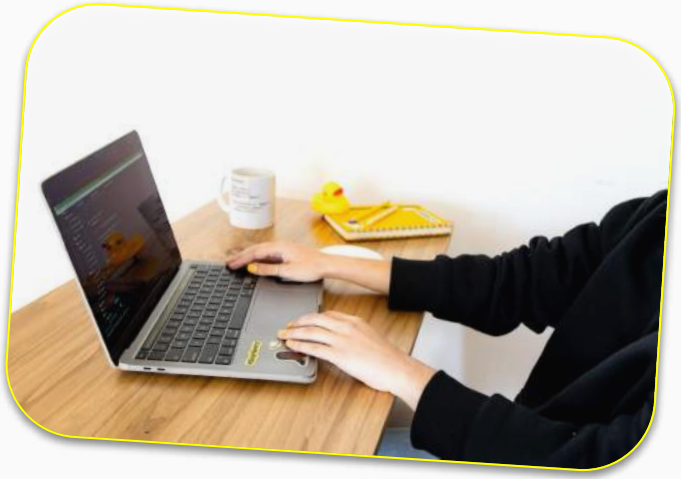
HENRY

Modelos de Clasificación

Data Science

Agenda



- Modelos de Clasificación
- Regresión Logística
- Árboles de decisión
- Vecinos más cercanos
- Support Vector Machine

OBJETIVOS DE LA CLASE

Al finalizar esta lecture estarás en la capacidad de...

- ➜ Adaptar el modelo de regresión lineal a un problema de Clasificación (Regresión Logística).
- ➜ Comprender y diferenciar el funcionamiento de los algoritmos: Árbol de decisión, Vecinos más cercanos y Support Vector Machine.

Al **finalizar** cada uno de los temas, tendremos un **espacio de consultas**.

Hay un **mentor** asignado para responder el **Q&A**.

¡Pregunta, pregunta, pregunta! :D

Modelos y algoritmos

Características

- Basado en un determinado número de ejemplos en caso de supervisión, o agrupando por similitud.
- Se busca generalizar, aprender conceptos a partir de un conjunto de ejemplos y sus características. Cuantos más ejemplos, probablemente sea más fácil la tarea.
- Son robustos sistemas de regresión, capaces de ajustarse a una altísima dimensionalidad y una enorme complejidad, difícil de entender.
- El aprendizaje inductivo consiste en construir un modelo general a partir de información específica (instancias).

Características

- El Sesgo Inductivo de un algoritmo de aprendizaje es el conjunto de afirmaciones que el algoritmo utiliza para construir un modelo.
- Forma de las hipótesis (número y tipo de parámetros).
- Características del funcionamiento del algoritmo. Como principio metodológico, ante igualdad de condiciones (por ejemplo, igual desempeño), debemos elegir al modelo más simple porque esperamos que generalice mejor.

Trabajo con algoritmos

1. Selección del algoritmo: Elegir por diferentes criterios el que se va a emplear y testear.
2. Entrenamiento: Conforme al algoritmo escogido y los datos que se tienen hay que ver si el entrenamiento da resultados.
3. Evaluación de Calidad: Se utilizan métricas y métodos para decidir si el algoritmo es adecuado o se debe modificar.

Trabajo con algoritmos

4. Ajuste de hiper parámetros: Se los modifica según el tipo de situación, los datos y las métricas arrojadas durante y tras el entrenamiento realizado. Volver al paso 2.

5. Objetivos y Métricas: Si se está satisfecho, fin de la tarea y modelo entrenado, si no es así, debemos volver al paso 1.

Modelos de **clasificación**

¿Cómo?

En los problemas de clasificación utilizamos algoritmos de ML que nos permiten diferenciar si un conjunto de datos pertenece a una determinada clase o a otra/s.

El resultado de nuestra función:

$$f(X) = y$$

es una etiqueta de categoría, por lo que el algoritmo debe discernir si ciertos valores/atributos pertenecen a cierta categoría o no.

Aprendizaje supervisado: pasos indispensables

1. En la primera etapa, tomamos nuestros datos e identificamos como variables predictoras X a algunas características, y en una variable y al atributo que queremos predecir.

   Cada componente x (instancia) de X (todas las instancias) tiene asociada una etiqueta y en Y.

Aprendizaje supervisado: pasos indispensables

2. Le "mostramos" pares (x,y) –pares asociados de variables predictoras y etiquetas– a un modelo preparado para aprender de nuestros datos, de forma tal que crea un conjunto de reglas o asociaciones para, dada una entrada x, predecir su salida y.

   Esta fase se conoce como entrenamiento.

