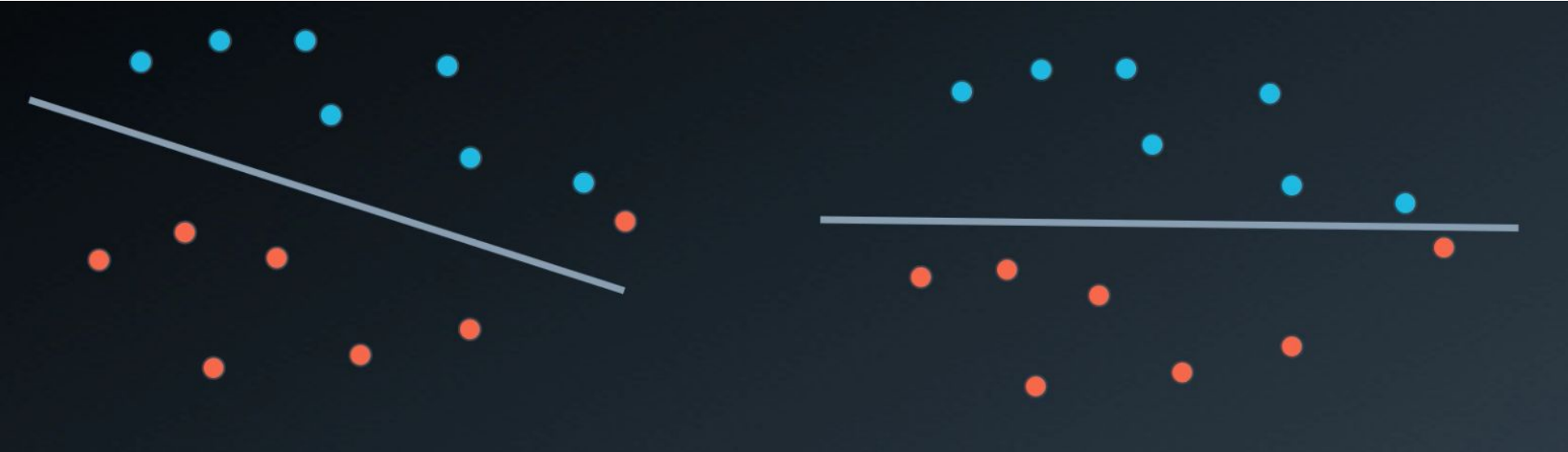
Parámetro C

- ¿Cuál es mejor?



Parámetro C

- Depende del tipo del problema



Parámetro C

- Necesitamos niveles de flexibilidad
 - El parámetro C es una constante del error de clasificación
-
- $E = C * \text{Error Clasificación} + \text{Error Margen}$
-
- C (es inversamente proporcional a lambda)
 - Largo -> propenso a overfitting, margen pequeño
 - Pequeño -> margen grande

