Fundamentos de Ciencias de Datos

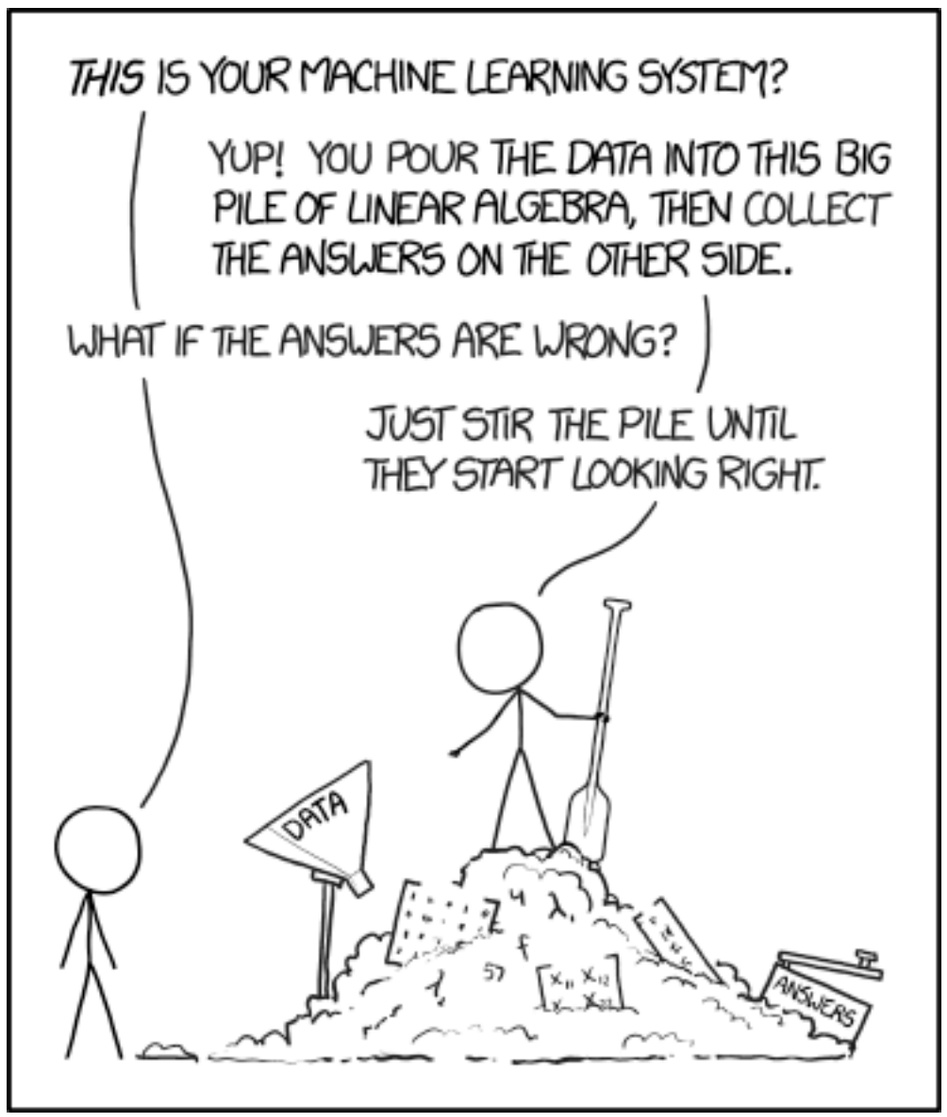
Semana 09 - Introducción a Machine Learning

Machine Learning

¿Qué es *Machine Learning*?

Machine Learning

¿Qué es *Machine Learning*?



Machine Learning

Machine Learning es la ciencia (y el arte) de programar computadores de manera tal que ellos puedan aprender de los datos sin haber dado instrucciones en concreto

Fuente: Hands-on Machine Learning with Scikit Learn and Tensorflow

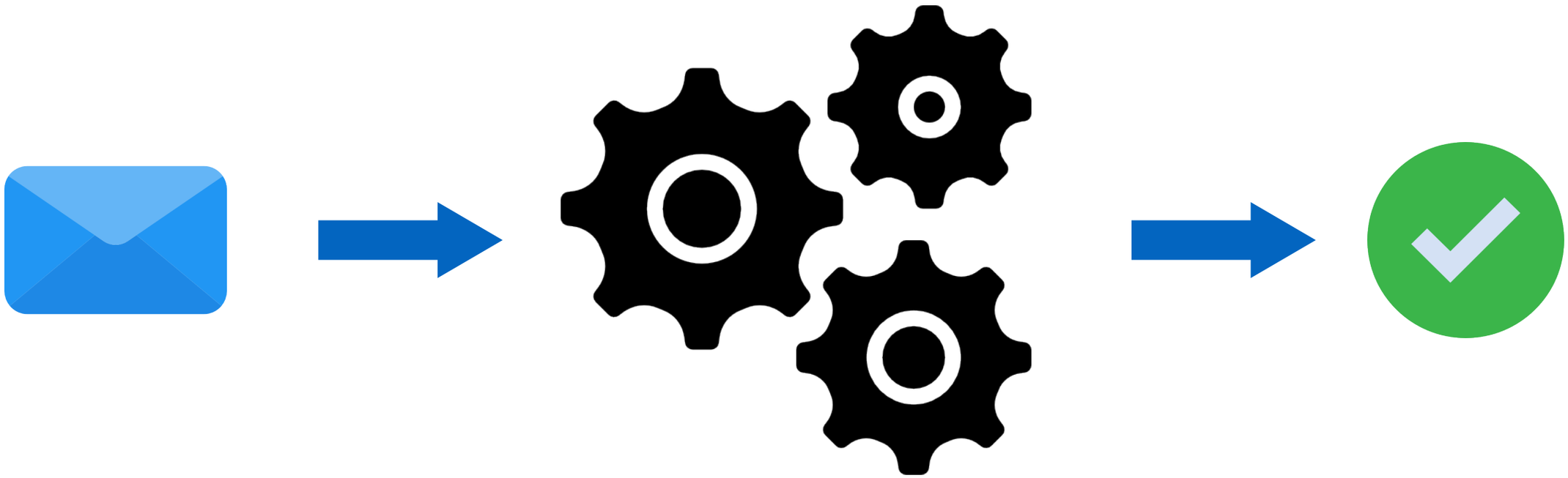
Machine Learning

Expectativa

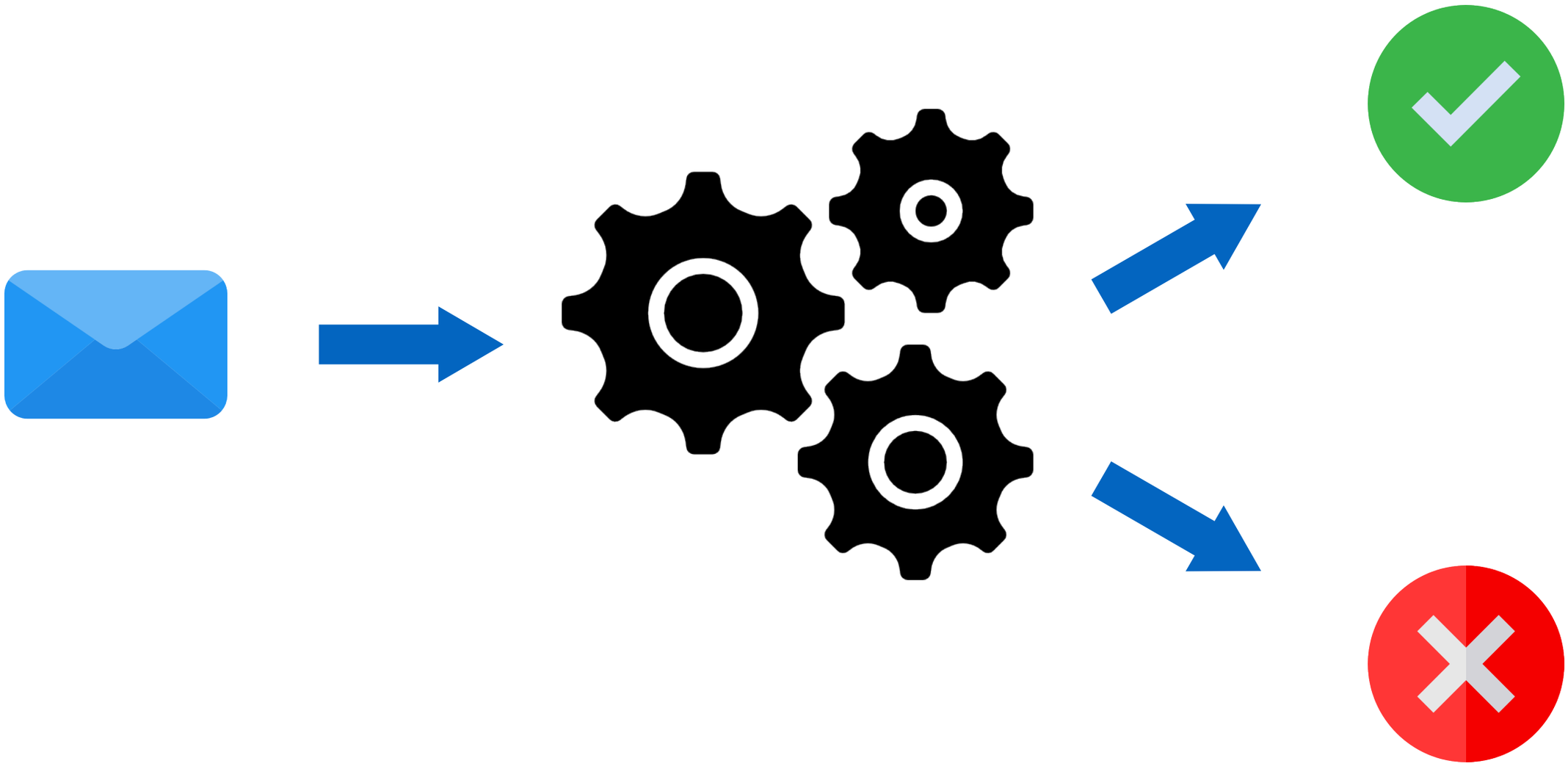
Machine Learning

Sin embargo, uno de los primeros modelos que se hizo famoso fue el clasificador de spam!