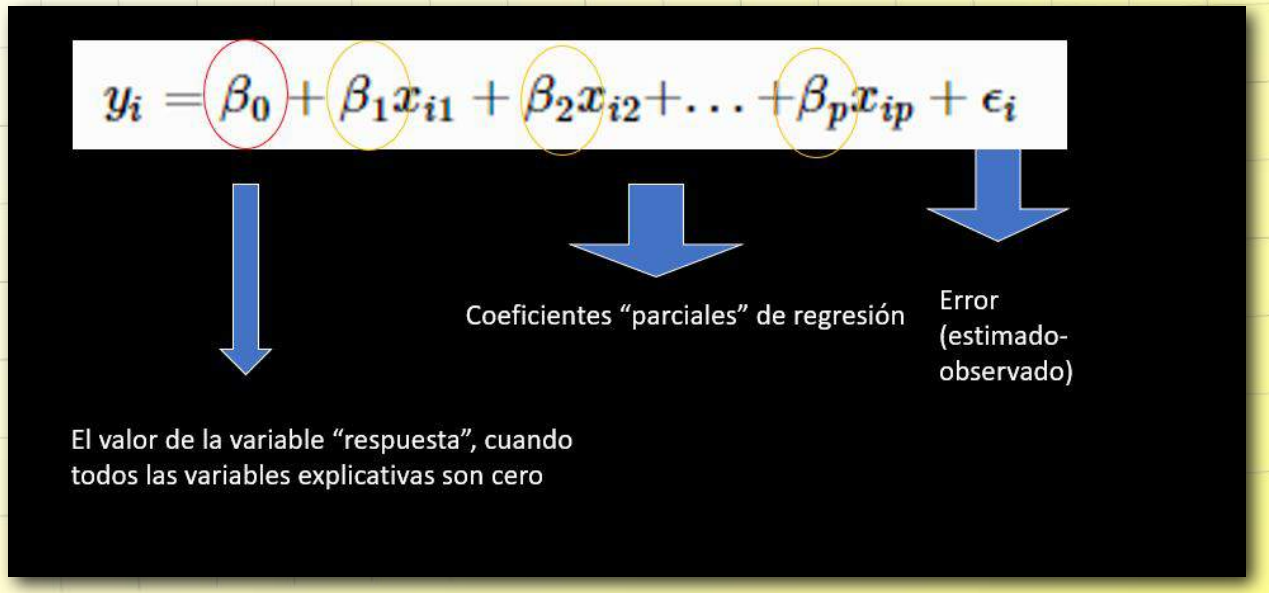
Aprendizaje supervisado: pasos indispensables

3. Una vez que el modelo está entrenado, podemos utilizarlo.

   En esta etapa -denominada predicción-, lo que esperamos es que el modelo nos diga la etiqueta correspondiente a instancias cuya etiqueta no conocemos.

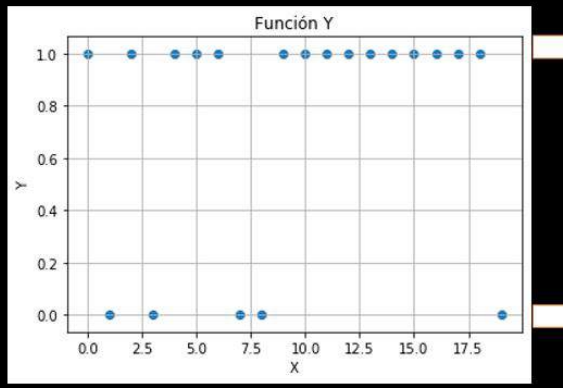
Regresión logística

Regresión Lineal Vs. Regresión Logística

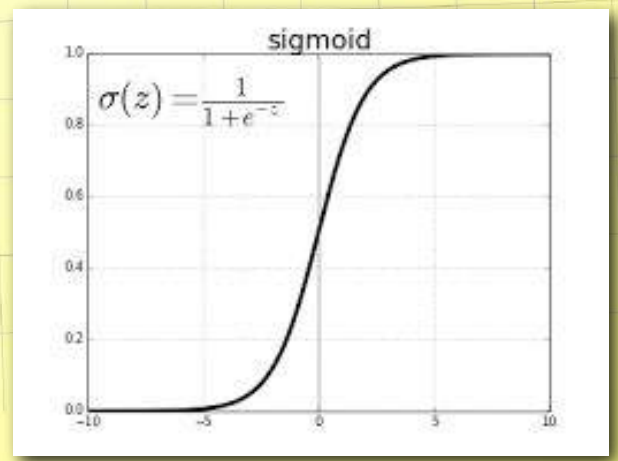


Regresión logística

$$Y_i = \begin{cases} 1 : \text{Evento exitoso} \\ 0 : \text{en caso contrario} \end{cases}$$



La variable objetivo es dicotómica (0 para ausencia y 1 para presencia). En este caso, se utiliza la función sigmoidal cuyo rango está dado entre 0 y 1.



Regresión logística

Al reemplazar en la función sigmoidal la ecuación de regresión lineal, nos queda...

$$P(y = 1|X = x) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}$$

La probabilidad de que la variable objetivo valga 1, para unos valores determinados de las variables explicativas (features).

$$P = \frac{e^y}{1 + e^y}$$

Árbol de decisión

¿Qué es?