Kernel Polinomial

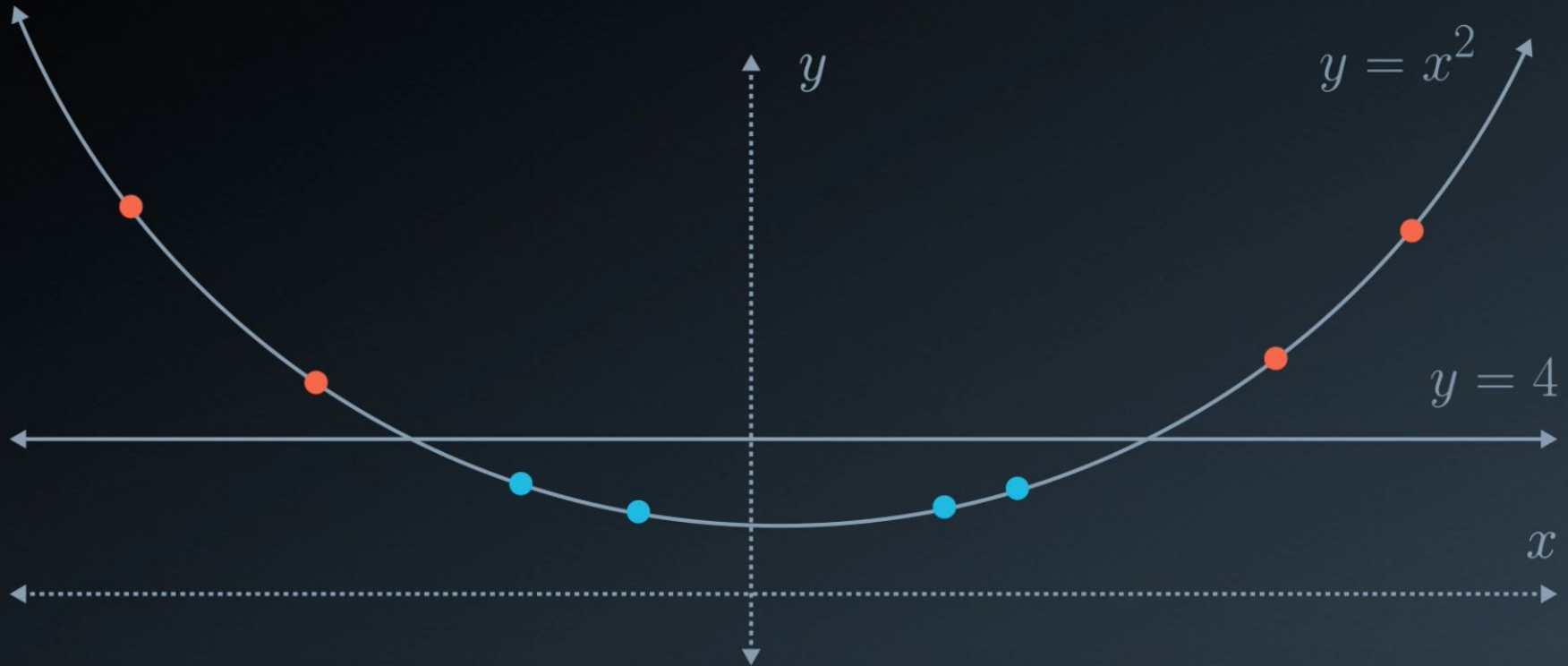
- Podemos agregar una nueva dimensión con el truco del kernel

Kernel Polinomial

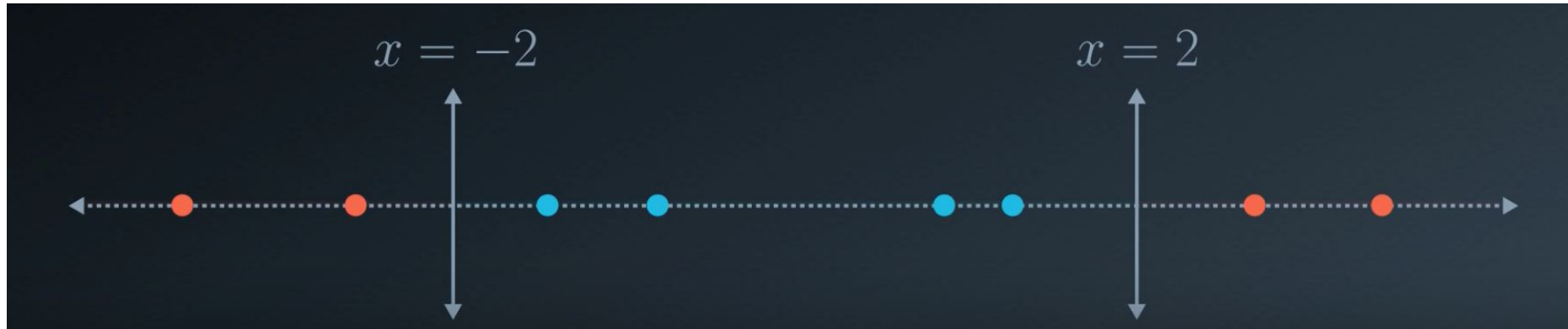
- Necesitamos un modelo más complejo



Kernel Polynomial



Kernel Polynomial



Método del Kernel

- Otro método de separar info
- ¿Cómo podemos separar los puntos?



