

Datos no estructurados, evaluación & mejora de modelos

Taller comercio y desarrollo regional

Módulo Aprendizaje de máquinas

Curso - Taller

Hoy

1. Datos no estructurados

¿Qué significa no estructurado? ¿Qué tipo de preguntas se puede responder con estos datos?

Caso: minería de texto

2. Evaluación de modelos

¿Cómo escogemos entre diferentes máquinas?

3. Script de modelos de ML

¿Con qué paquetes corremos los modelos en R?

4. Taller diferenciado

Avance de informe vs script de mejoramiento de modelos



Photo by [The Coach Space](#) from [Pexels](#)

Datos no estructurados

Un comentario menor...



Photo by [Julian V](#) from [Pexels](#)

Datos no estructurados

- No todos los datos en el mundo vienen en filas y columnas...
- Es muy común encontrar

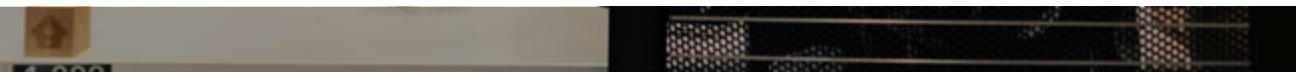


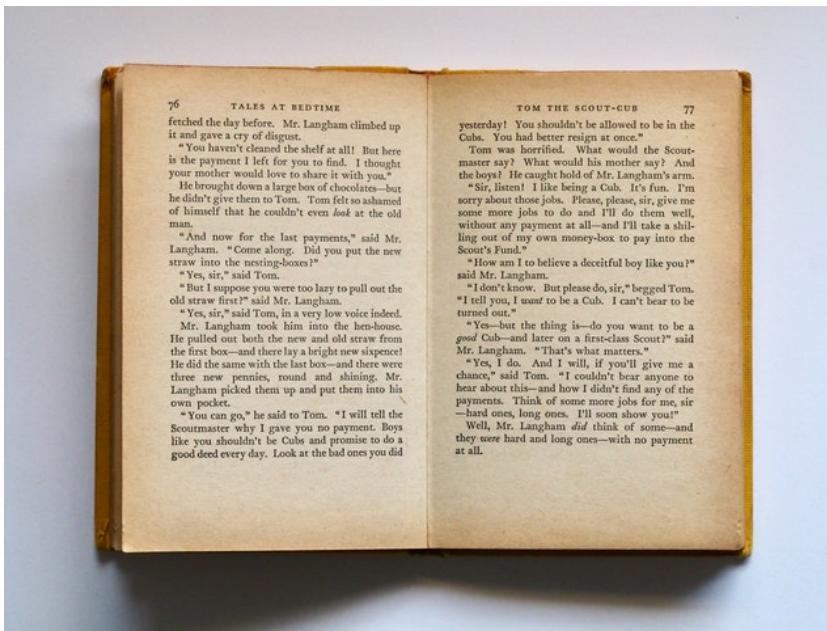
Photo by [Jess Bailey Designs](#) from [Pexels](#)

Photo by [Tirachard Kumtanom](#) from [Pexels](#)

Photo by [Solen Feyissa](#) from [Pexels](#)

Datos no estructurados

- No todos los datos en el mundo vienen en filas y columnas...
- Es muy común encontrar



Datos no estructurados

- No todos los datos en el mundo vienen en filas y columnas...
- Es muy común encontrar

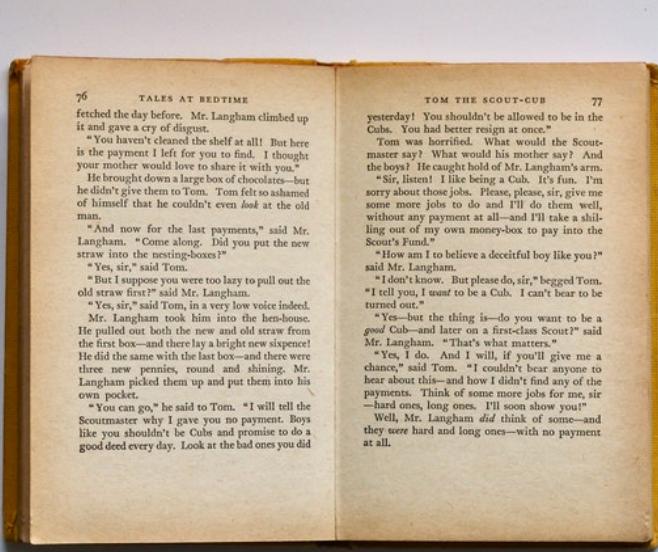


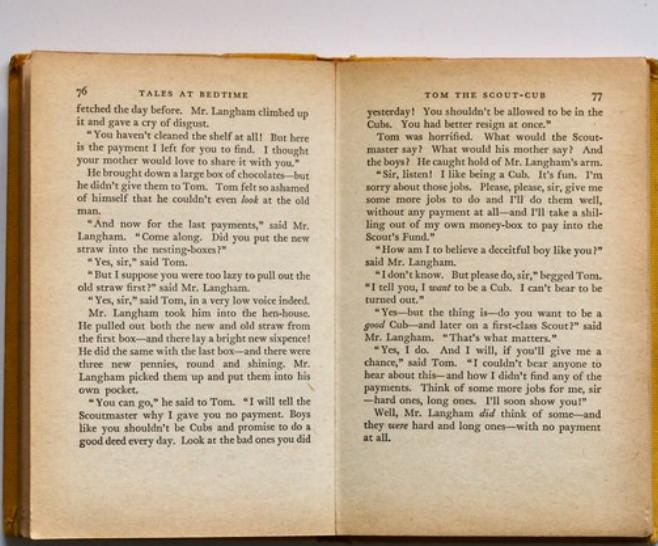
Photo by [Jess Bailey Designs](#) from [Pexels](#)