Un árbol de decisión es una estructura compuesta de nodos, ramas y hojas. Dada una nueva instancia, ésta es clasificada recorriendo el árbol de decisión: en cada nodo, el árbol hace una pregunta a la instancia sobre alguno de sus atributos.

En esencia, el árbol hace preguntas y va clasificando de acuerdo a las respuestas. Podría pensarse como una combinación if-elif-else y return.

División... ¿Para qué?

La división está basada en el criterio más significativo para diferenciar los elementos:

- ➜ Entropía(t): También conocido como ganancia de información, se utiliza para medir el grado de impureza de una muestra y elegir el atributo que más la reduce.
- ➜ Impureza de Gini: Cuantifica cuán puro es un conjunto. A más homogeneidad en las etiquetas, más puro es, cuanto más alto es su valor, más impura es la muestra.
- ➜ Prune tree: El modelo debe ser lo más simple posible, se poda el árbol hasta que sacar un nodo no mejore la performance del modelo en un test set.

Ejemplo: polillas

Tomemos de ejemplo el caso de las mediciones de las dos especies de polillas.

Queremos analizar esos datos para que, ante una nueva medición, se pueda predecir de qué especie se trata.

masa	envergadura	especie
747	43	Luna
723	47	Luna
760	42	Luna
718	43	Luna
736	43	Emperador
727	48	Luna
719	42	Emperador
765	48	Luna
753	41	Emperador
740	40	Emperador
740	47	Luna
736	46	Luna
760	44	Emperador
740	47	Emperador

Ejemplo: polillas

Se comienza construyendo una pregunta por cada columna y evaluando cuál deja mejor separadas las instancias...

Pregunta 1

masa	envergadura	especie
747	43	Luna
723	47	Luna
760	42	Luna
718	43	Luna
736	43	Emperador
727	48	Luna
719	42	Emperador
765	48	Luna
753	41	Emperador
740	40	Emperador
740	47	Luna
736	46	Luna
760	44	Emperador
740	47	Emperador

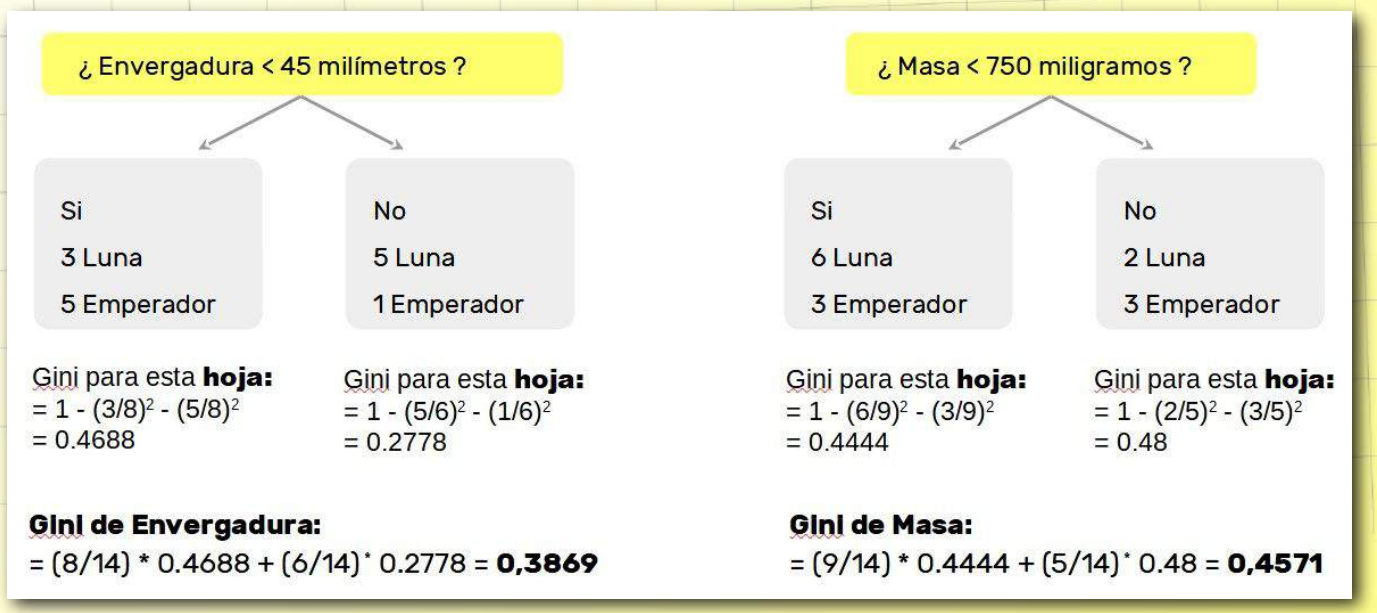
Ejemplo: polillas

Pregunta 2

masa	envergadura	especie
747	43	Luna
723	47	Luna
760	42	Luna
718	43	Luna
736	43	Emperador
727	48	Luna
719	42	Emperador
765	48	Luna
753	41	Emperador
740	40	Emperador
740	47	Luna
736	46	Luna
760	44	Emperador
740	47	Emperador