Machine Learning



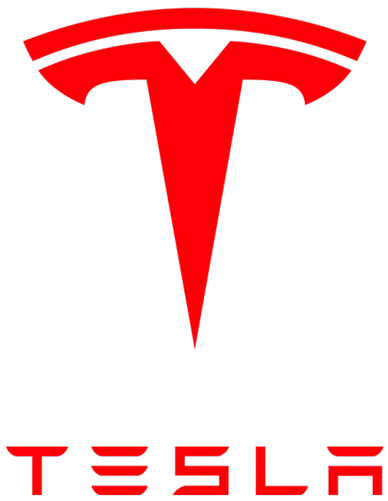
Machine Learning

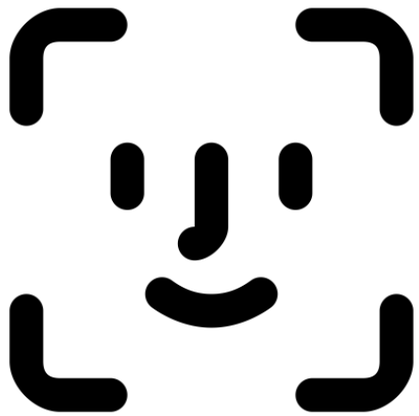


Machine Learning

En la actualidad tenemos aplicaciones en varios campos

Machine Learning





Machine Learning

NETFLIX

TESLA

Machine Learning

NETFLIX

TESLA

Cambridge Analytica

Pero profesor, ¿hacer este tipo de cosas está a nuestro alcance?

Machine Learning

En la actualidad existen muchas librerías y frameworks que ponen a nuestro alcance las herramientas del área de *Machine Learning*

Aunque hay que recordar que existe fundamento teórico importante que viene del campo de la estadística

Machine Learning

El flujo de trabajo en ML

Hay que aprender a realizar un proyecto desde el inicio hasta el final:

- Recolectar datos
- Limpiar datos
- Entender los datos (visualizar, correlaciones, ...)
- Entrenar el modelo (o los modelos)
- Entender su rendimiento
- Analizar errores y mejorar
- Llevar a producción

Machine Learning

El modelo

El modelo es el algoritmo de ML particular que vamos a utilizar y lo vamos a entrenar sobre algunos datos conocidos para hacer predicciones

Machine Learning

Ejemplo de modelo - clasificador de spam

Partimos con un clasificador que no hace nada o funciona mal

Machine Learning

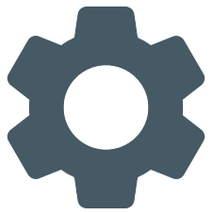
Ejemplo de modelo - clasificador de spam

Pero le comenzamos a mostrar ejemplos

Machine Learning

Ejemplo de modelo - clasificador de spam

Pero le comenzamos a mostrar ejemplos



Machine Learning

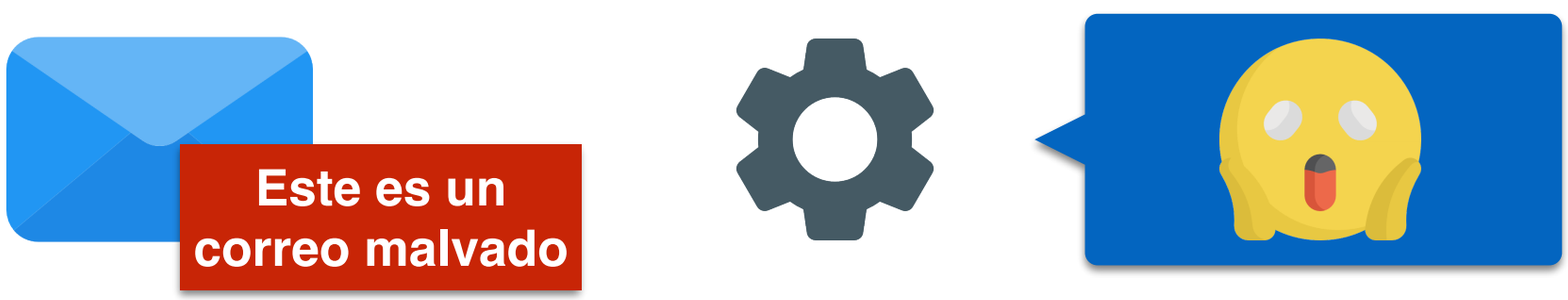
Ejemplo de modelo - clasificador de spam

Pero le comenzamos a mostrar ejemplos

Machine Learning

Ejemplo de modelo - clasificador de spam

Pero le comenzamos a mostrar ejemplos



Machine Learning

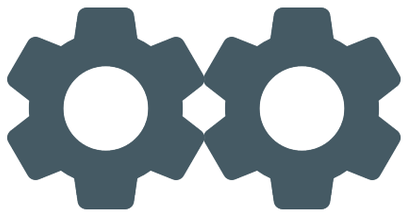
Ejemplo de modelo - clasificador de spam

A medida que nuestro clasificador ve muchos ejemplos, comienza a predecir mejor

Machine Learning

Ejemplo de modelo - clasificador de spam

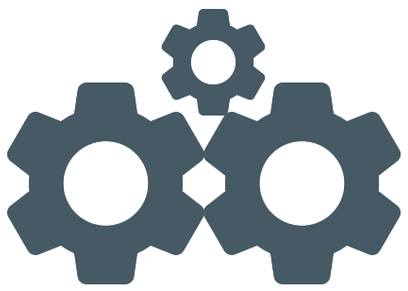
A medida que nuestro clasificador ve muchos ejemplos, comienza a predecir mejor



Machine Learning

Ejemplo de modelo - clasificador de spam

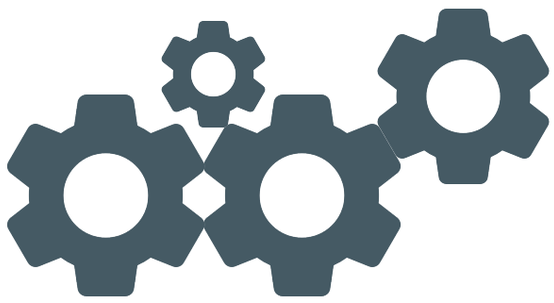
A medida que nuestro clasificador ve muchos ejemplos, comienza a predecir mejor



Machine Learning

Ejemplo de modelo - clasificador de spam

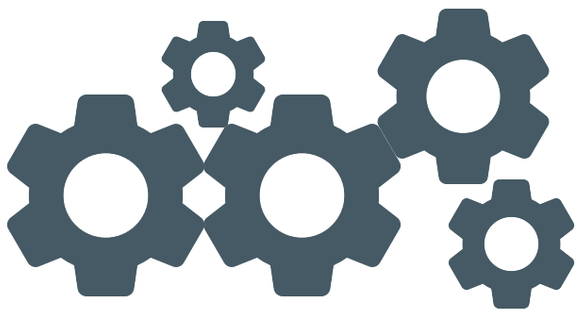
A medida que nuestro clasificador ve muchos ejemplos, comienza a predecir mejor



Machine Learning

Ejemplo de modelo - clasificador de spam

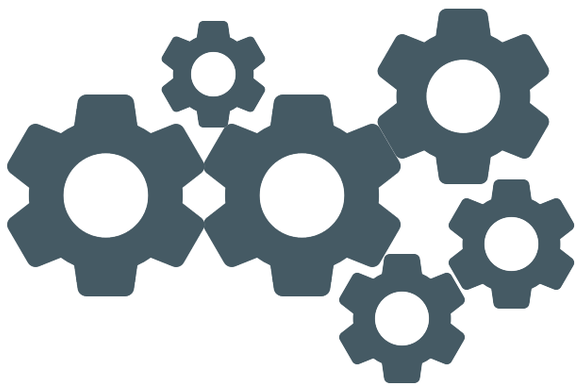
A medida que nuestro clasificador ve muchos ejemplos, comienza a predecir mejor



Machine Learning

Ejemplo de modelo - clasificador de spam

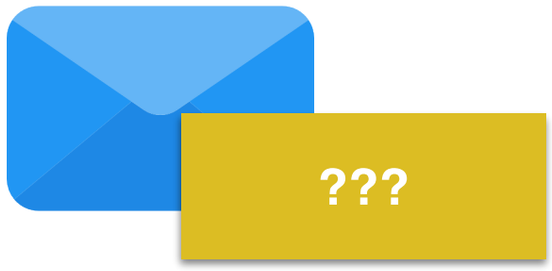
A medida que nuestro clasificador ve muchos ejemplos, comienza a predecir mejor

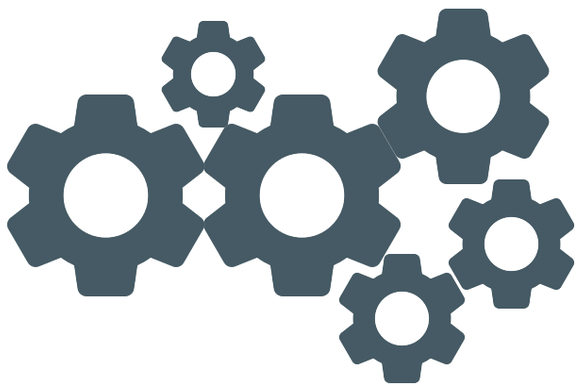


Machine Learning

Ejemplo de modelo - clasificador de spam

Y ahora cuando ve un correo desconocido:

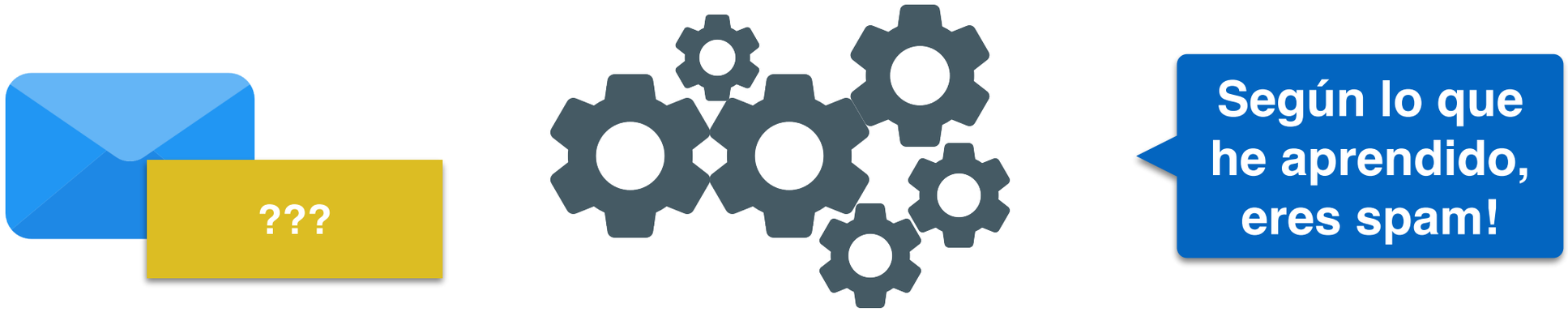




Machine Learning

Ejemplo de modelo - clasificador de spam

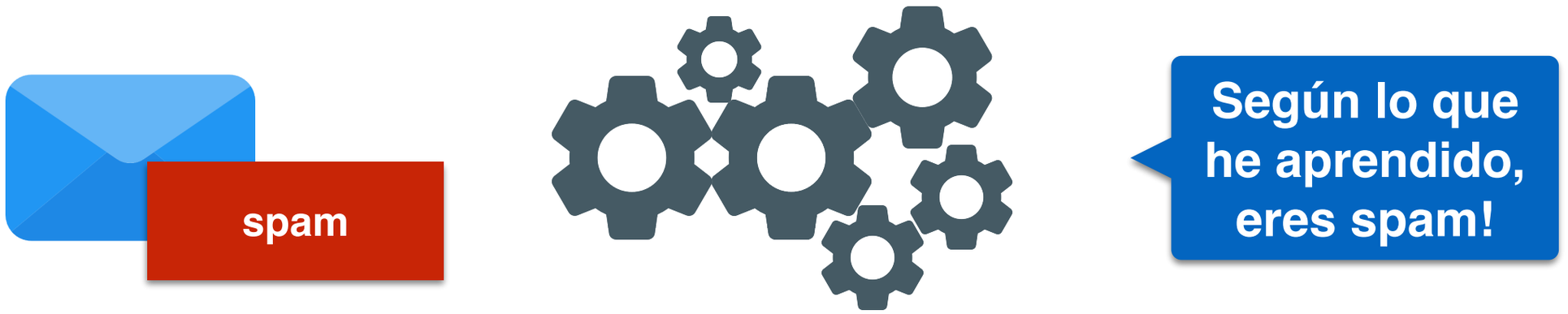
Y ahora cuando ve un correo desconocido:



Machine Learning

Ejemplo de modelo - clasificador de spam

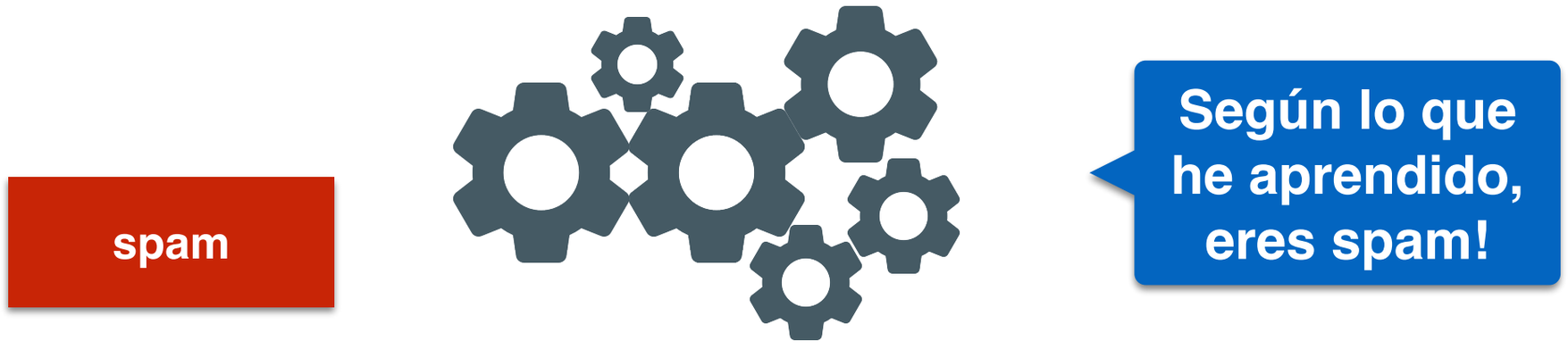
Y ahora cuando ve un correo desconocido:



Machine Learning

Ejemplo de modelo - clasificador de spam

Y ahora cuando ve un correo desconocido:



Machine Learning

Clasificación

En este caso estamos decidiendo si un correo que no hemos visto pertenece a alguna de estas dos clases:

- Clase 1: correo deseado
- Clase 2: correo no deseado

Pero también podemos hacer más cosas!

Machine Learning

Algunas tareas

- Clasificación binaria: clasificamos entre dos clases

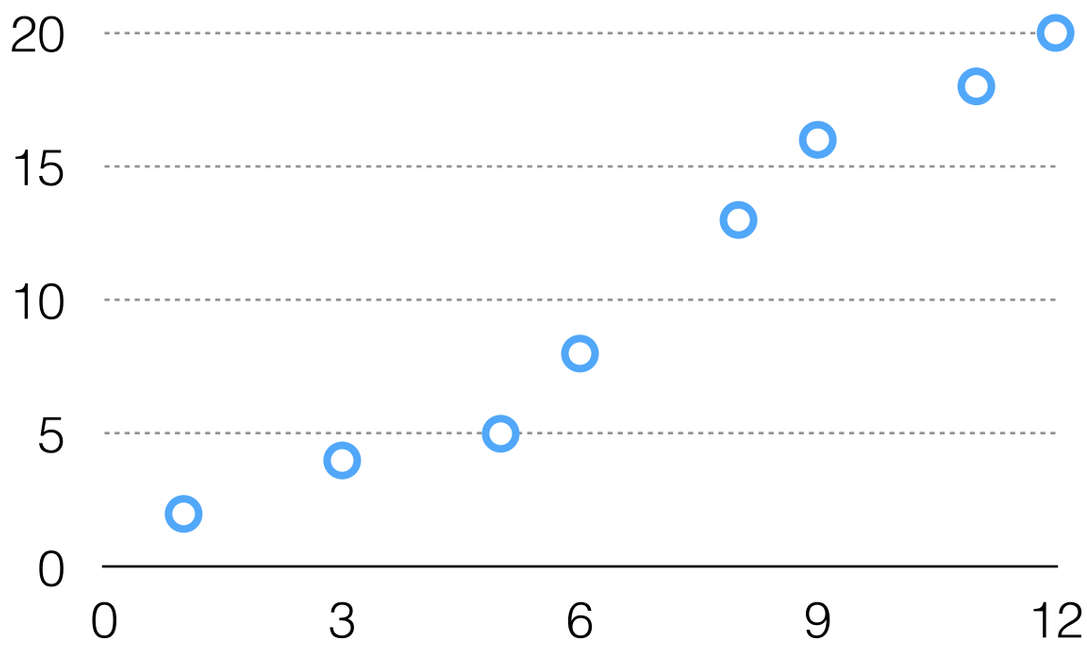
- Clasificación multiclase: clasificamos entre varias clases

- Regresión: buscamos un valor numérico

Machine Learning

Ejemplo - regresión lineal

Tenemos datos de los valores de viviendas en base a los m^2 de las mismas



Machine Learning

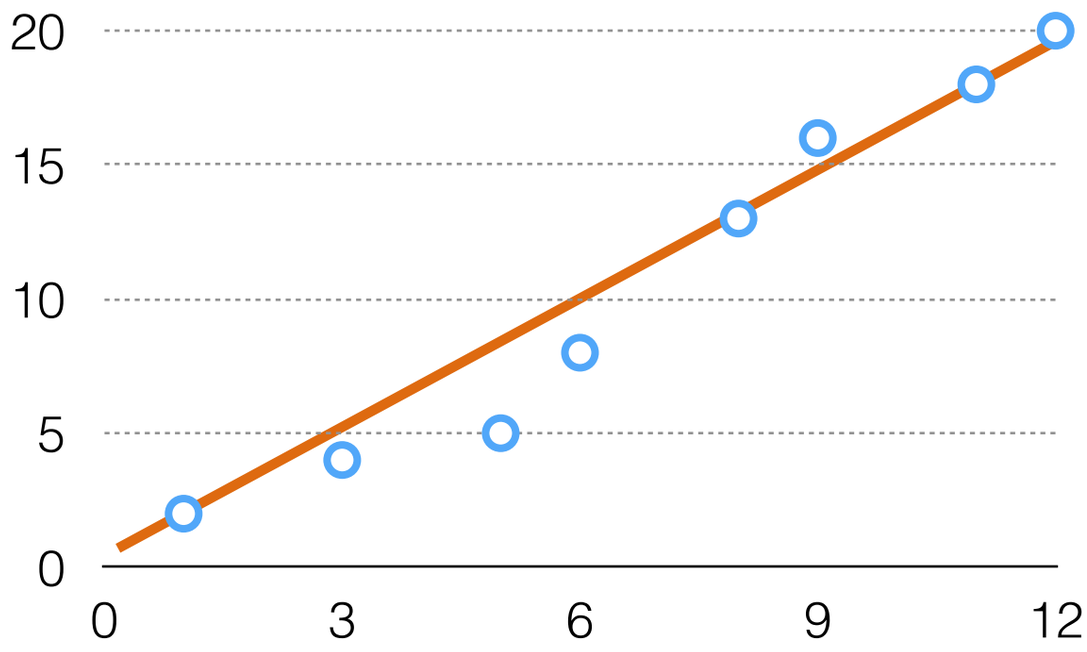
Ejemplo - regresión lineal

Si alguien nos entrega los m^2 de una vivienda que no conocemos, ¿cómo podemos calcular su valor?

Machine Learning

Ejemplo - regresión lineal

Si alguien nos entrega los m^2 de una vivienda que no conocemos, ¿cómo podemos calcular su valor?



Machine Learning

Modelos

Ahora estamos explicando esto como una caja negra: mostramos ejemplos y el modelo aprende

¿Qué tanto se aleja esto de la realidad?