Photo by [Tirachard Kumtanom](#) from [Pexels](#)

Photo by [Solen Feyissa](#) from [Pexels](#)

Datos no estructurados

- No todos los datos en el mundo vienen en filas y columnas...
- Es muy común encontrar



Para esto

- La metodología es por etapas:
 1. Hacer un **preprocesamiento** de la información
 2. Ese preprocesamiento estructura los datos de alguna forma
 3. Desarrollo de análisis con datos estructurados

Algunos ejercicios incluyen:

- Minería de texto: estructuración de textos
 - Análisis de sentimiento
 - Análisis de tópicos
 - ...
- *Scraping*: recolección sistemática de datos ubicados de forma dispersa

Érase una vez en un lugar muy muy lejano, había una familia que reinaba sobre un terreno lleno de personas. Había suficientes manzanas en esa tierra como para alimentar a un ejército por dos meses.



Tipo	Frecuencia
Sustantivos	12
Adjetivos	10
Artículos	7

Algunos ejercicios incluyen:

- Minería de texto: estructuración de textos
 - Análisis de sentimiento
 - Análisis de tópicos
 - ...
- *Scraping*: recolección sistemática de datos ubicados de forma dispersa

Érase una vez en un lugar muy muy lejano, había una familia que reinaba sobre un terreno lleno de personas. Había suficientes manzanas en esa tierra como para alimentar a un ejército por dos meses.



Tipo	Palabra 1	Palabra 2...
Frase 1	1	0
....	0	1
Frase 234	0	2

En fin...

- A partir de los datos pre-procesados se pueden correr análisis estadísticos.
- Las formas en que se pueden aplicar estos datos son muy amplias, para otros cursos.

Evaluación de modelos

Escogiendo entre máquinas
En aprendizaje supervisado



Photo by [Fernando Arcos](#) from [Pexels](#)

Es importante poder comparar

- Parte del ejercicio del ML es matemática exacta
- ... y parte es actitud de juego para probar muchas variaciones.

Es importante poder comparar

- Parte del ejercicio del ML es matemática exacta
- ... y parte es actitud de juego para probar muchas variaciones.

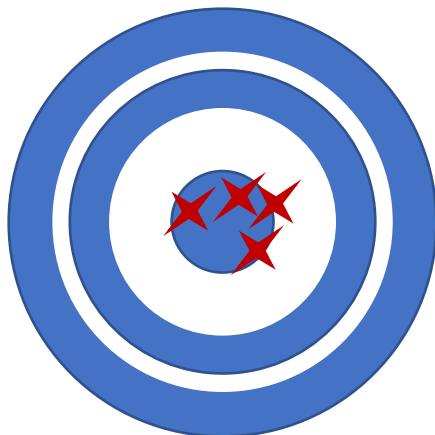
¿Cómo escoger?

Dos consideraciones

- Sesgo: qué tanto se equivoca al predecir nuestra máquina.
- Varianza: problema de generalización - ¿se descontrola con datos desconocidos?

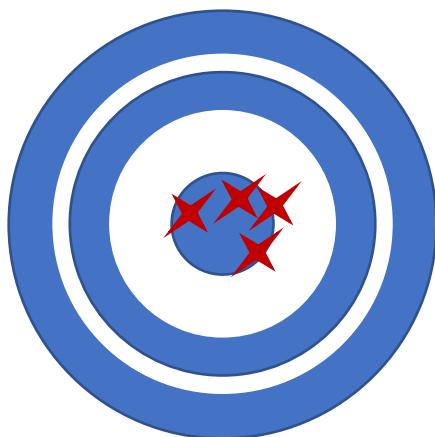
Dos consideraciones

- Sesgo: qué tanto se equivoca al predecir nuestra máquina.
- Varianza: problema de generalización - ¿se descontrola con datos desconocidos?



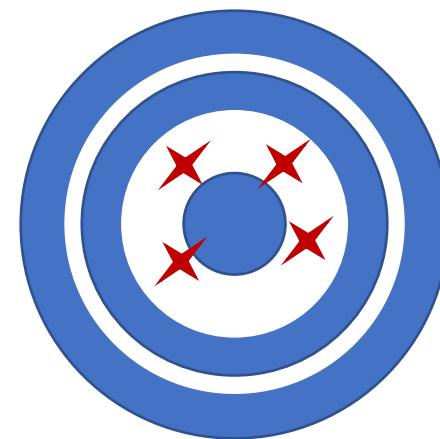
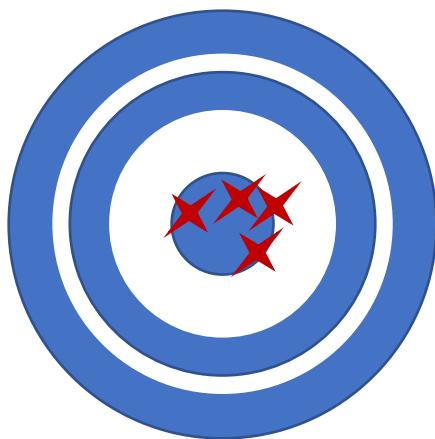
Dos consideraciones

- Sesgo: qué tanto se equivoca al predecir nuestra máquina.
- Varianza: problema de generalización - ¿se descontrola con datos desconocidos?