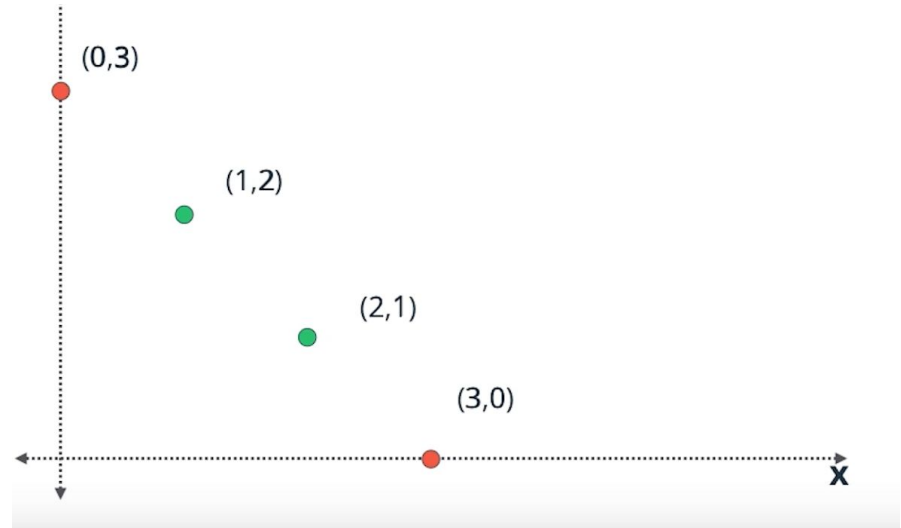
Método del Kernel

- Otro método de separar info
- ¿Qué ecuación nos puede ayudar?

$x + y$

xy

y^2

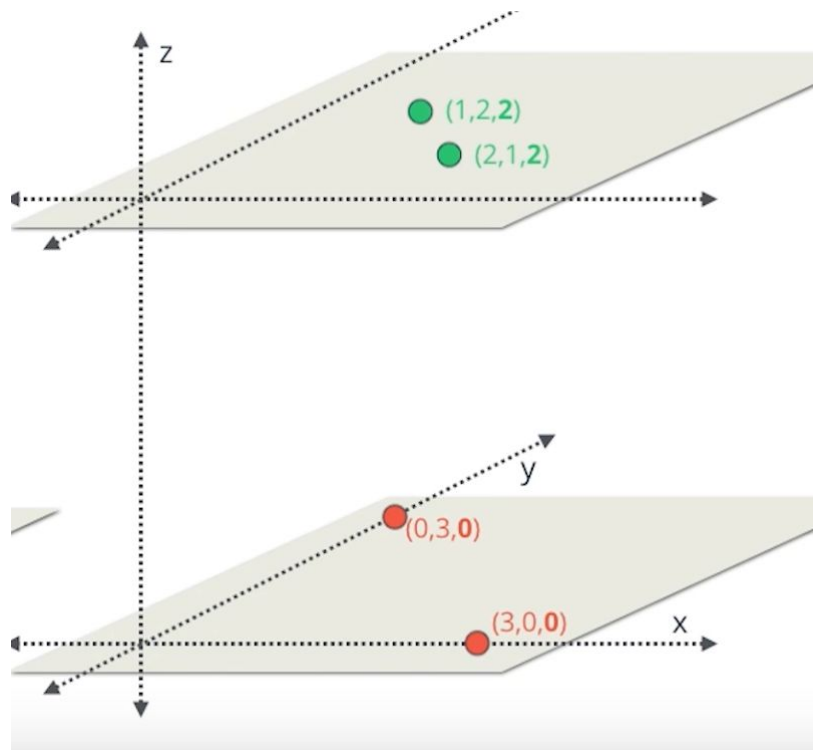


Método del Kernel

- Otro método de separar info
- ¿Qué ecuación realmente separa?
- Utilicemos el resultado como z (una nueva dimensión)

	(0,3)	(1,2)	(2,1)	(3,0)
$x+y$	3	3	3	3
xy	0	2	2	0
x^2	0	1	4	9

Método del Kernel



	$(0,3)$	$(1,2)$	$(2,1)$	$(3,0)$
$x+y$	3	3	3	3
xy	0	2	2	0
x^2	0	1	4	9

SVM

- Podemos hacerlo polinomial
- Podemos agregar más dimensiones
- ¡Pero es lo mismo!

Grado del kernel

- Si $d=2$

$$\begin{array}{cc} x & y \\ x^2 & xy & y^2 \end{array}$$

- Si $d=3$

$$\begin{array}{cc} x & y \\ x^2 & xy & y^2 \\ x^3 & x^2y & xy^2 & y^3 \end{array}$$