Frameworks de ML

Scikit Learn

```
# Create linear regression object
regr = linear_model.LinearRegression()

# Train the model using the training sets
regr.fit(X_train, y_train)

# Make predictions using the testing set
y_pred = regr.predict(X_test)
```

Frameworks de ML

Existe una amplia variedad de *frameworks* que tienen muchas soluciones implementadas:

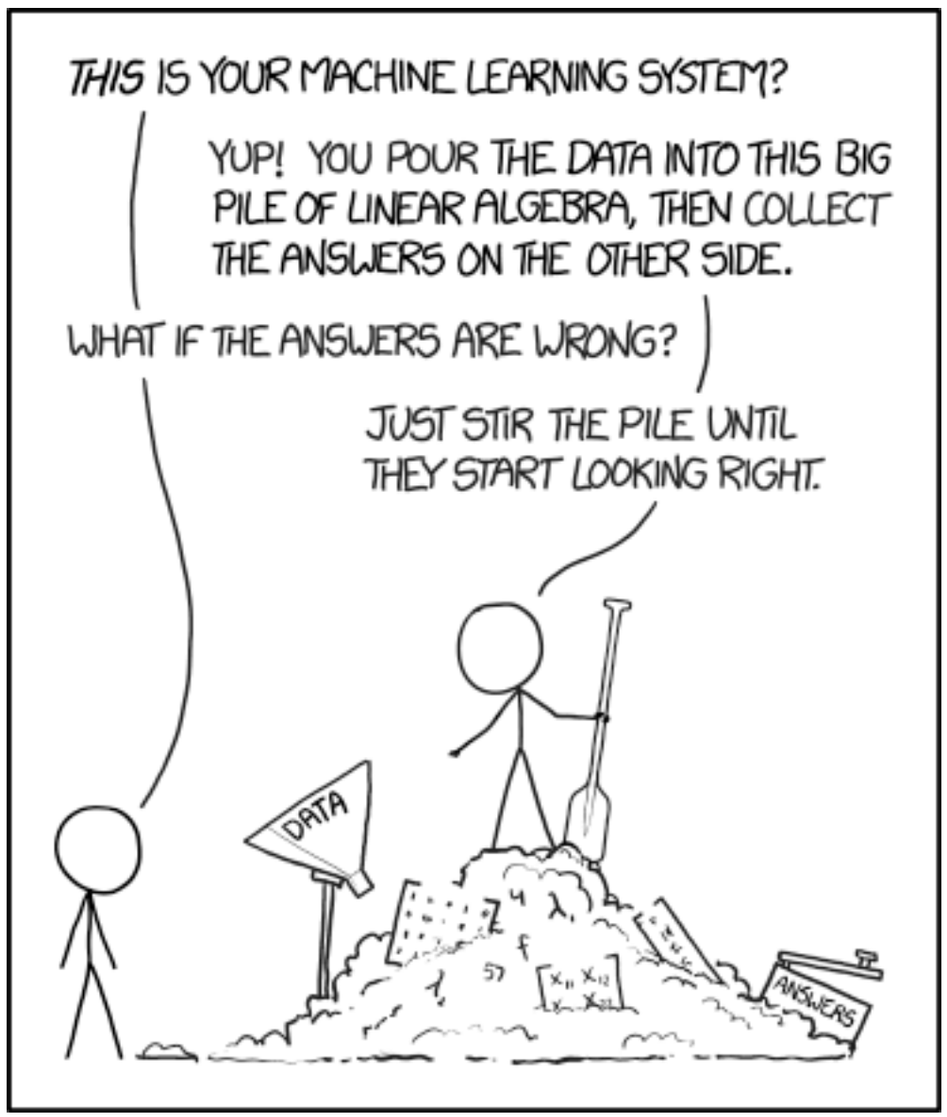
- Scikit Learn
- Tensorflow
- Keras
- Pytorch
- ...

Machine Learning

¿Se entiende mejor?

Machine Learning

¿Se entiende mejor?

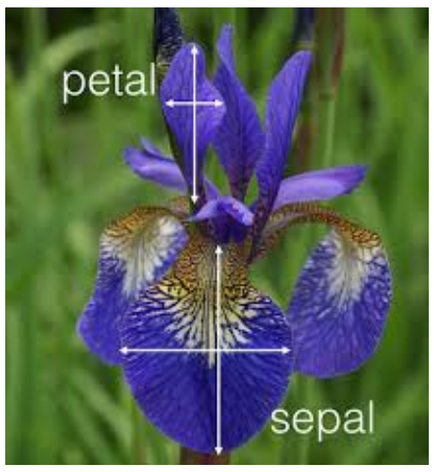


¿Cómo aprende un programa?

Regresión lineal

Para entender cómo aprende un programa, vamos a partir con un ejemplo sencillo: hacer una regresión lineal

Vamos a partir por el *hello world* en el área de *Machine Learning*: el *dataset* de las flores Iris



Regresión lineal

Este dataset se ve como una tabla con las siguientes columnas:

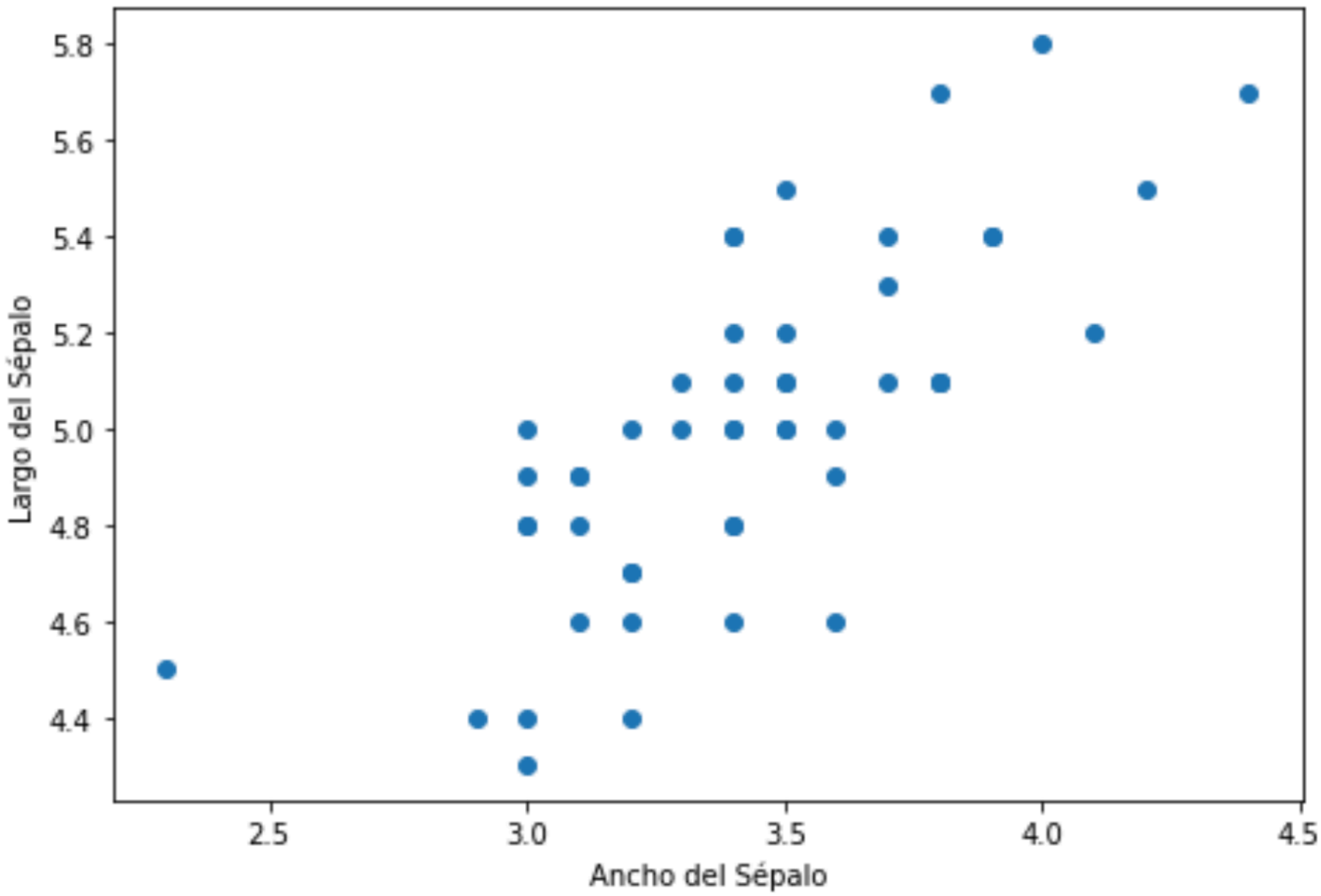
Largo del sépalo	Ancho del sépalo	Largo del pétalo	Ancho del pétalo	Tipo de flor
5.1	3.5	1.4	0.2	Iris Setosa
4.9	3.0	1.4	0.2	Iris Setosa
4.7	3.2	1.3	0.2	Iris Setosa
...	...	...	...	...