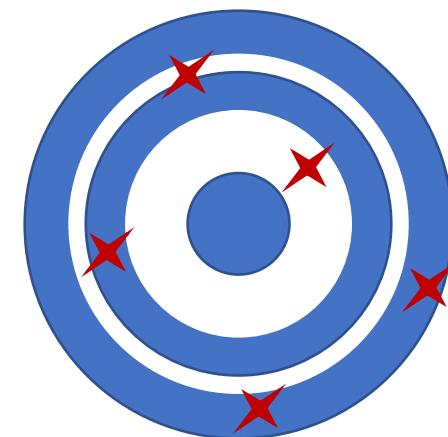
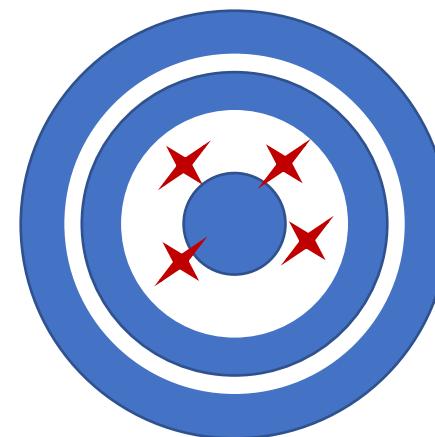
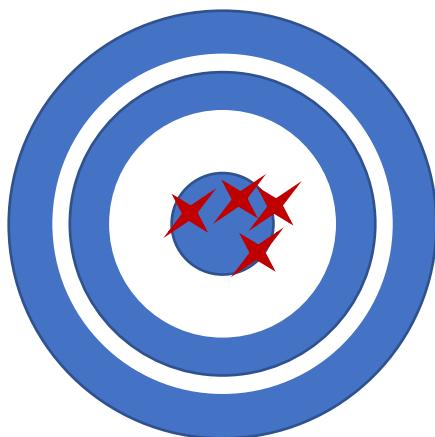
Dos consideraciones

- Sesgo: qué tanto se equivoca al predecir nuestra máquina.
- Varianza: problema de generalización - ¿se descontrola con datos desconocidos?



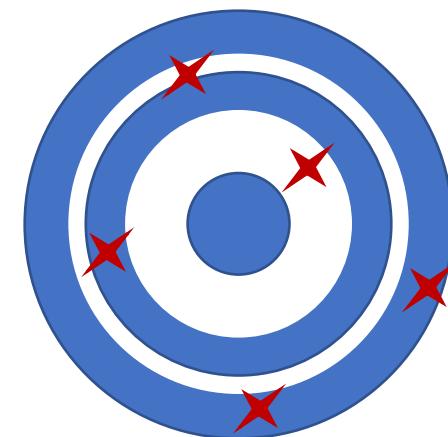
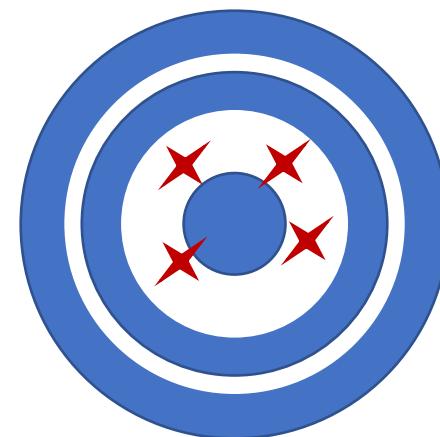
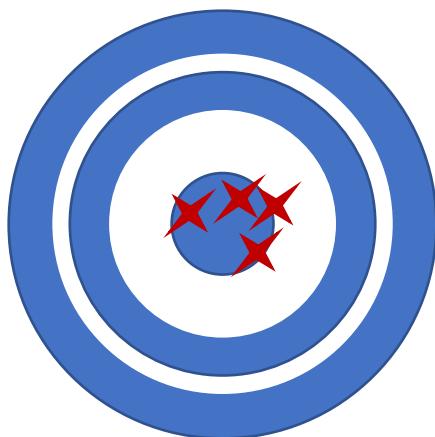
Dos consideraciones

- Sesgo: qué tanto se equivoca al predecir nuestra máquina.
- Varianza: problema de generalización - ¿se descontrola con datos desconocidos?



