SolidQ
Think Big. Move Fast.

Módulo 3: Entendiendo los algoritmos de Machine Learning

Primera Parte

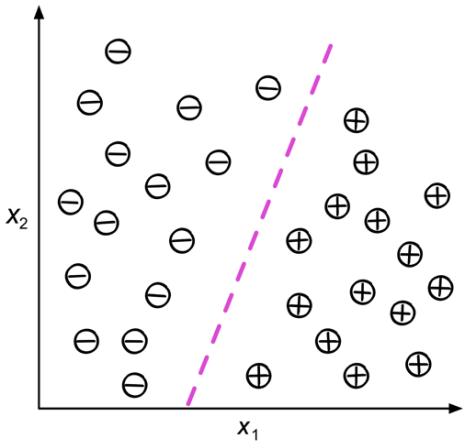
Agenda

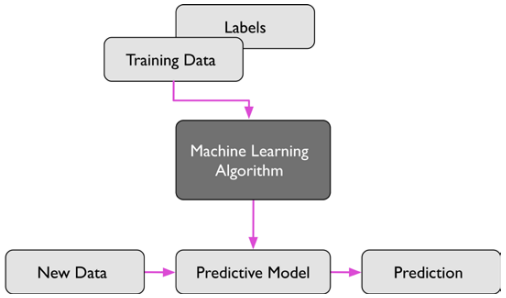
- Introducción a los algoritmos
- Datos de Entrenamiento y Prueba
- Clasificación
- Regresión
- Clustering
- Anomalías
- Recomendaciones
- Conjuntos de datos No-balanceados
- Como interpretar modelos

Introducción a Algoritmos

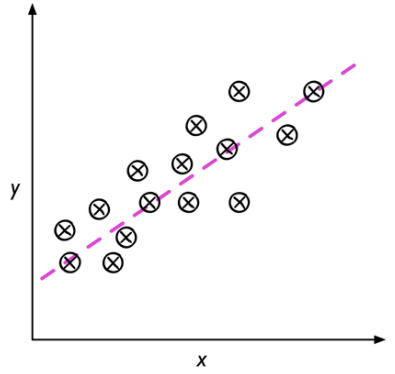
Haciendo predicciones con aprendizaje supervisado