Regresión lineal

En este ejemplo vamos a hacer un predictor de los valores del **largo del sépalo** en función del ancho del sépalo para las flores de tipo **Setosa**

Un buen punto de partida es graficar en un plano las flores Iris Setosa, dejando en el eje x el ancho del sépalo y en el eje y el largo del sépalo

Regresión lineal

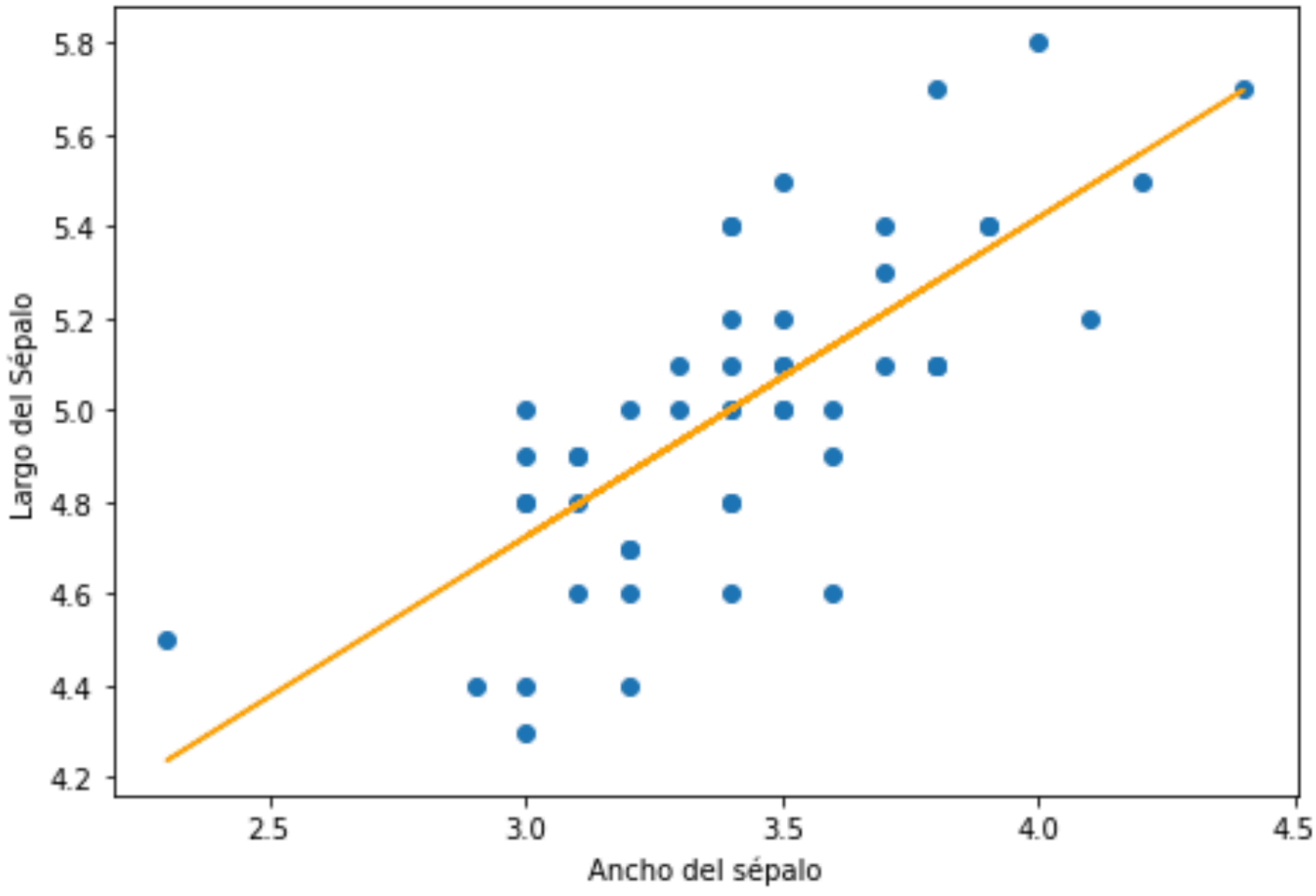


Regresión lineal

Ahora supongamos que salimos a pasear y encontramos una flor setosa cuyo ancho del sépalo es 5 cm

¿Cómo podemos predecir su largo?

Regresión lineal



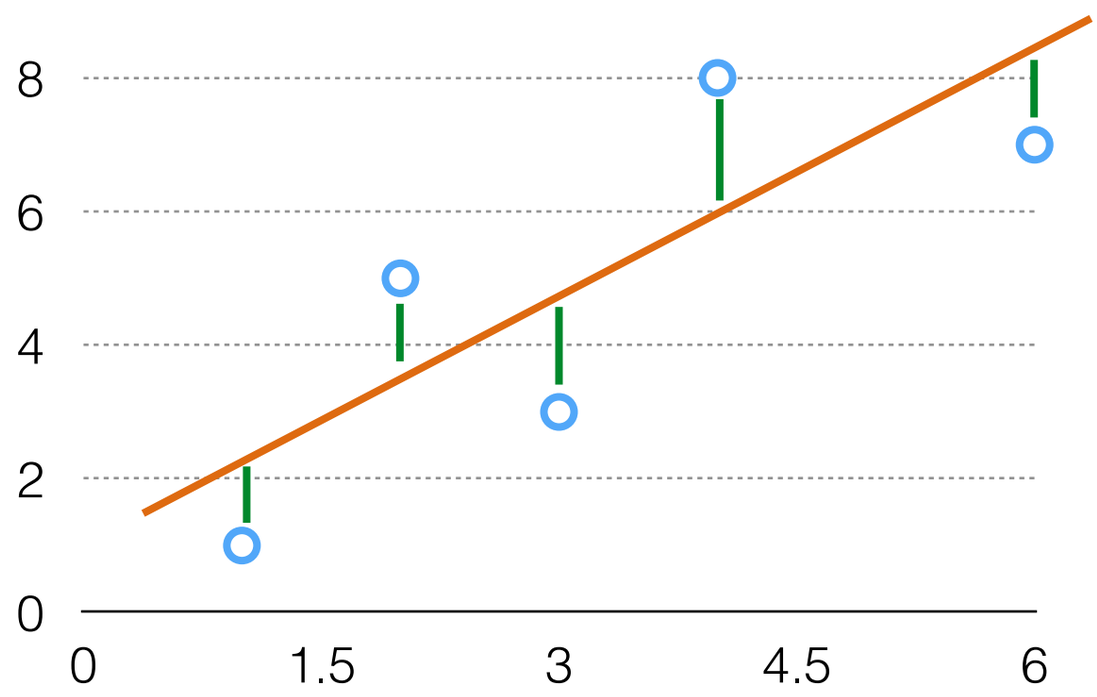
Regresión lineal

Podemos ajustar una regresión lineal, y ver el valor estimado para cuando el ancho del sépalo vale 5

Lo que estamos haciendo es descubrir los valores de $\hat{m}$ y $\hat{n}$ de forma tal que la recta $\hat{y} = \hat{m}x + \hat{n}$ pasa lo más cerca posible de todos los puntos del *dataset*

Regresión lineal

Para lograr esto, buscamos la recta que minimiza los errores



Regresión lineal

Matemáticamente, queremos resolver el siguiente problema de optimización:

$$\min \sum_{i=1}^{n} |y_i - \hat{y}_i| = \min \sum_{i=1}^{n} |y_i - (\hat{m}x_i + \hat{n})|$$

Donde nuestro dataset tiene n filas (no confundir con $\hat{n}$ que es el coeficiente de posición de la recta que queremos aprender)

Además y_i es el valor observado para la fila i, mientras que $\hat{y}_i$ es el valor que vamos a predecir para esa fila

Regresión lineal

¿Cómo resolvemos este problema de optimización?

Regresión lineal

Lo primero es cambiar el problema de optimización

En general no nos gusta el valor absoluto porque no es derivable en el origen, así que en vez de eso vamos a minimizar el error al cuadrado

$$\min \sum_{i=1}^{n} |y_i - (\hat{m}x_i + \hat{n})| \rightarrow \min \sum_{i=1}^{n} (y_i - (\hat{m}x_i + \hat{n}))^2$$

Regresión lineal

Para resolver el problema tenemos muchas opciones, por ejemplo, derivar en función de $\hat{m}$ y $\hat{n}$, igualar a 0 y resolver el sistema de ecuaciones; con eso obtenemos la fórmula:

$$\hat{m} = \frac{\sum_{i=1}^{n} (x_i y_i - \bar{y} x_i)}{\sum_{i=1}^{n} (x_i^2 - \bar{x} x_i)} \qquad \hat{n} = \bar{y} - \hat{m}\bar{x}$$

Donde $\bar{x}$ y $\bar{y}$ es el promedio de todos los x e y respectivamente

Regresión lineal

En resumen, lo que hicimos fue **aprender** los valores de $\hat{m}$ y $\hat{n}$ en función de nuestros datos

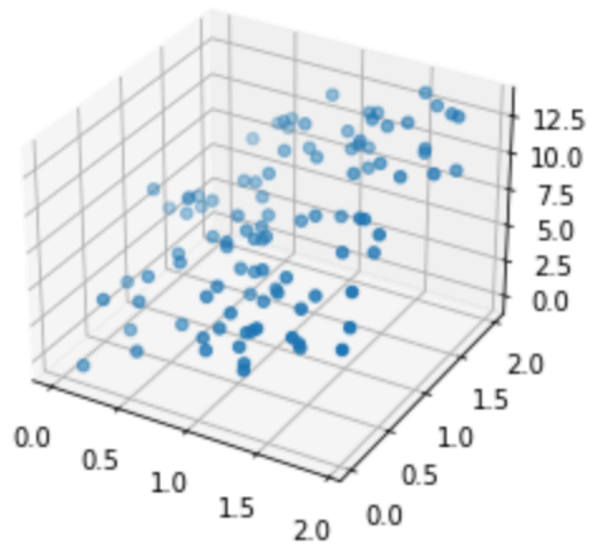
Esta idea es fundamental en el área de *Machine Learning*

Regresión lineal

¿Qué pasa si queremos predecir en función de más variables? Por ejemplo en este caso podría ser los valores del sépalo junto con el ancho del pétalo