Dos consideraciones

- Sesgo: **errores de predicción**
- Varianza: **predicciones de entrenamiento vs prueba**



Evaluación de modelos

CLASIFICACIÓN



Photo by [Fernando Arcos](#) from [Pexels](#)

Sesgo: error de predicción - clasificación

Tasa de error de predicción:

$$TasaPrediccion = \frac{TotalErrores}{TotalPredicciones}$$

Sesgo: error de predicción - clasificación

Matriz de confusión

		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

Tasa de error de predicción:

$$TasaPrediccion = \frac{TotalErrores}{TotalPredicciones}$$

Evaluación de modelos

REGRESIÓN



Photo by [Fernando Arcos](#) from [Pexels](#)

Sesgo: error de predicción - regresión

Promedio de errores

Real	Predicho
1	2
2	4
3	5
4	6
5	8
6	6

Sesgo: error de predicción - regresión

Promedio de errores

Real	Predicho	Error
1	2	1
2	4	2
3	5	2
4	6	2
5	8	3
6	1	-5

Sesgo: error de predicción - regresión

Promedio de errores:

$$(1+2+2+2+3-5)/6$$

Real	Predicho	Error
1	2	1
2	4	2
3	5	2
4	6	2
5	8	3
6	1	-5

Sesgo: error de predicción - regresión

Promedio de errores:

$$(1+2+2+2+3-5)/6$$

Promedio de errores al cuadrado

para tener un valor positivo:

$$(1^2+2^2+2^2+2^2+3^2+(-5^2))/6$$

Real	Predicho	Error
1	2	1
2	4	2
3	5	2
4	6	2
5	8	3
6	1	-5

Variaciones adicionales

Del maravilloso mundo del machine
learning



Photo by [Matthew Montrone](#) from [Pexels](#)

Una vez inventado los modelos...

- Existen muchas formas de utilizar las herramientas
 - De validar las herramientas
 - De mejorar las herramientas
-
- No es necesario conocer todo el abanico de opciones; sólamente entender bien

Variaciones adicionales

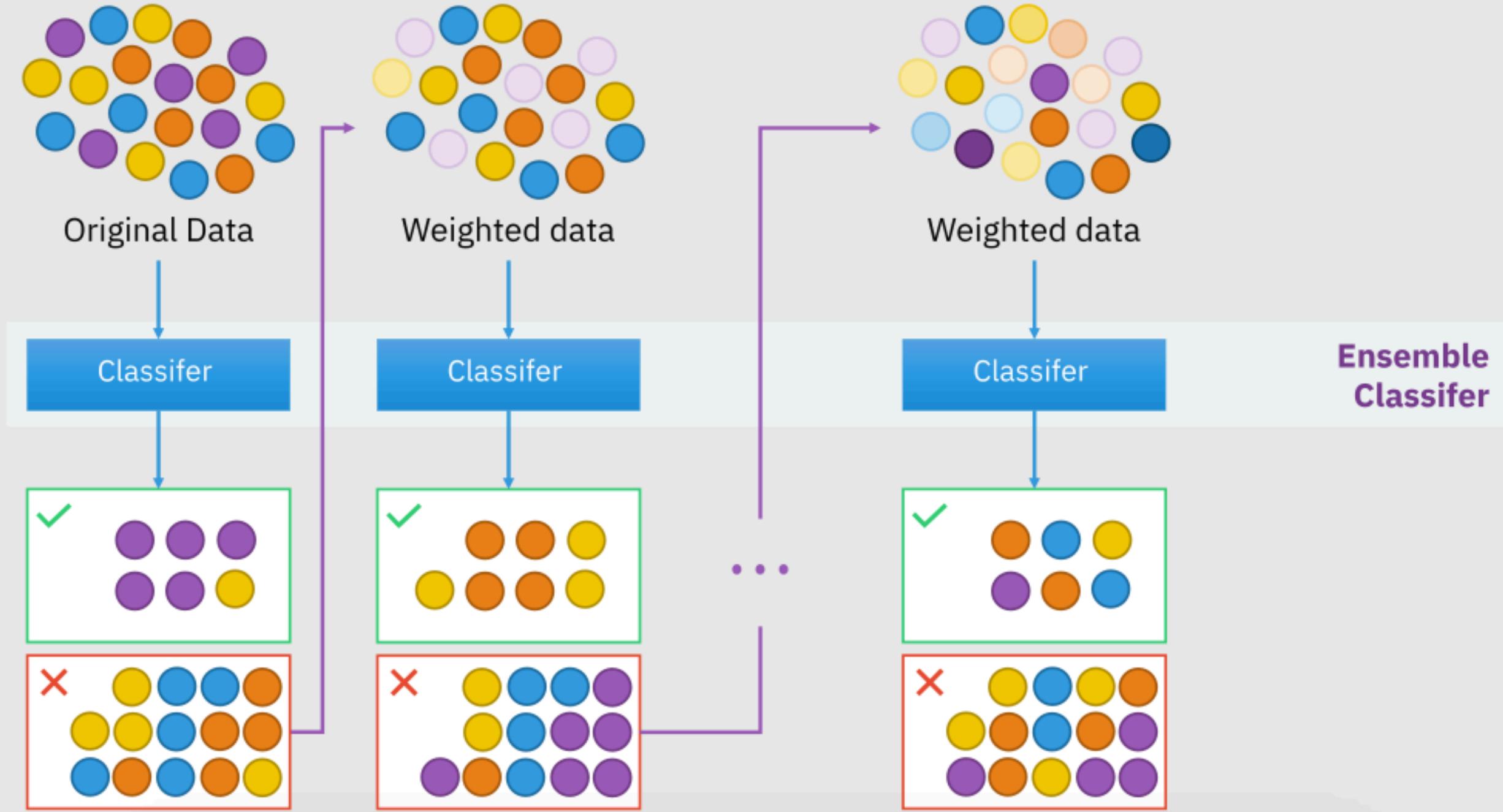
BOOSTING



Photo by [Matthew Montrone](#) from [Pexels](#)