- Clasificación para predecir etiquetas de clases
  - Binaria
  - Multiclase



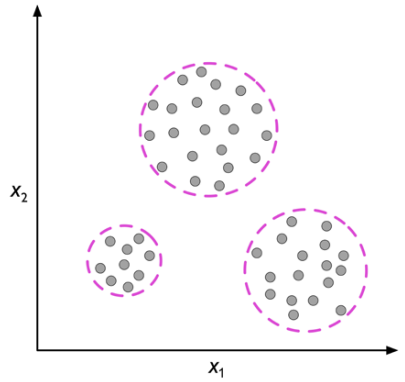


- Regresión para predecir salidas continuas



Descubriendo estructuras ocultas con aprendizaje no supervisado

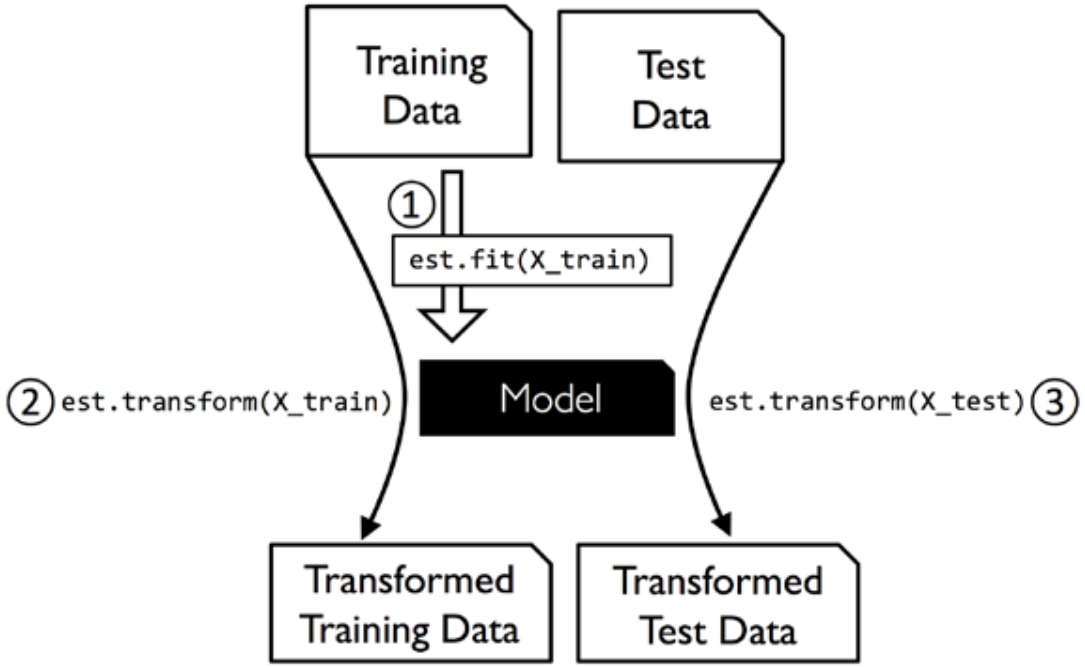
- Encontrando subgrupos con Clustering



- Reducción de la dimensionalidad para compresión de datos

Demo 03 – A: El primer algoritmo

- Scikit-learn



Datos de Entrenamiento y Prueba

¿Por qué?

- Debemos de dejar datos que algoritmo no conozca para poder evaluar su rendimiento
- Debemos de intentar no introducir sesgos a la hora de realizar esta división
- Tercer grupo: Datos de seguridad
  - Entreno el modelo con un conjunto de datos de entrenamiento
  - Pruebo con el conjunto de datos de prueba
  - Entreno el modelo con datos de entrenamiento + prueba
  - Compruebo rendimiento con Datos de seguridad

¿Cómo?

- Submódulo sklearn.model_selection
- train_test_Split
  - Validación de entrada
  - next(ShuffleSplit().split(X, y))

Demo 03- B Train y Test

Dividiendo nuestro conjunto de datos

Regresión

Regresión

- Para predecir resultados con valores reales:
  - ¿Cuántos clientes llegarán a nuestro sitio web la próxima semana?
  - ¿Cuántos televisores venderemos el año que viene?
  - ¿Podemos predecir los ingresos de alguien desde sus clics de navegación a través de la información?

Regresión

- Cada observación se representa como un conjunto de números

Representamos una persona : [5 3 120 12 1 0 …..]

[0 0 89 5 1 1 …..]

[1 0 20 0 0 1 …..]

:

Income

84

32

-10

:

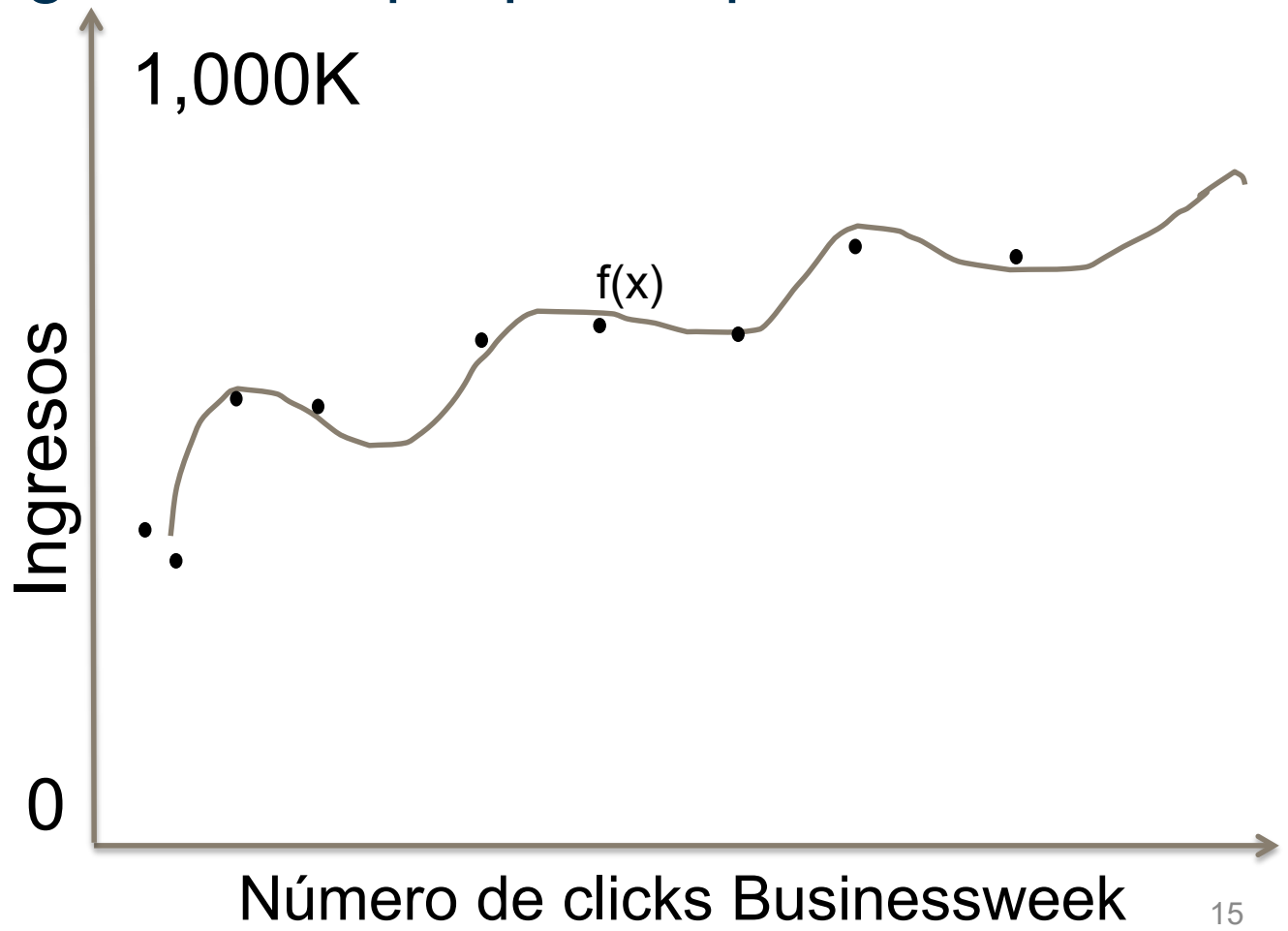
Características, llamadas X

Etiquetas, llamdas Y

Regresión

- Formalmente, dado un conjunto de entrenamiento (x_i, y_i) para i=1…n, queremos crear un modelo de regression f que pueda predecir la etiqueta y para un nuevo x

f(x) = function(Número de clics Businessweek)