Regresión lineal

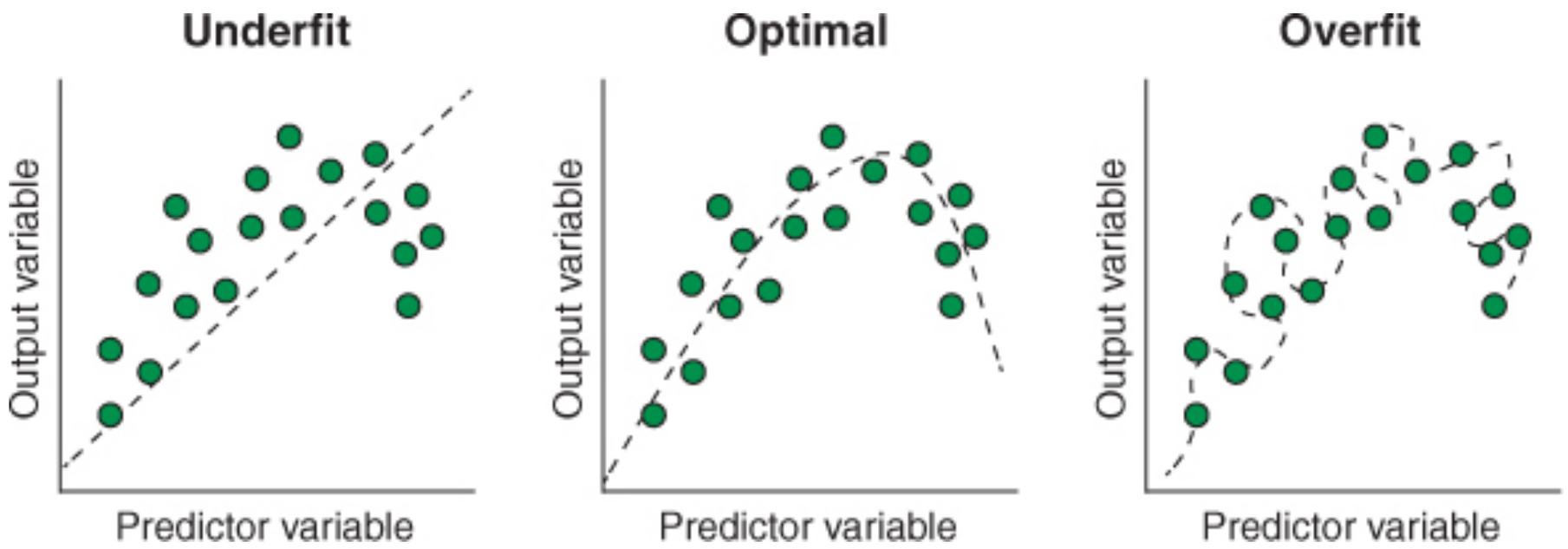
El modelo de regresión lineal se puede extender para más variables; por ejemplo, si queremos obtener un valor z en función de valores x e y, podemos generar un plano que pasa en 3 dimensiones



Regresión polinomial

Además, podemos generar modelos para ajustar polinomios en vez de rectas

Si bien el modelo se ajustará más a los datos, existe el riesgo de generar modelos sobre-ajustados (*overfitted*)


Underfit
Optimal
Overfit
Output variable
Predictor variable
Output variable
Predictor variable
Output variable
Predictor variable

Clasificación

Ahora, ¿qué pasa si dado los datos de una flor queremos predecir su tipo?

En este *dataset* los tipos son Iris Setosa, Versicolor y Virginica

Primera idea. Usar un modelo de regresión y asignar para cierto rango de valores un tipo

Clasificación

El problema es que para clasificar, nos gustaría tener un valor acotado, por ejemplo, una probabilidad de que una instancia pertenezca a una clase

Posible solución. El resultado de la regresión puede pasar por la función $\sigma : \mathbb{R} \rightarrow [0,1]$

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

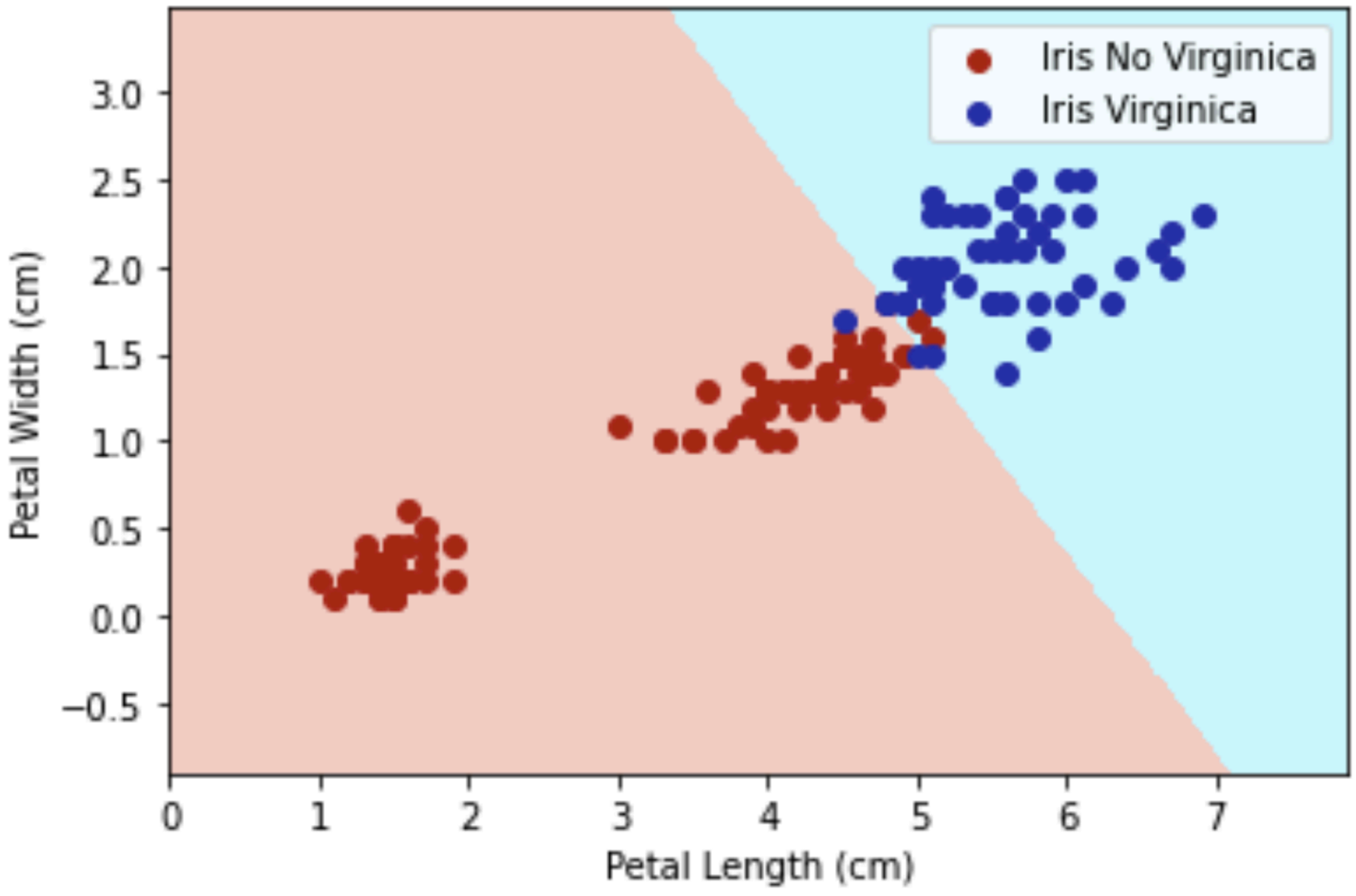
Regresión logística

Modelo utilizado para clasificación binaria, en que una instancia pasa por una regresión lineal, y ese resultado por una sigmoide

Es un modelo de clasificación lineal, en el que **vamos a aprender los coeficientes de una recta que divide el plano**

Para encontrar los coeficientes, resolvemos un problema de optimización utilizando *Gradient Descent*

Regresión logística



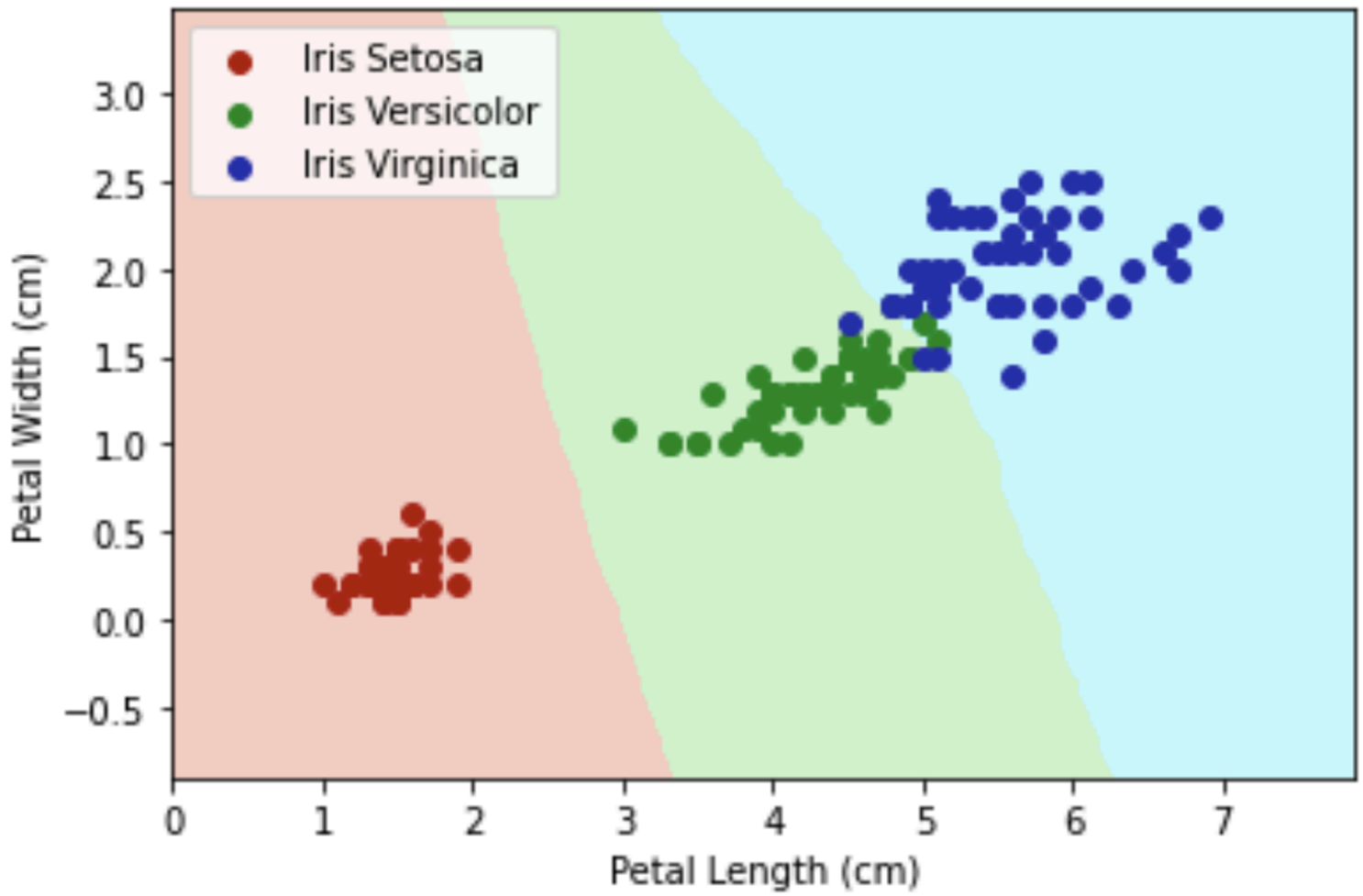
Clasificación

En general, los distintos modelos de *Machine Learning* van a generar distintas **fronteras de decisión** (i.e. formas de separar el espacio)