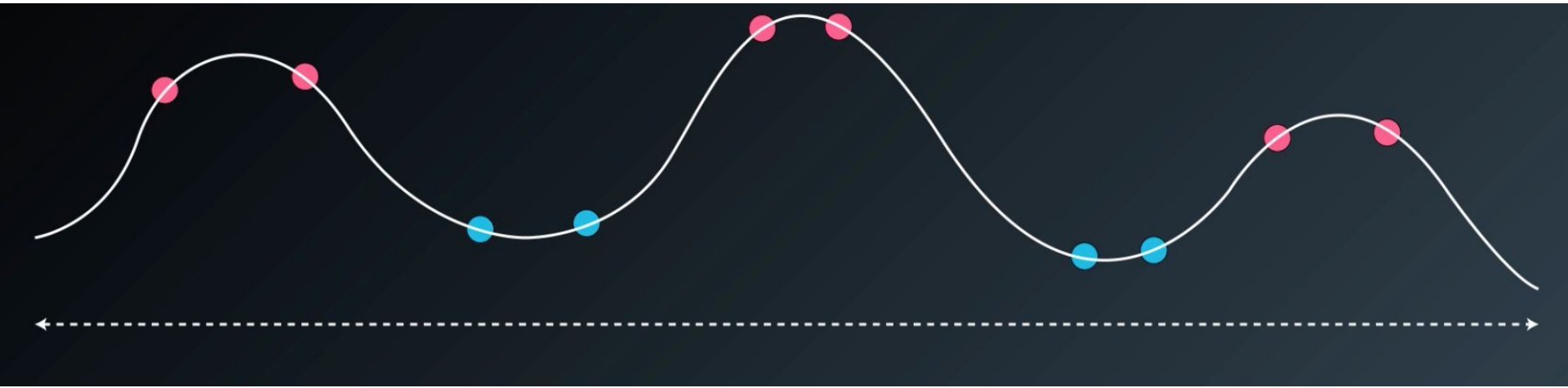
Kernel RBF

- Podemos crear una función base radial sobre cada punto



Kernel RBF

- Podemos crear una función base radial sobre cada punto



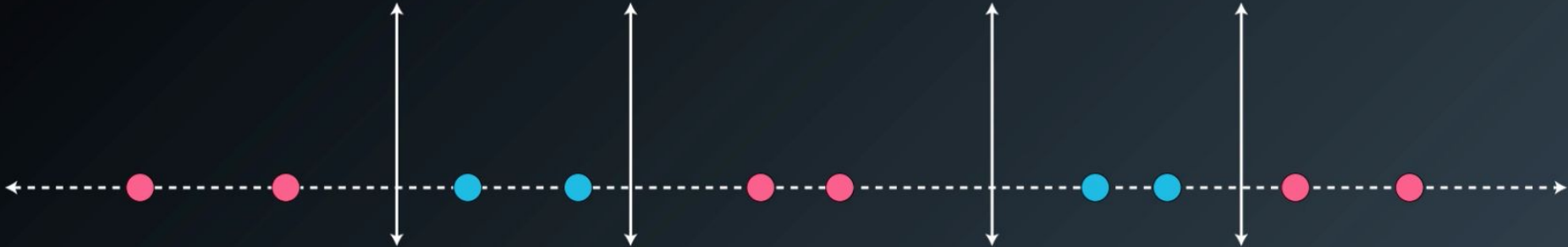
Kernel RBF