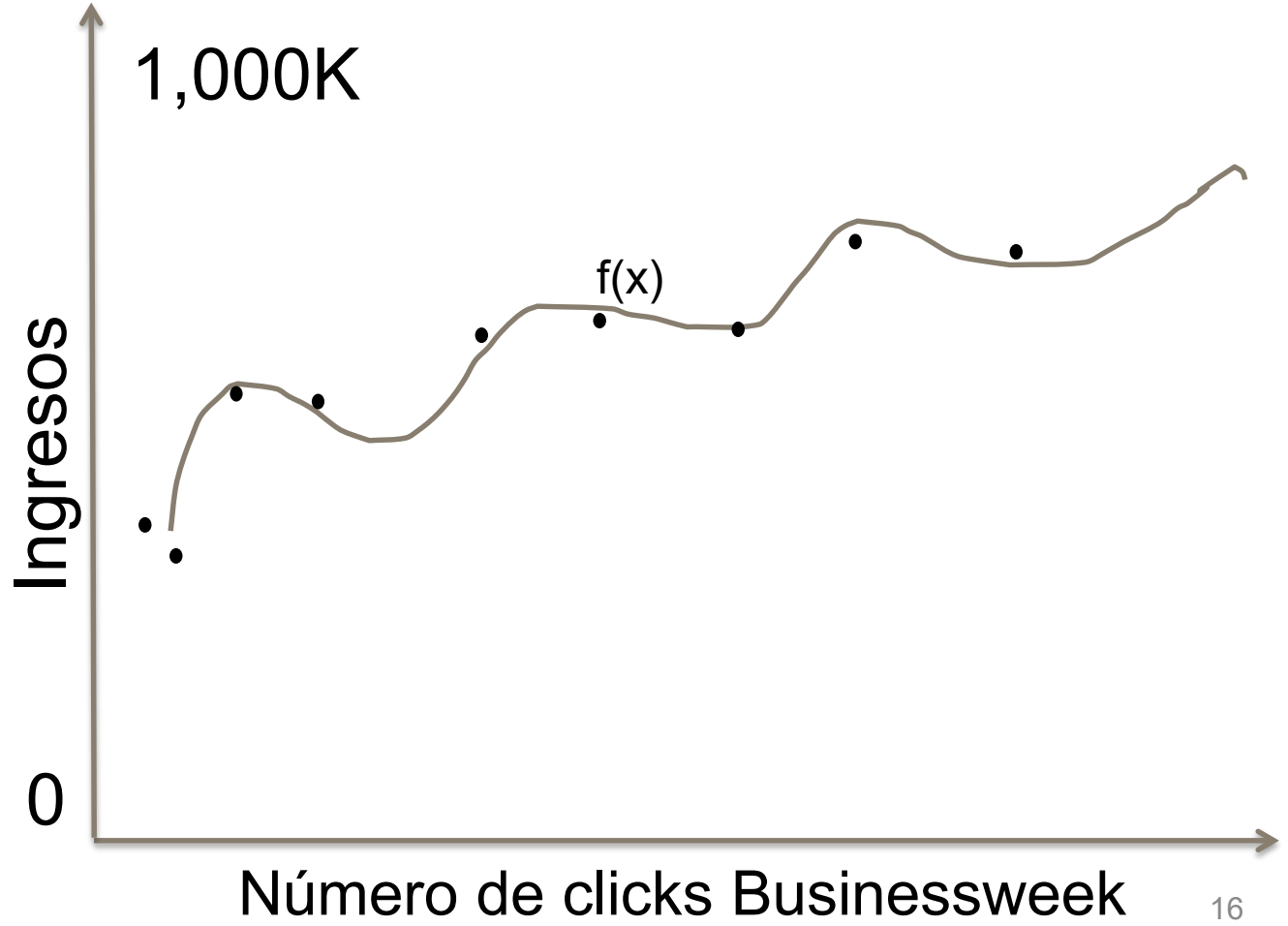


Regresión

- Formalmente, dado un conjunto de entrenamiento (x_i, y_i) para i=1…n, queremos crear un modelo de regression f que pueda predecir la etiqueta y para un nuevo x

f(x) = function(Número de clics Businessweek)

(Overfitting?)