Boosting

- ¿Puede un grupo de clasificadores débiles convertirse en un clasificador fuerte?
- Intuición similar a la de bosques aleatorios...
- Agregaciones de votación como las de bosques aleatorios no necesariamente permiten aprovechar las propiedades de cada clasificador débil



Boosting

- ¿Puede un grupo de clasificadores débiles convertirse en un clasificador fuerte?
- Intuición similar a la de bosques aleatorios...
- Agregaciones de votación como las de bosques aleatorios no necesariamente permiten aprovechar las propiedades de cada clasificador débil
- Existen varias formas de agregar clasificadores, boosting es una estrategia general que conduce a potenciar clasificadores pequeños.

Variaciones adicionales

VALIDACIÓN CRUZADA



Photo by [Matthew Montrone](#) from [Pexels](#)

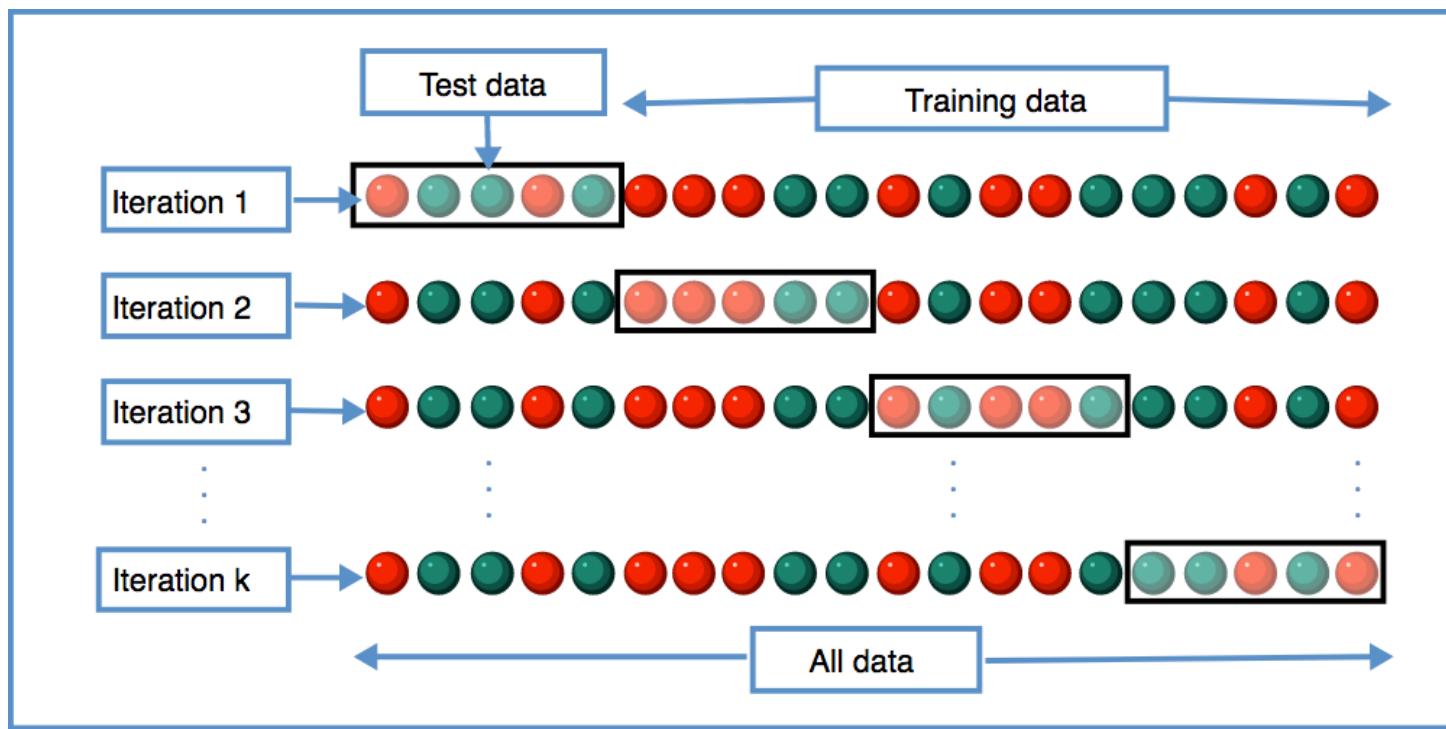
Validación cruzada

- ¿Qué tan bien generaliza?



Validación cruzada

- ¿Qué tan bien generaliza?