- Podemos crear una función base radial sobre cada punto



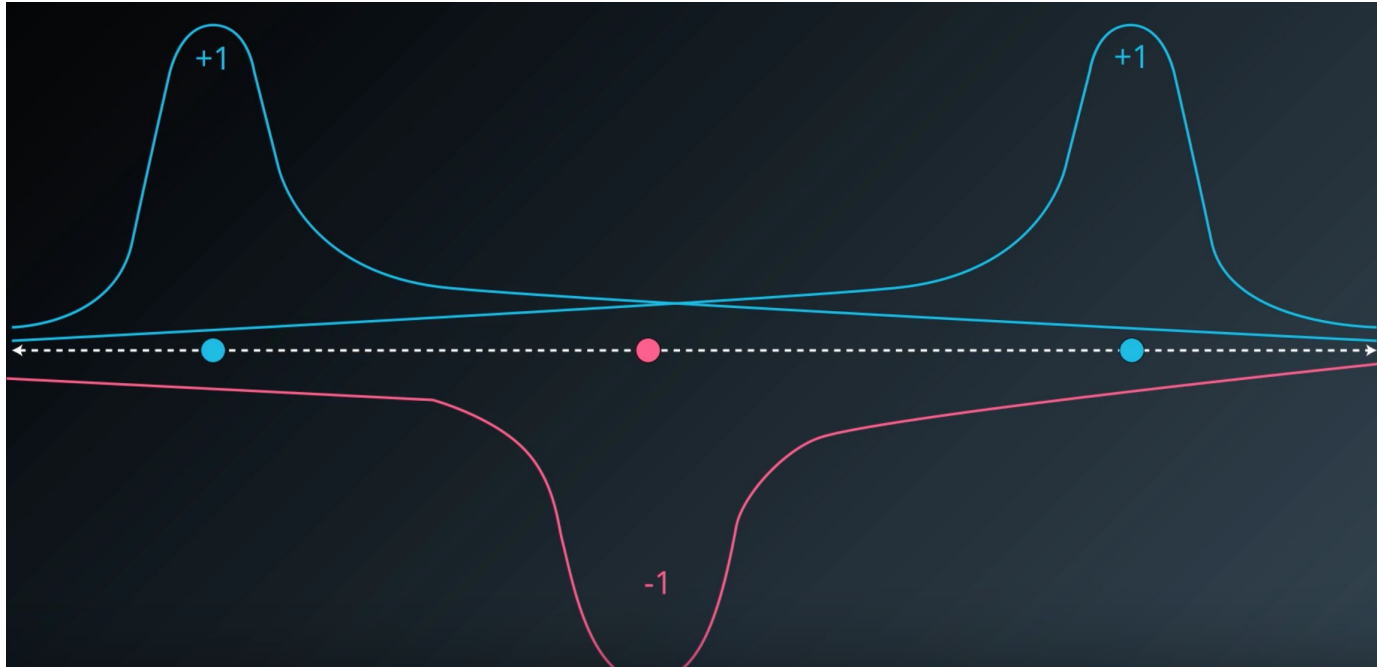
Kernel RBF

- Podemos crear una función base radial sobre cada punto

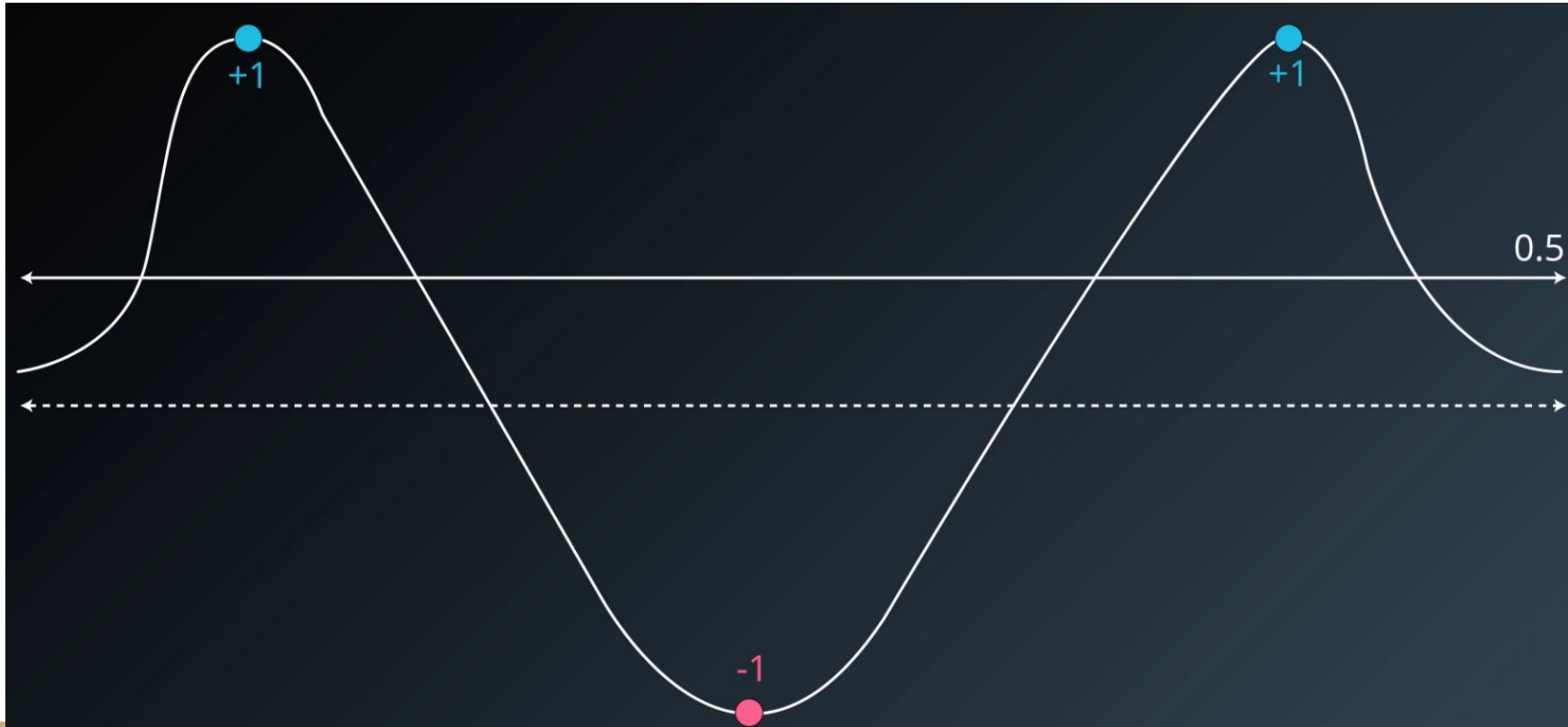


Kernel RBF

- Combinamos los RBFs de cada punto (volteamos para otras clases)