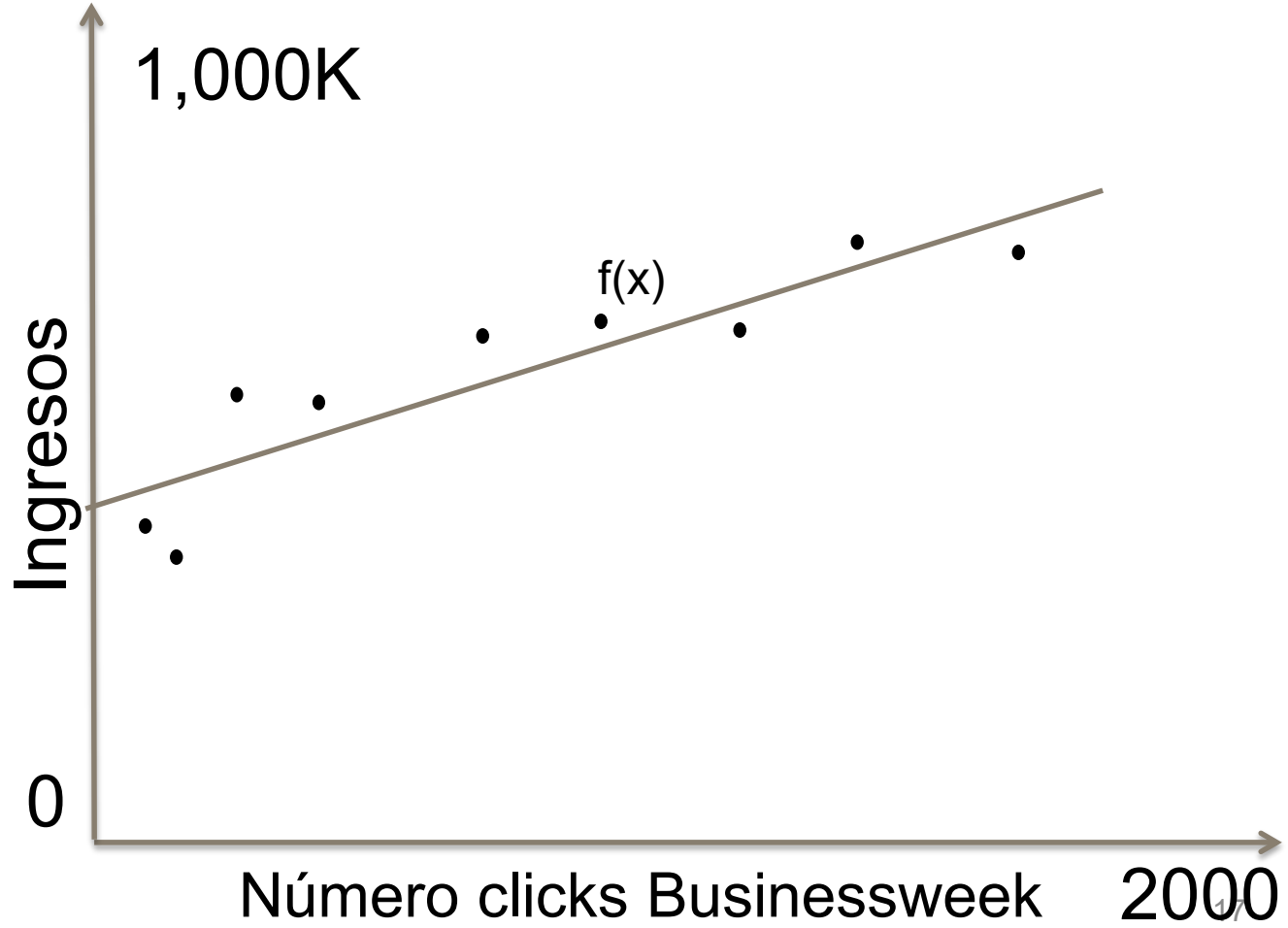


Regresión

- Formalmente, dado un conjunto de entrenamiento (x_i, y_i) para i=1…n, queremos crear un modelo de regression f que pueda predecir la etiqueta y para un nuevo x

f(x) = function(Número de clics Businessweek)
= 5K*Número clicks Businessweek + 100K

(Underfitting?)