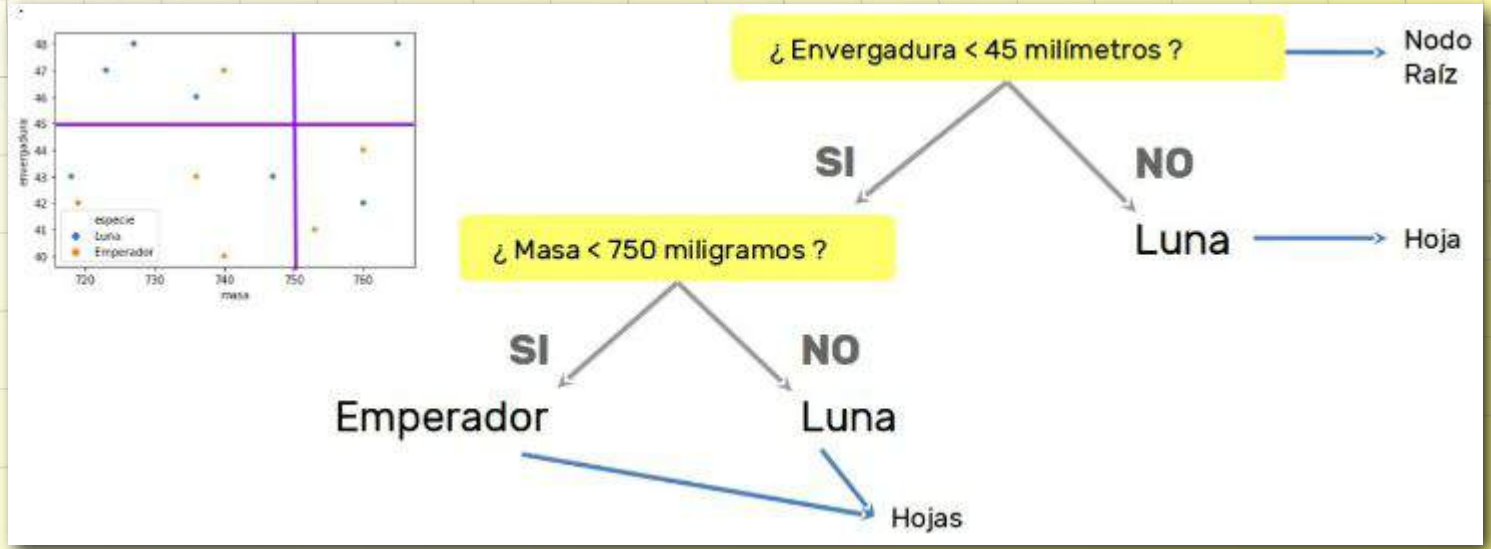
Ejemplo: polillas

¿Cuál de las dos preguntas separó mejor las clases?



Ejemplo: polillas

El árbol de decisión tendría la siguiente forma, con profundidad 2:



Ventajas y desventajas

Ventajas:

- ➜ Fáciles de entender, interpretar y visualizar. Esto ayudará a la hora de comunicar nuestro trabajo.
- ➜ Entrenamiento rápido.
- ➜ Modelo base para otros más complejos (Random Forest, XGBoost, etc.).

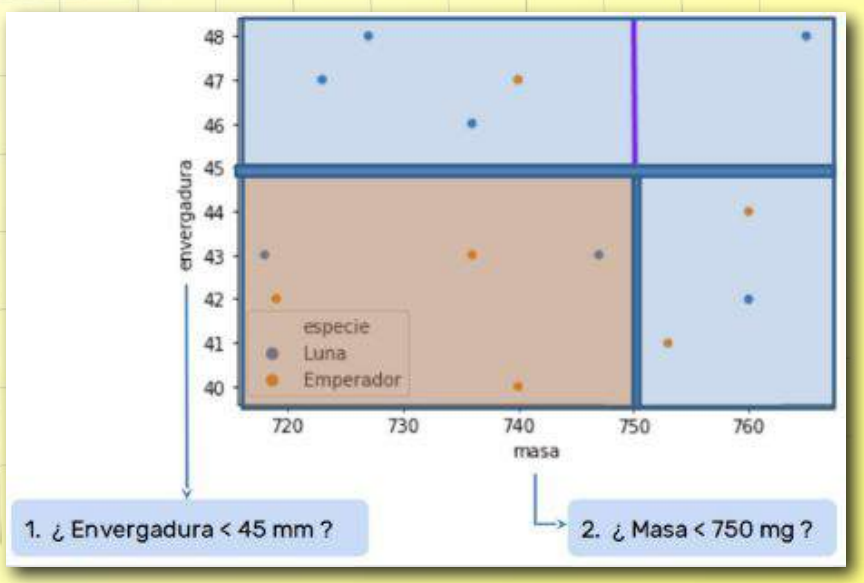
Desventajas:

- ➜ Poder de generalización relativamente bajo en ciertas circunstancias.
- ➜ Desempeño inferior a modelos más modernos.