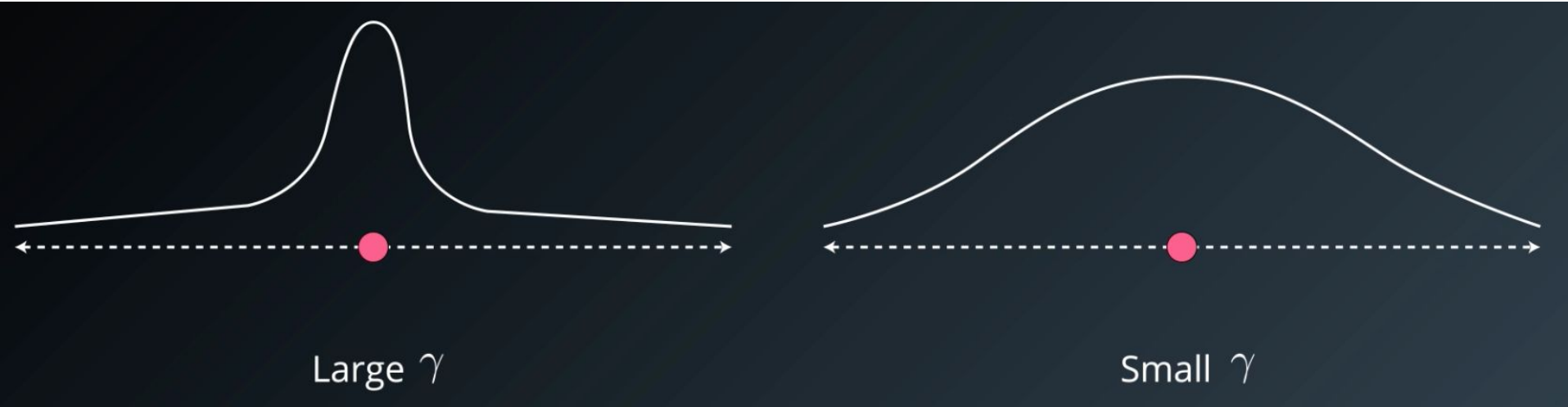


Kernel RBF



Gamma

- Gamma nos dice qué tan delgado son las curvas
 - Más propenso a overfitting



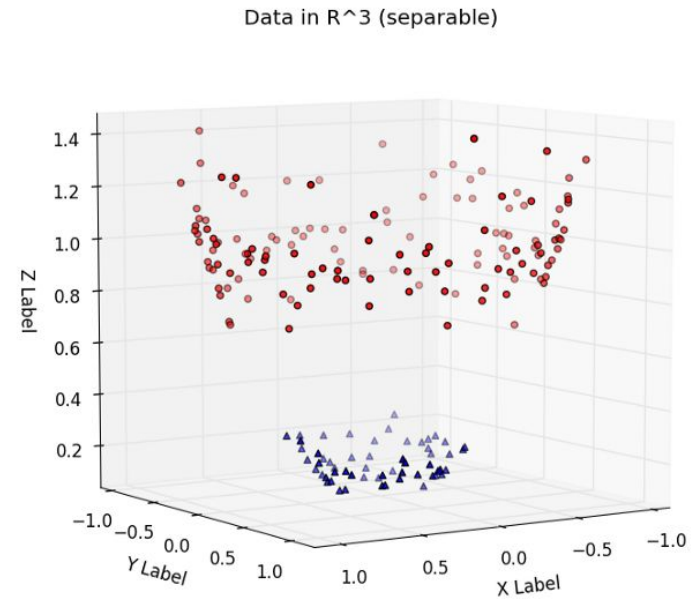
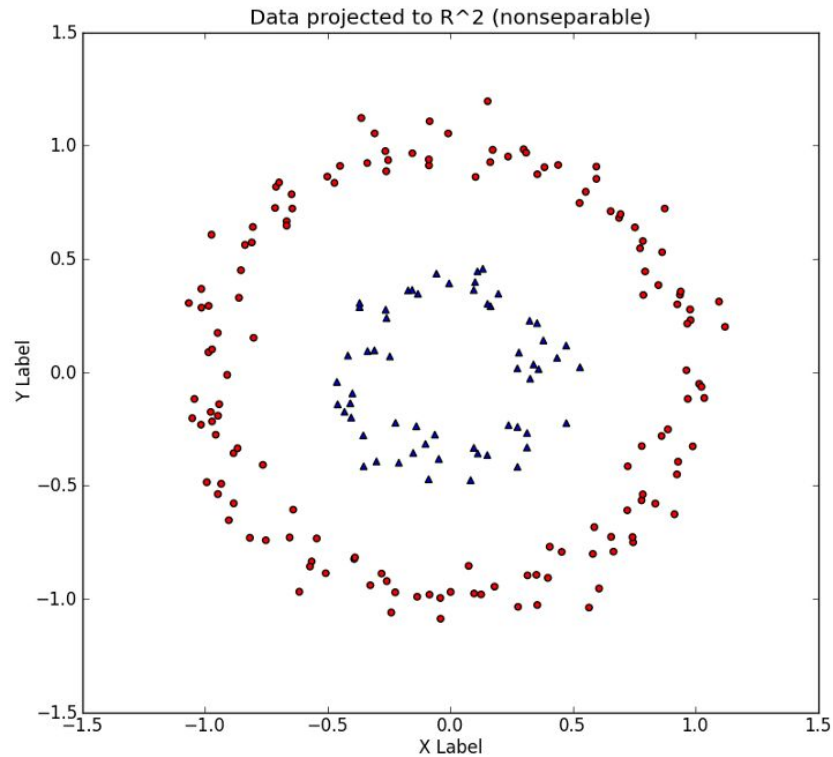