Esto se logra ya que cada modelo resuelve un problema de optimización distinto

KNN

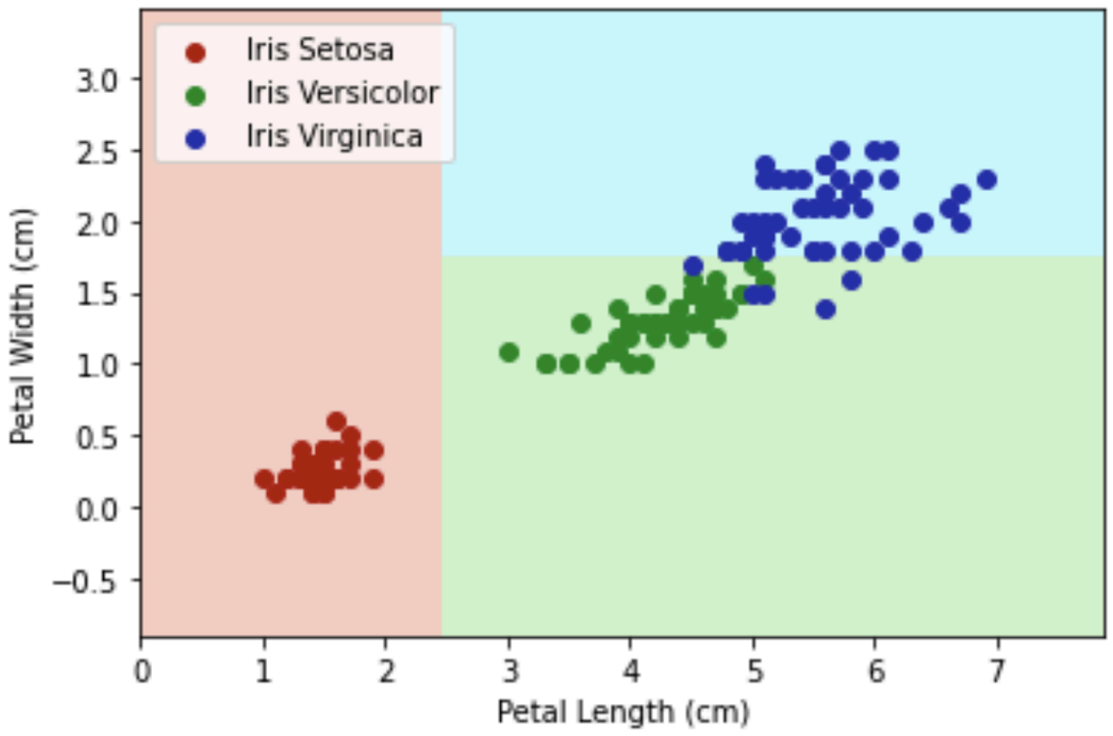
KNN es un modelo basado en distancias



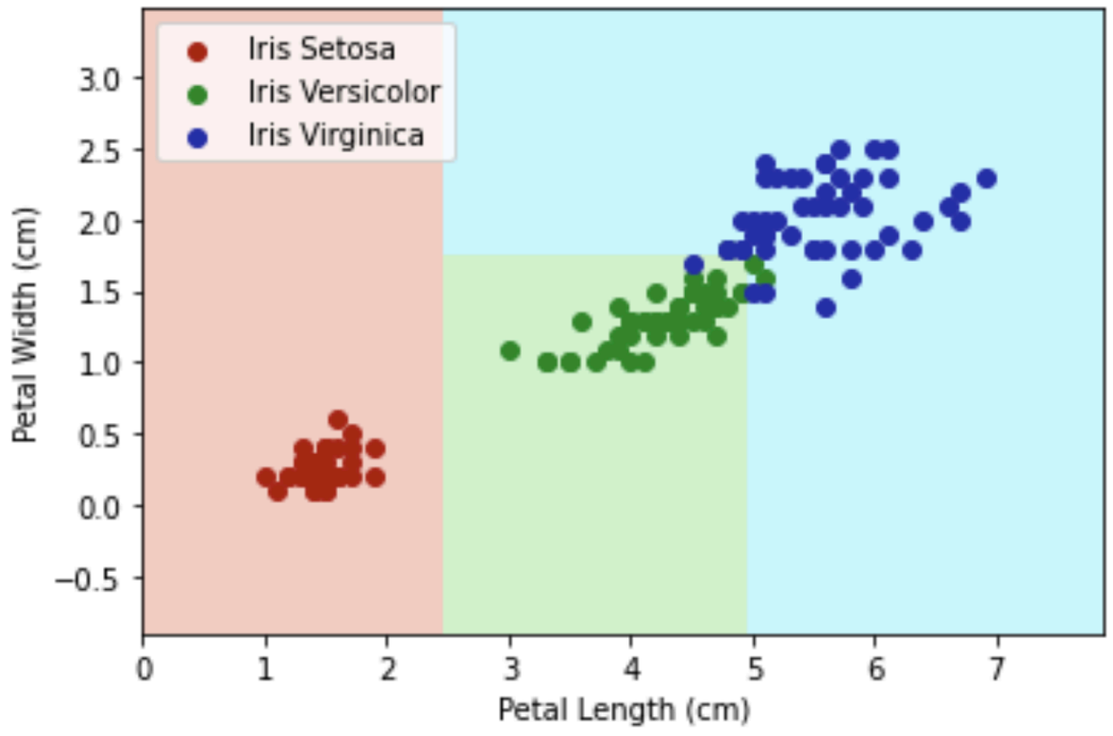
Decision Trees

Un árbol de decisión clasifica una instancia de acuerdo a distintas reglas

Árbol de prof. 2



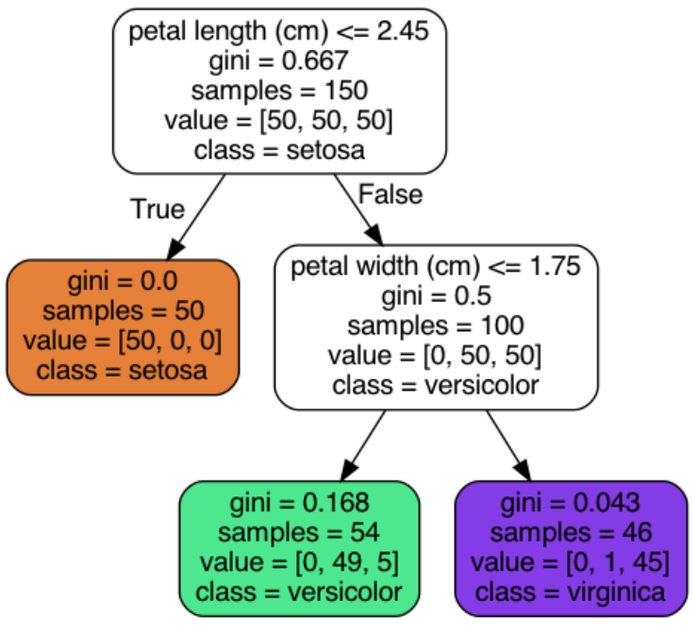
Árbol de prof. 3



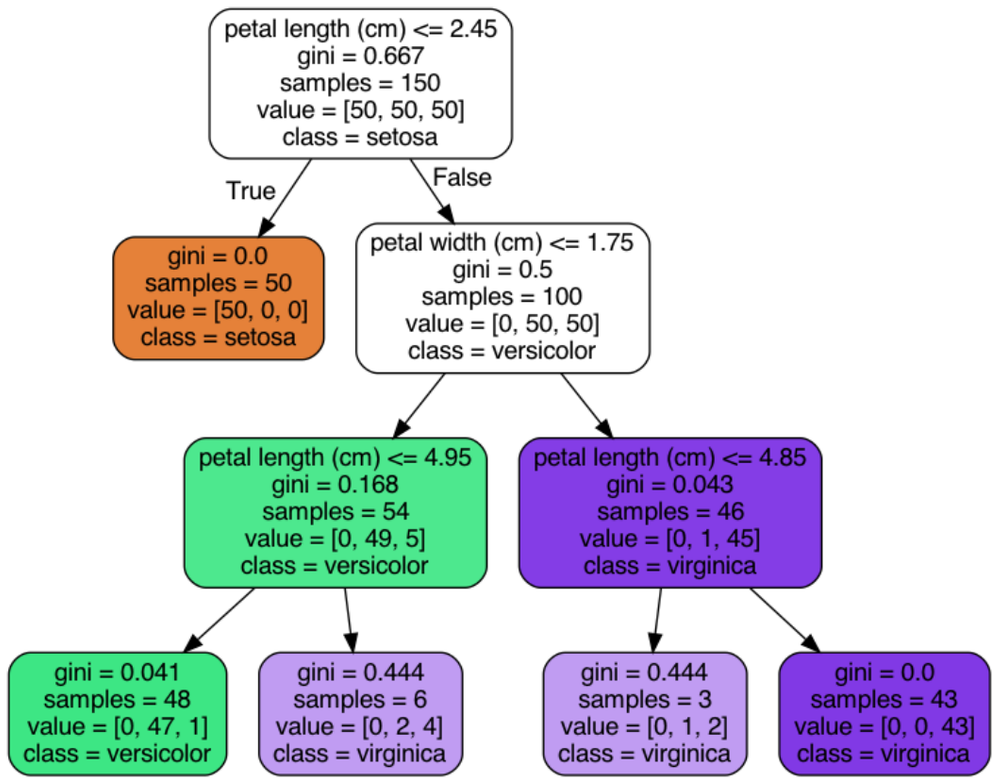
Decision Trees

Un árbol de decisión clasifica una instancia de acuerdo a distintas reglas