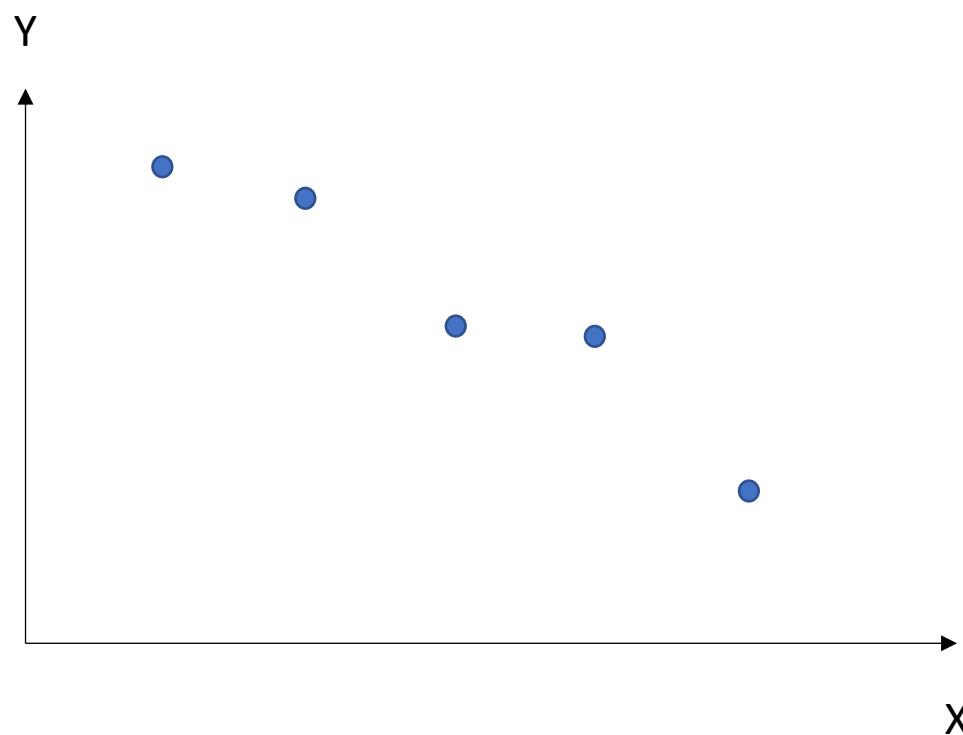


Se repite la separación train-test múltiples veces.. 1000 o 100 veces...

Se promedian u obtiene sd los errores de predicción, por ejemplo.

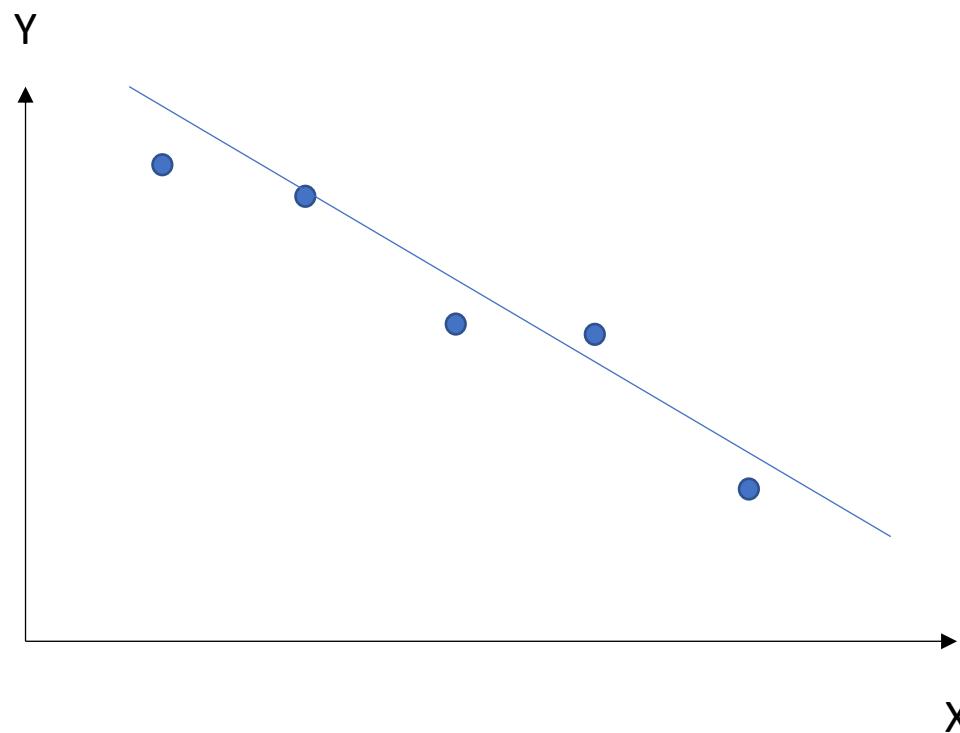
Regularización para regresión lineal

Recuerden que las regresiones toman un grupo de datos...