Gamma

- Gamma alto es más propenso a overfitting
- Relacionado con la desviación estándar (inversamente proporcional)

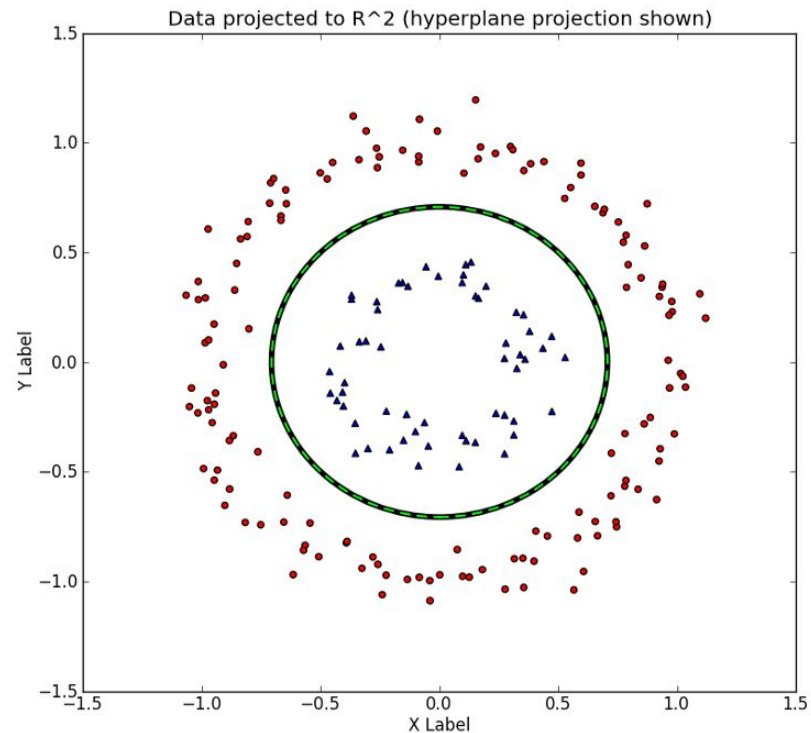
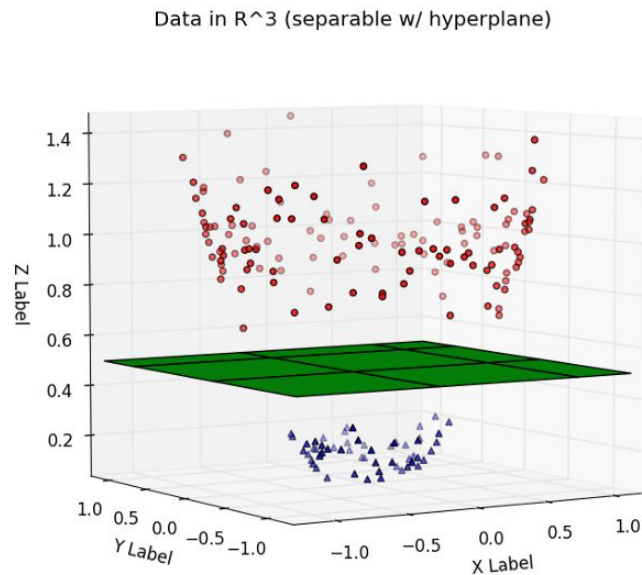
SVMs