Árbol de prof. 2

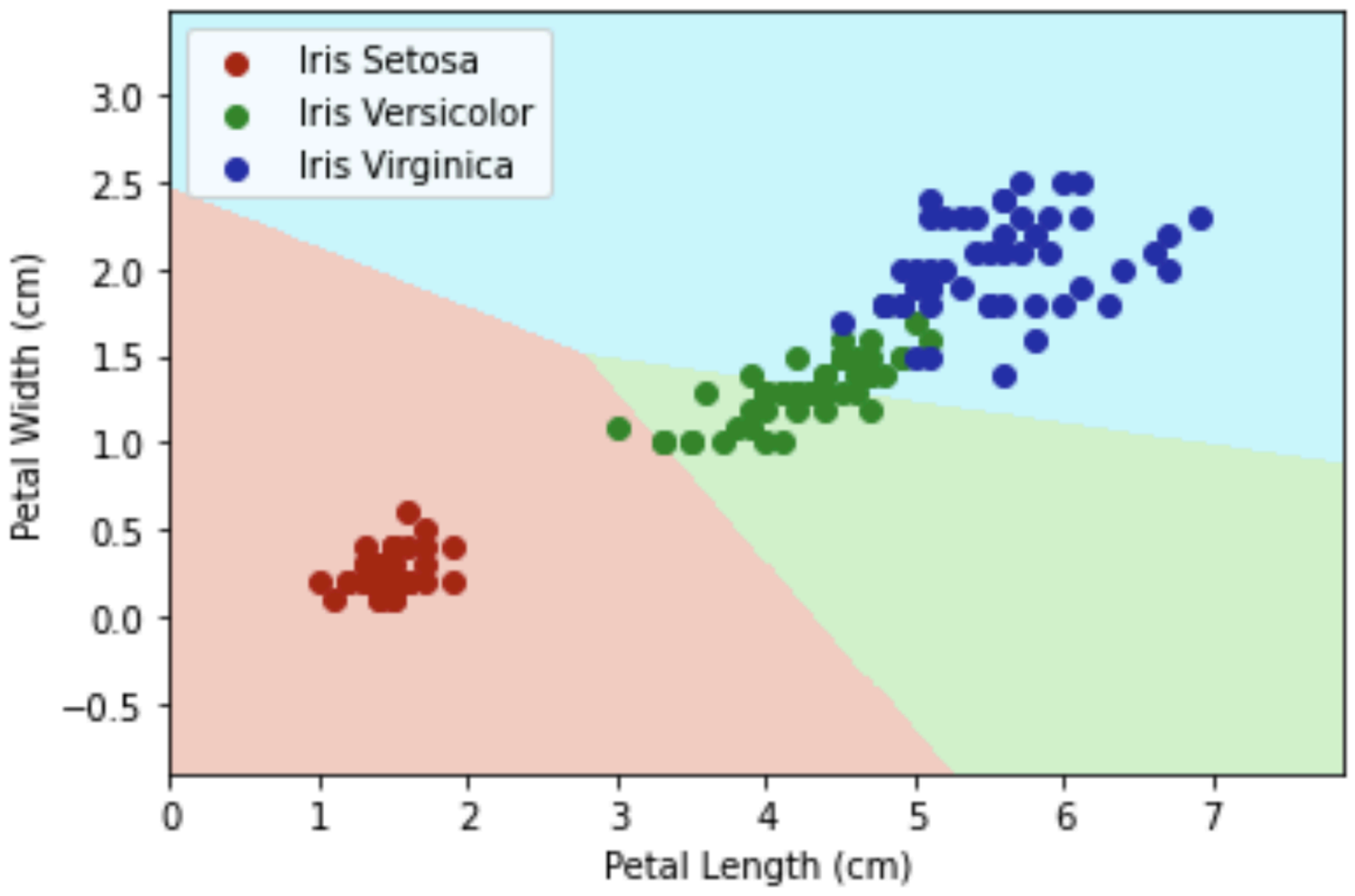


Árbol de prof. 3



SVM

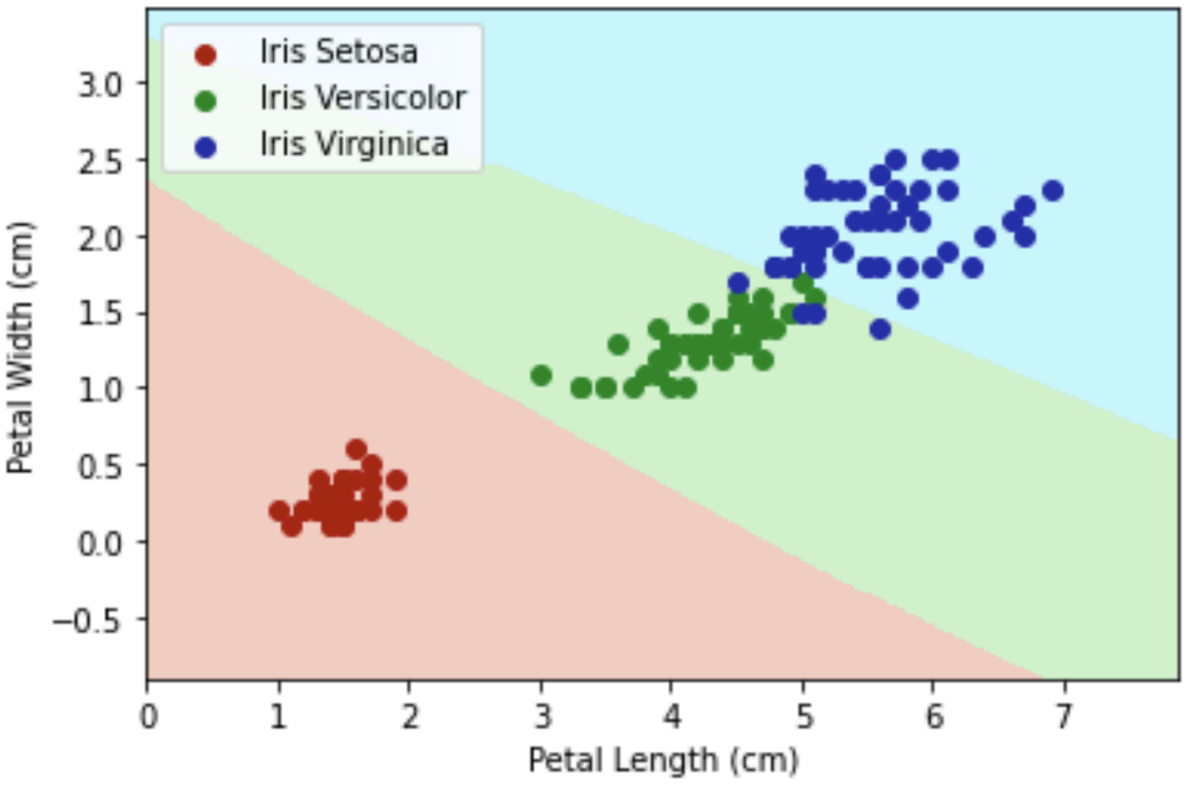
SVM busca la recta de mayor margen entre instancias



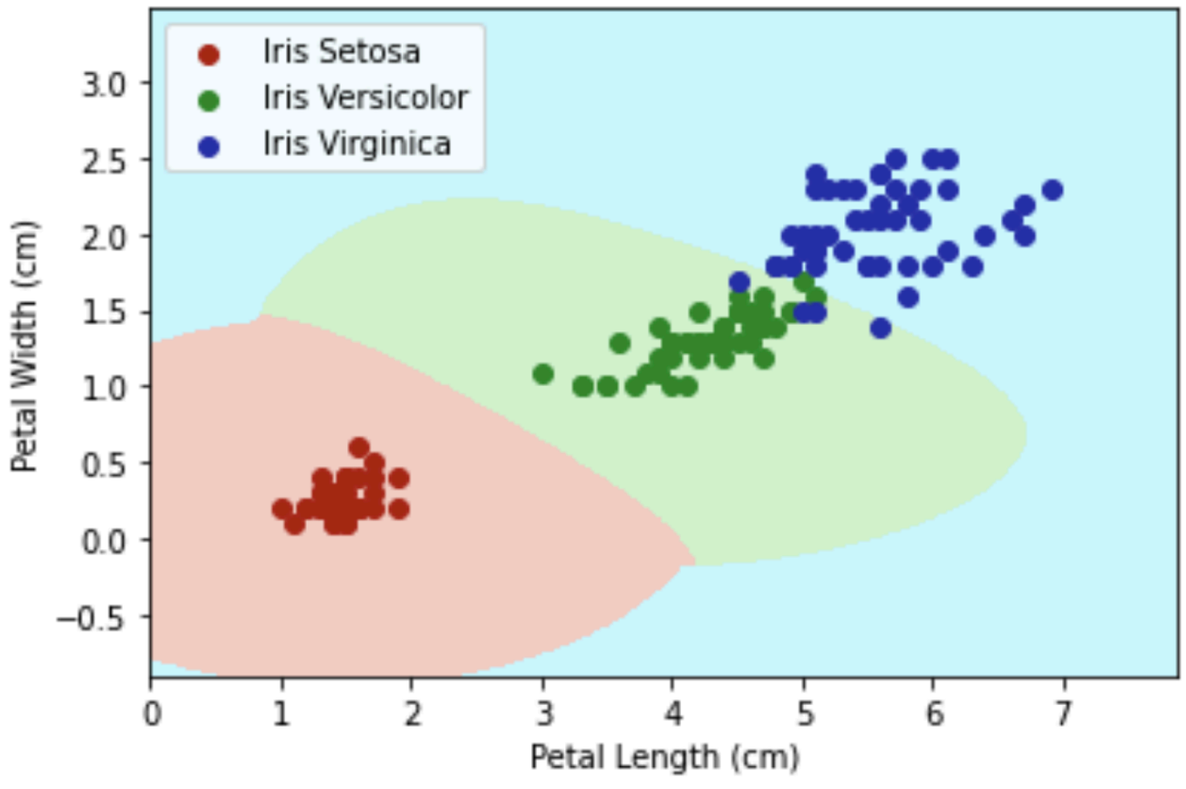
SVM

Pero SVM también puede hacer clasificación no lineal gracias al *truco del kernel*

Kernel RBF $\gamma = 0.1$



Kernel RBF $\gamma = 1$



Fundamentos de Ciencias de Datos

Semana 09 - Introducción a Machine Learning