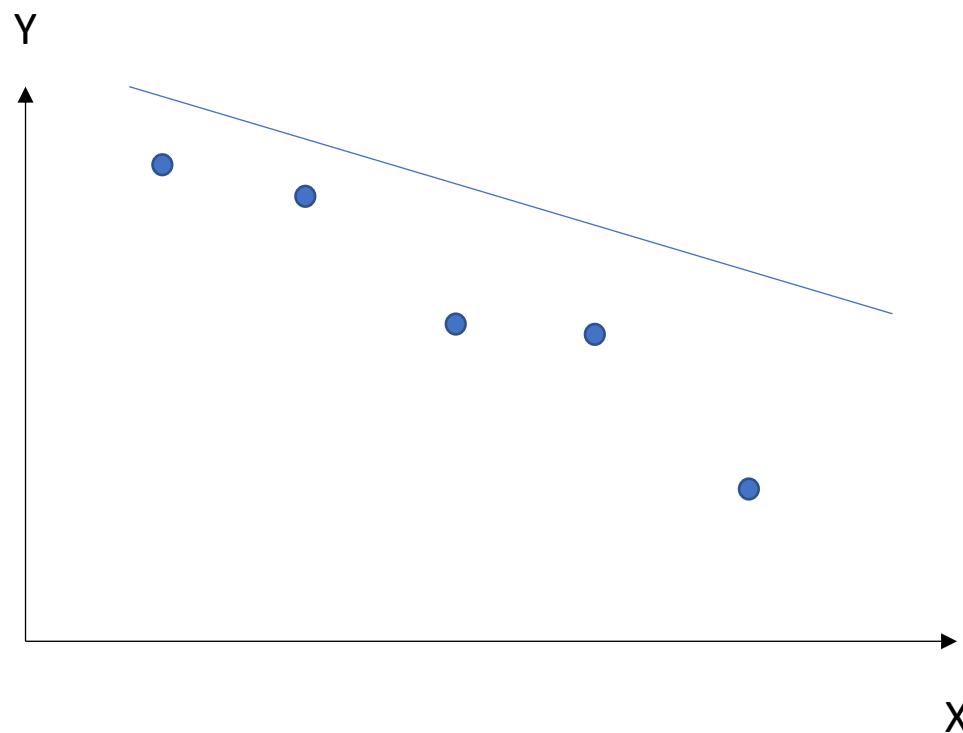
Regularización para regresión lineal

... y ajustamos una línea que “describe bien los datos”



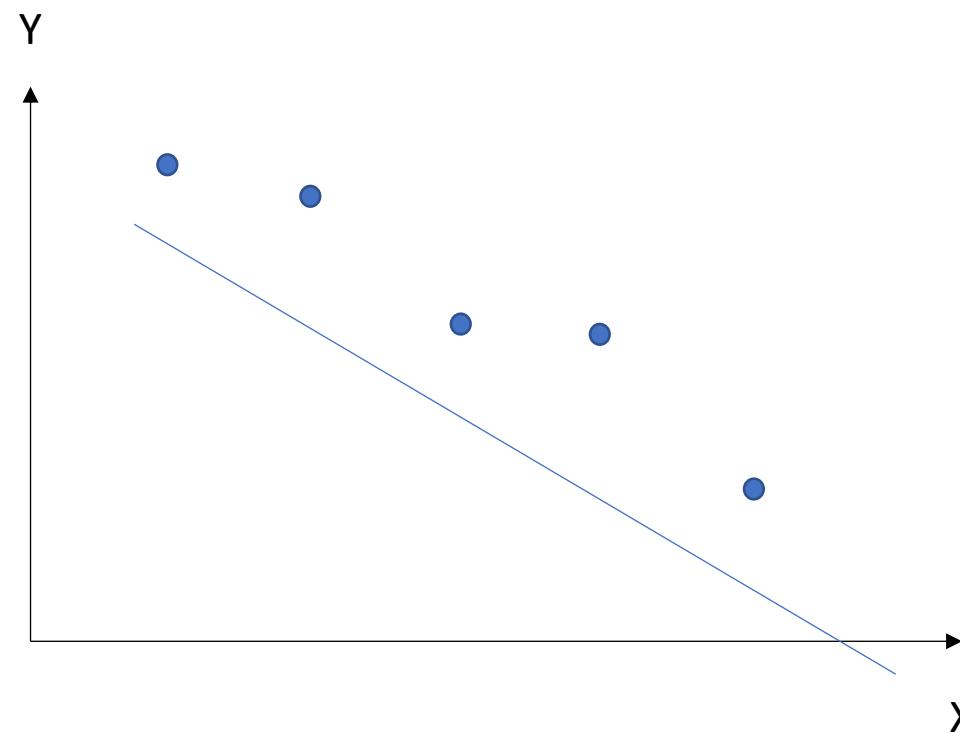
Regularización para regresión lineal

... se ajusta su pendiente...