- Excelente manera de aprender funciones no lineales
- Uno de los métodos más comunes es con optimización de Lagrange.
 - Se obtiene un hiperplano que separa basándose en los vectores de soporte
- Si no se puede separar la información limpiamente
 - Permitimos más errores o
 - Aumentamos dimensiones

SVM

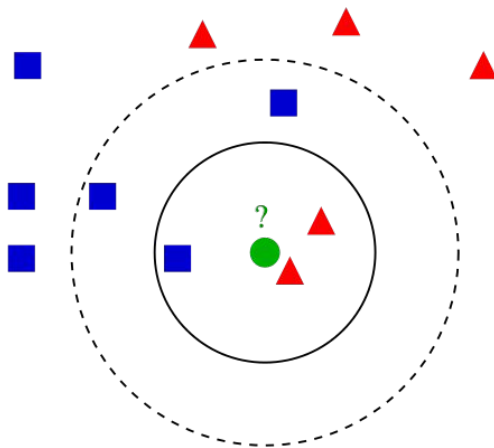


SVM



k-nearest neighbors (k-NN)

- “Eres el promedio de tus k amigos más cercanos”
- Extremadamente sencillo
- Usamos el promedio o la moda del label para los k puntos más cercanos



k-NN

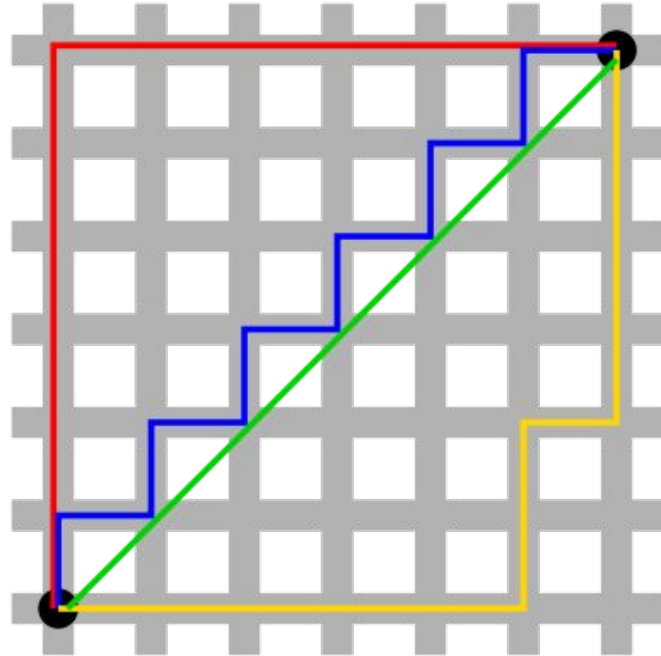
- Sorprendentemente, puede funcionar mejor que modelos paramétricos

k-NN

- Guardamos la información de entrenamiento
 - Ordenamos información según similitud
 - Saca el promedio de los k puntos más cercanos.
-
- Bueno para información compleja donde una simple función no puede describir la relación entre X y Y.

Calculando distancia

- Distancia Euclidiana
- Distancia Manhattan



¿Cómo elegir k?

- Usamos **cross-validation**
- k más alta previene overfitting
- Si es muy alto, el modelo tendrá bias

Usos de k-NN

- Detección de fraude
 - Es extremadamente rápido
- Predicción de costos de casa
- Completar info de entrenamiento que falta
 - Si falta info de una columna, puedes rellenarla con k-NN