Fronteras de Decisión

En los distintos modelos, podemos tener visualmente la distribución de los datos y también el resultado del proceso de aplicación algoritmo.

Tomando el caso del ejemplo, podemos observar las líneas de división de ambas especies, aplicando las dos preguntas tenemos los cuadrantes, uno de los cuales pertenecen a cada especie.



vecinos más cercanos

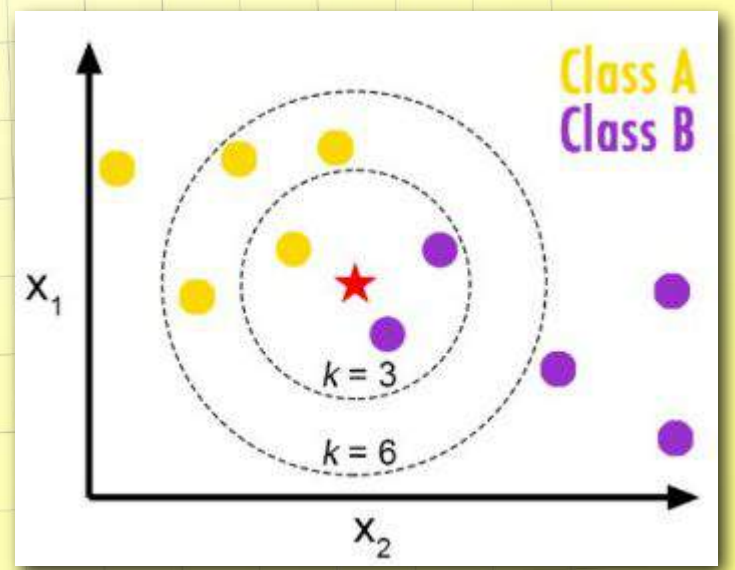
¿Qué es?

Este modelo, también conocido como KNN (K-Nearest-Neighbor), es considerado un aprendiz perezoso, ya que no hay aprendizaje propiamente dicho, sino que predice la clasificación para un nuevo dato buscando a los K vecinos más parecidos.

Dada una nueva instancia de la cual no conocemos la etiqueta objetivo, vamos a asumir que su etiqueta será igual a la de las instancias “vecinas” en el conjunto de datos que tenemos.

Ejemplo

Si tenemos los features X1 y X2, y además un tercer feature que es la clase, que puede ser A o B. Dada una instancia nueva de la que no se conoce su clase, se recurre a sus vecinos más cercanos para clasificarla y K es la cantidad de vecinos que se evalúan para saber la clase de la nueva instancia.



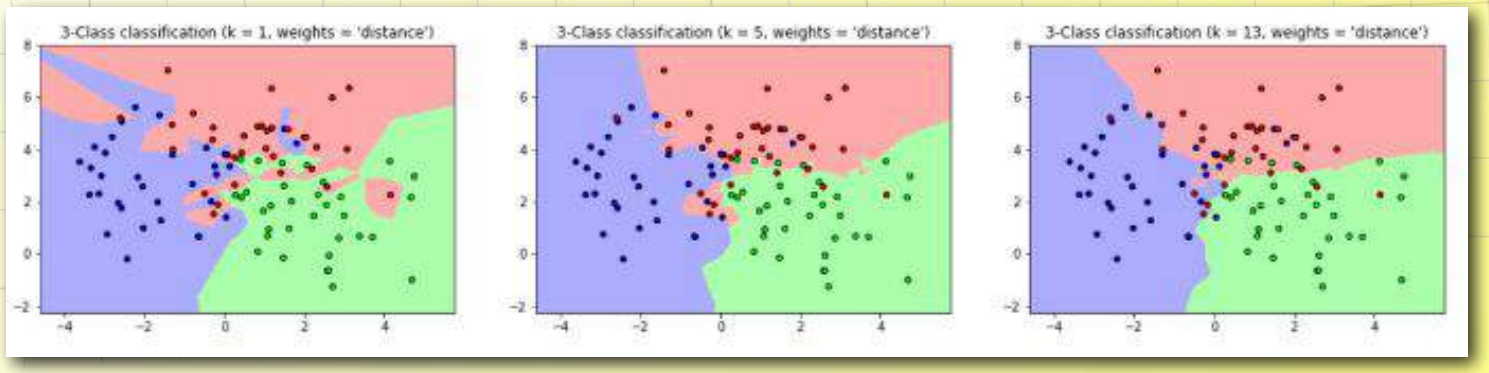
Ejemplo

Siguiendo con el ejemplo anterior:

- K es el hiperparámetro.
- Si tomamos K = 3, entonces vamos a clasificar la nueva instancia como de clase B, ya que habrá dos instancias de B y una de A.
- Si tomamos K = 6, entonces vamos a clasificar la nueva instancia como de clase A, ya que habrán cuatro instancias de A y dos de B.
- No hay una receta para elegir K de antemano, depende del problema. Normalmente la solución es probar varios valores y ver cuál modelo se desempeña mejor.