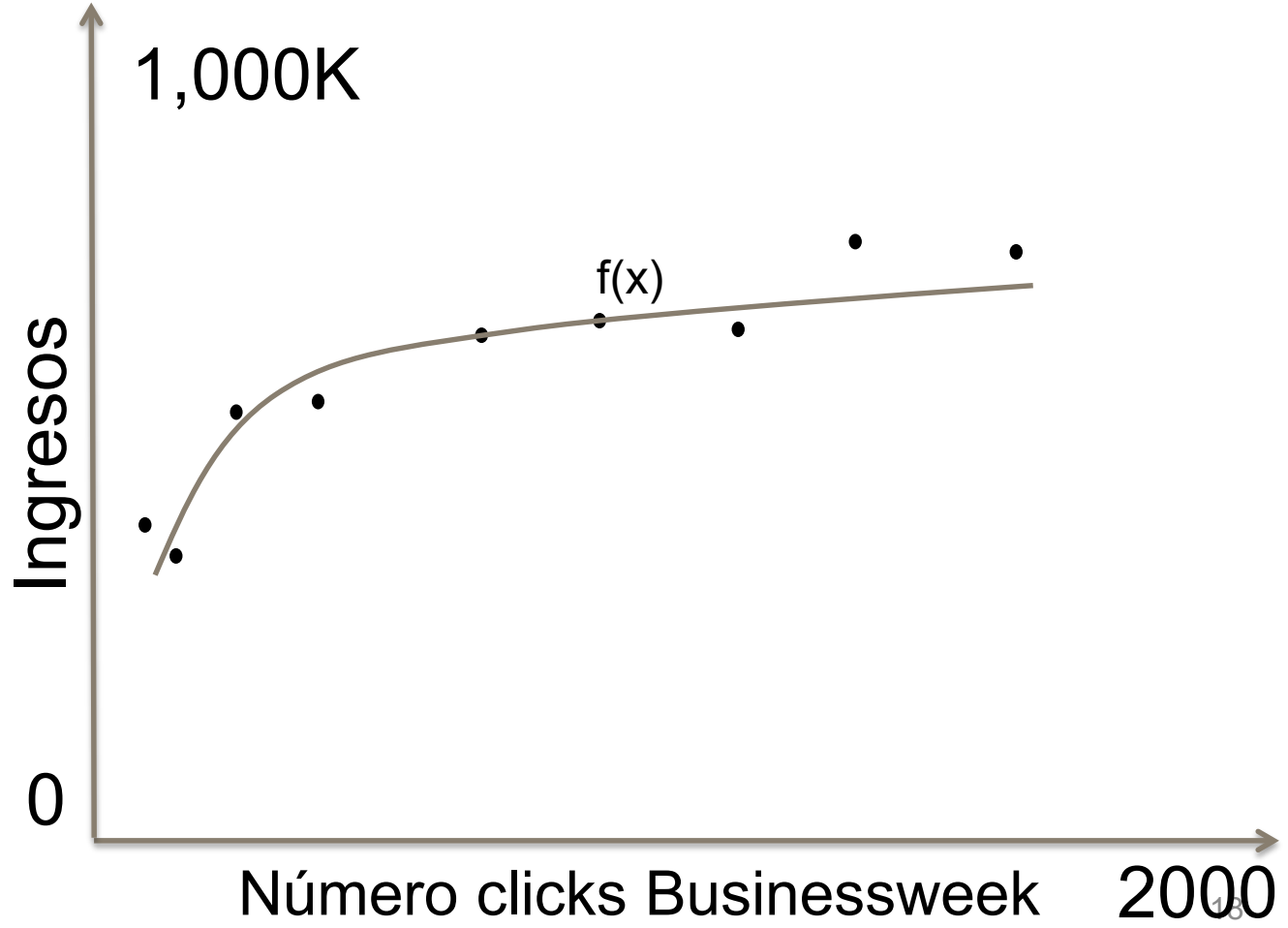
Regresión

- Formalmente, dado un conjunto de entrenamiento (x_i, y_i) para i=1…n, queremos crear un modelo de regression f que pueda predecir la etiqueta y para un nuevo x

f(x) = function(Number of Businessweek clicks)

Correcto?

Hablermos más tarde sobre ello



Regresión

- Formalmente, dado un conjunto de entrenamiento (x_i, y_i) para i=1…n, queremos crear un modelo de regression f que pueda predecir la etiqueta y para un nuevo x

Estimated income:

f(x) = function(Number of visits to upscale furniture websites, Number of Businessweek clicks, Number of distinct people emailed per day, Number of purchases of over 5K within the last month, Number of visits to airlines, etc.)

For instance,

f(x) = 3*Number of visits to upscale furniture websites
+10*Number of Businessweek clicks
+100*Number of distinct people emailed per day
+2*Number of purchases of over 5K within the last month
+10*Number of visits to airlines

But f(x) could be much more complicated

Aplicaciones de la Regresión

- Predecir cantidades monetarias
- Predecir el consumo o demanda de productos / energía

Aprendizaje Supervisado - Resumen

- "Supervisado" significa que los datos de entrenamiento tienen etiquetas de verdad para aprender de ellas. La clasificación y la regresión son problemas de aprendizaje supervisados.
- La clasificación (supervisada) a menudo tiene etiquetas + 1 o-1.
- La regresión (supervisada) tiene etiquetas numéricas.
- Los algoritmos de aprendizaje supervisado son mucho más fáciles de evaluar que los no supervisados.

www.solidq.com

info@solidq.com