Regularización para regresión lineal

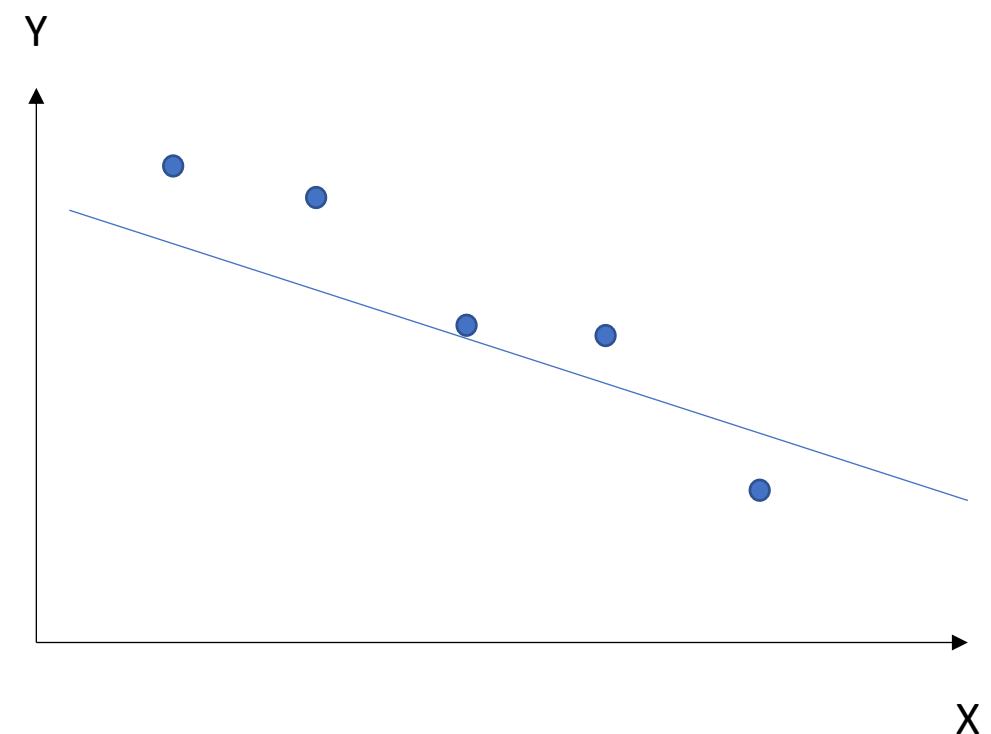
... y se ajusta su “altura”



Regularización para regresión lineal

Con regularización escogemos los valores **tales que**:

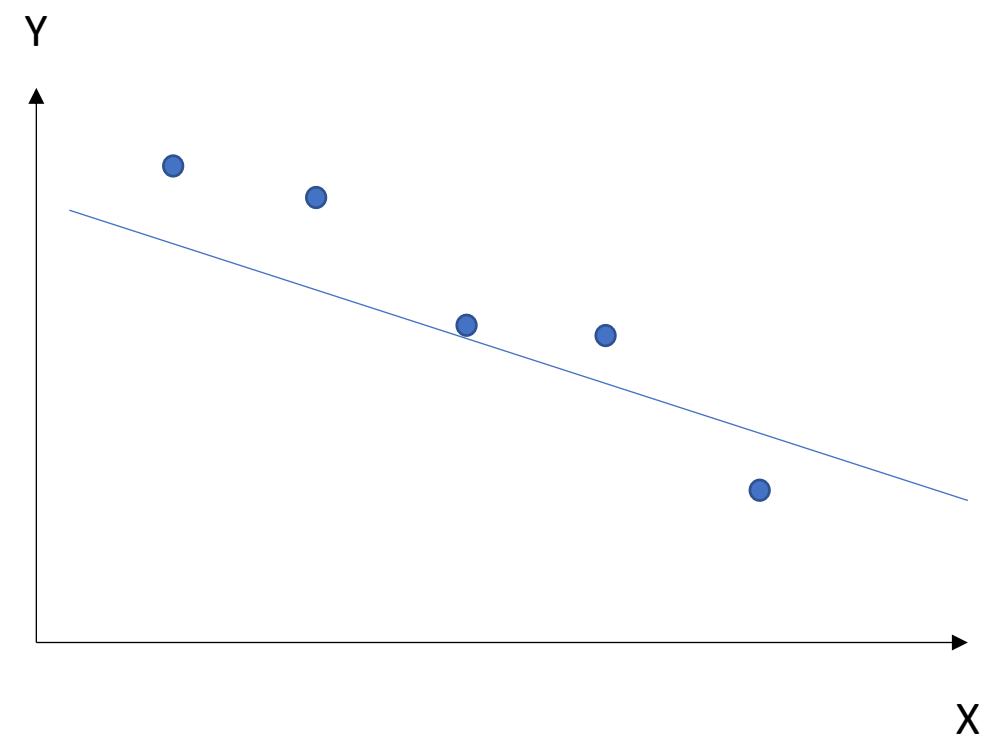
- Describen bien los datos...
- Y hace al modelo “un poquito” más rígido, de forma que sólo cosas muy estructurales muevan la línea



Regularización para regresión lineal

Con regularización escogemos los valores **tales que**:

- Describen bien los datos...
- Y hace al modelo “un poquito” más rígido, de forma que sólo cosas muy estructurales muevan la línea
- Esto hace un poco menos preciso el ajuste, pero incrementa la generalización



En conclusión

- No hay una forma estricta y única de aplicar las herramientas de aprendizaje de máquinas.
- Una vez tenemos **datos históricos** sobre los cuales aprender para decir cosas cuando aparezcan **datos nuevos** en el futuro...
- ... para entrenar herramientas para usar a futuro...
- Hay una cantidad de formas de probar, jugar, explorar profundizar.
- Invitación abierta a probar cosas y empaparse de **cómo lo hacen otros**: comunidades de aprendizaje en internet.

iScript!

Paquetes para correr modelos en R



Photo by [Christina Morillo](#) from [Pexels](#)