Valores de k

- Puede ser muy importante reescalar los valores de los features, por ejemplo llevando a escala de z-score, antes de usar este modelo.
- El valor que se elija para K va a ser determinante para el desempeño del modelo.



Ecuaciones

- Variables categóricas: distancia de Hamming (diferencia entre los caracteres de las palabras).

$$D_H = \sum_{i=1}^{k} |x_i - y_i|$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

X	Y	Distance
Male	Male	0
Male	Female	1

- Variables numéricas: distancia Euclidiana, o también puede hacerse mediante la distancia de Manhattan o la de Minkowski.

Euclidean	$\sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$
Manhattan	$\sum_{i=1}^{k} \lvert x_i - y_i \rvert$
Minkowski	$\left(\sum_{i=1}^{k} \lvert x_i - y_i \rvert^q \right)^{1/q}$

Ventajas y desventajas

Ventajas:

- ➜ Simple de interpretar.
- ➜ Rápido para entrenar.
- ➜ Nuevos datos no impactan en la precisión del algoritmo.
- ➜ El modelo se comporta bien ante un problema de clasificación multiclase.

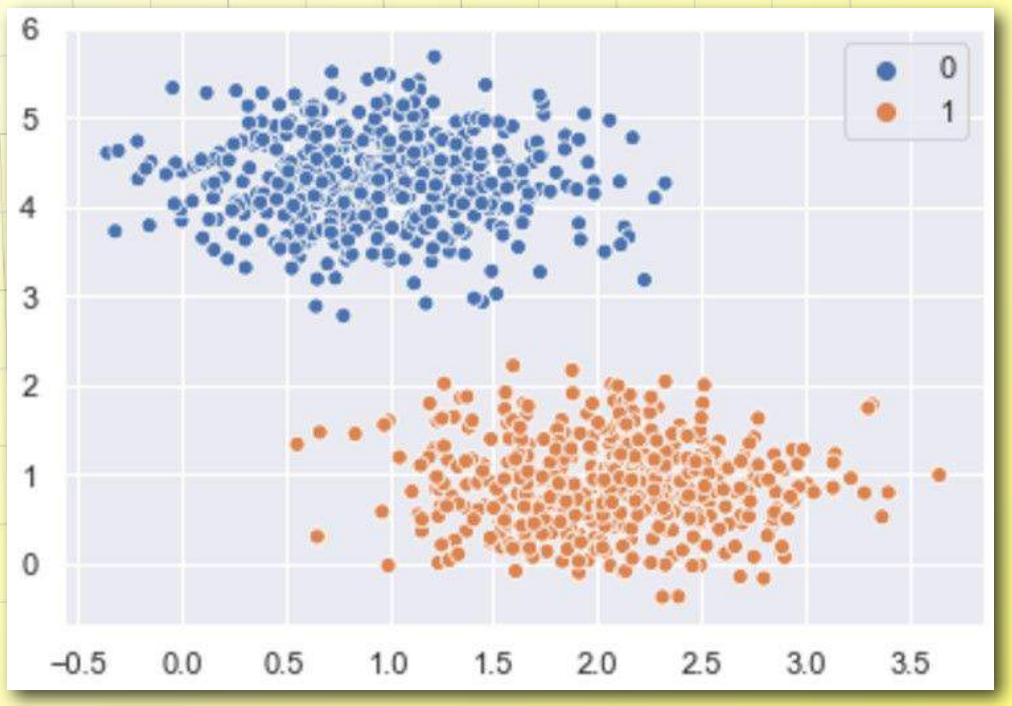
Desventajas:

- ➜ Lento para predecir.
- ➜ Ocupa mucho espacio en disco (guarda todo el set de entrenamiento).
- ➜ La métrica de distancia a elegir no es manifiesta.
- ➜ No se comporta de manera adecuada en datasets con muchas dimensiones.

Support Vector Machine

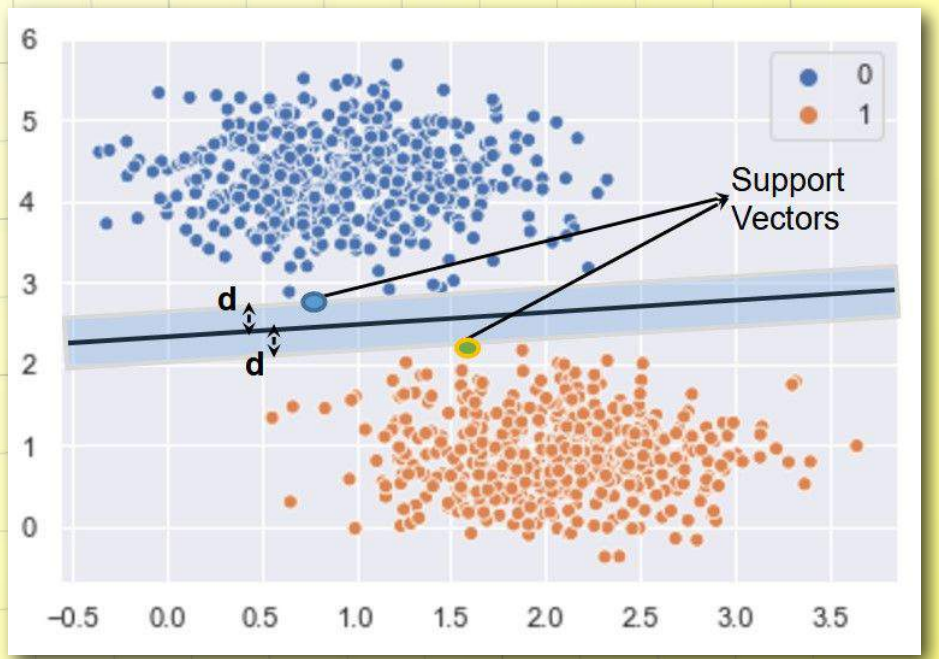
¿Qué es?

El gráfico representa dos grupos de puntos de clases distintas, que se intenta separar...



¿Qué es?

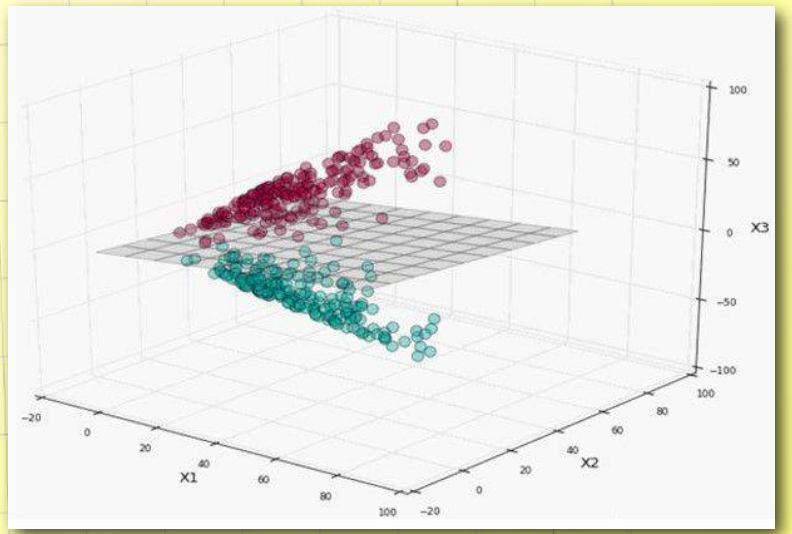
Para hacerlo, la solución más sencilla puede ser con una recta, sin embargo, existen infinitas rectas posibles que separan perfectamente los grupos de datos.



¿Qué es?

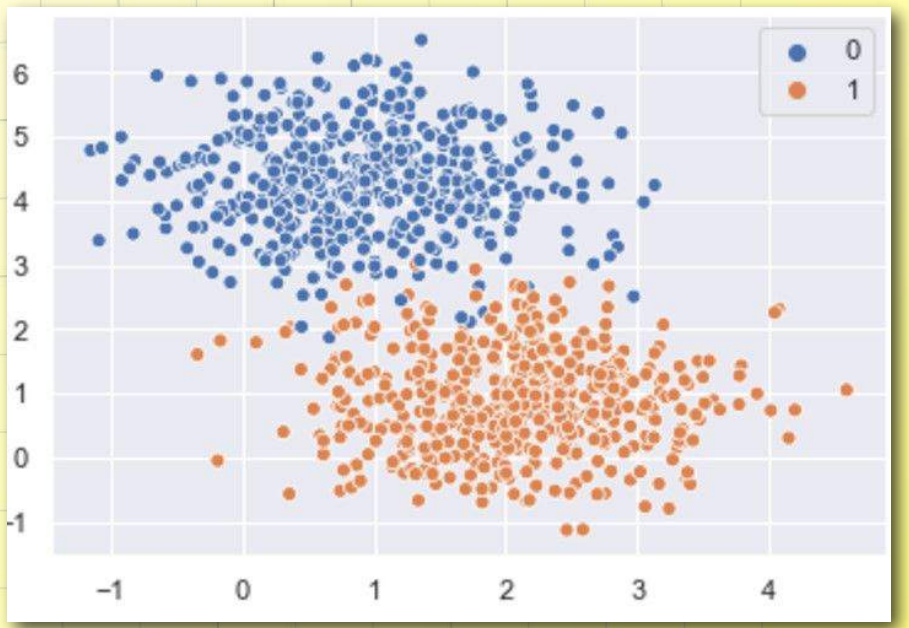
Los datos que pueden ser separados por una recta (o un hiperplano) se conocen como datos linealmente separables. SVM es extensible para n dimensiones.

En este caso, la recta de decisión se transformó en un plano de decisión, por tratarse de un problema de 3 dimensiones.



¿Qué es?

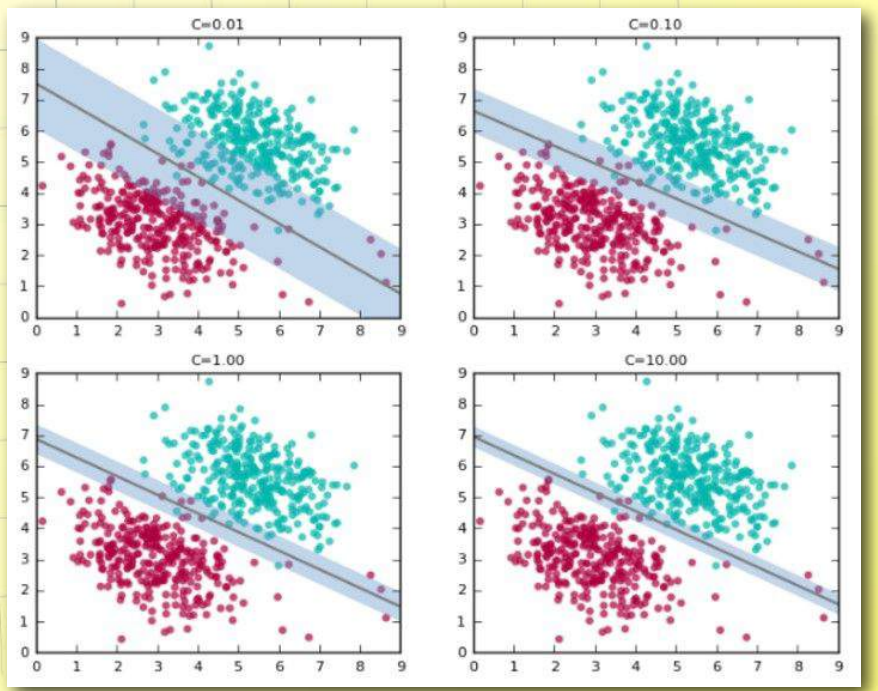
En la realidad, los datos NO suelen ser separables con una recta. Aunque tampoco se quiere descartar por completo al clasificador lineal, ya que parece adecuado para el problema, excepto por algunos puntos erróneos.



¿Qué es?

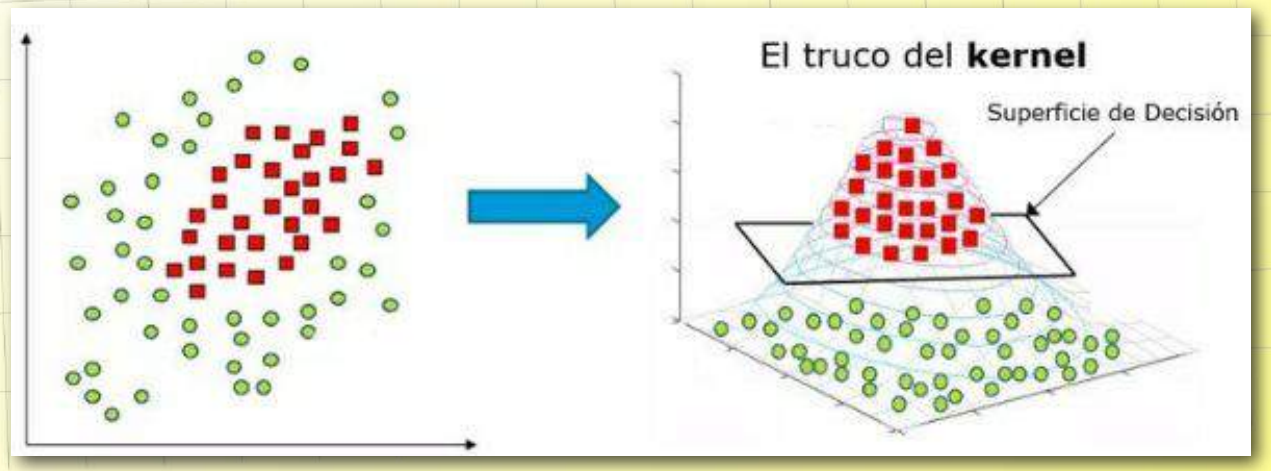
En SVM, se puede especificar cuántos errores estamos dispuestos a aceptar mediante un parámetro llamado C, lo que permite dictaminar la relación entre:

- ➜ Tener un amplio margen.
- ➜ Clasificar correctamente la mayor cantidad de puntos de entrenamiento (C más alto, menos errores en los datos de entrenamiento).



Truco de kernel

Cuando no podemos separar clases linealmente porque no existe un hiperplano de separación, usamos la función o truco de kernel.



¿PREGUNTAS?

¿Alguien dijo Homework?

HENRY

Próxima lecture

Evaluación de modelos I

¡Feedback!

Click on me

Dispones de un formulario en:

- Homeworks
- Guías de clase
- Slack


HENRY