

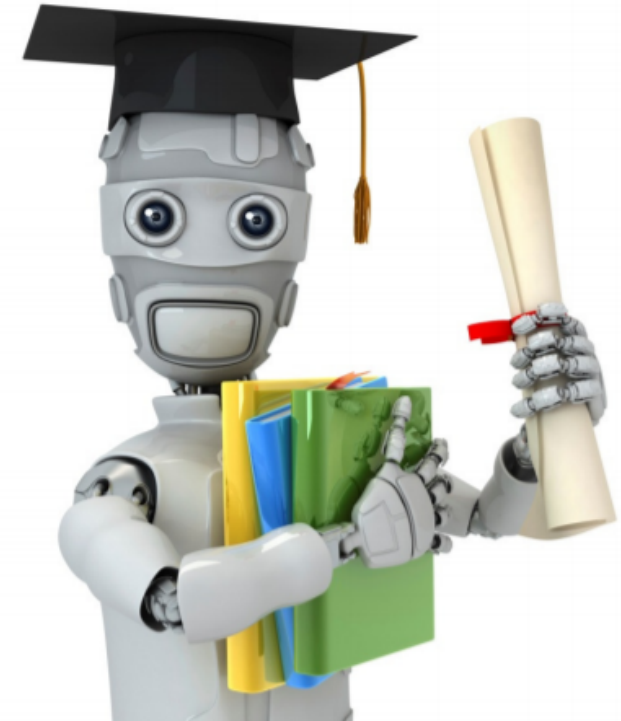
Aprendizaje Automático

Tecnológico de Costa Rica
Programa de Ciencia de Datos
Frans van Dunné

Agenda

- **Aprendizaje Automático**

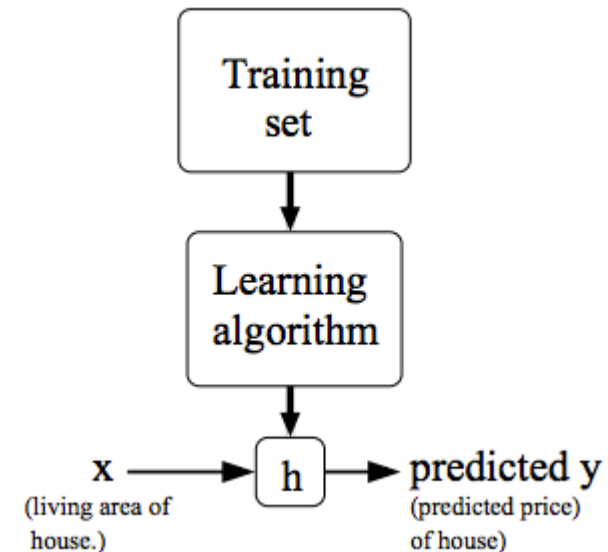
- Métodos de clasificación supervisada y no supervisada.
- Modelos paramétricos lineales de regresión: mínimos cuadrados y mínimos cuadrados regularizados.
- Ajuste polinomial de curvas.
- Selección del modelo (sobreajuste) y validación cruzada.
- La maldición de la dimensionalidad.



Aprendizaje supervisado versus aprendizaje no supervisado

Algoritmos supervisados

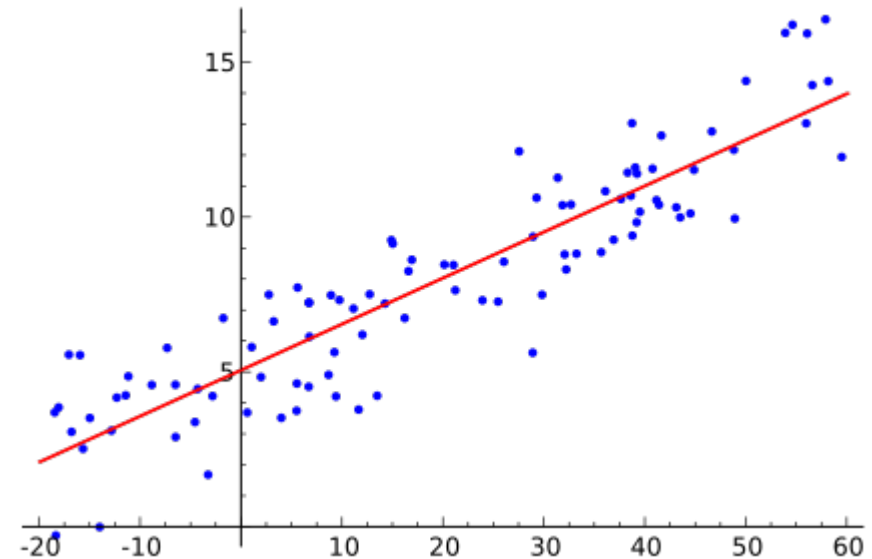
- Se entrena al algoritmo otorgándole las preguntas, denominadas características, y las respuestas, denominadas etiquetas. Esto se hace con la finalidad de que el algoritmo las combine y pueda hacer predicciones.



Dos tipos de aprendizaje supervisado

- **Regresión**

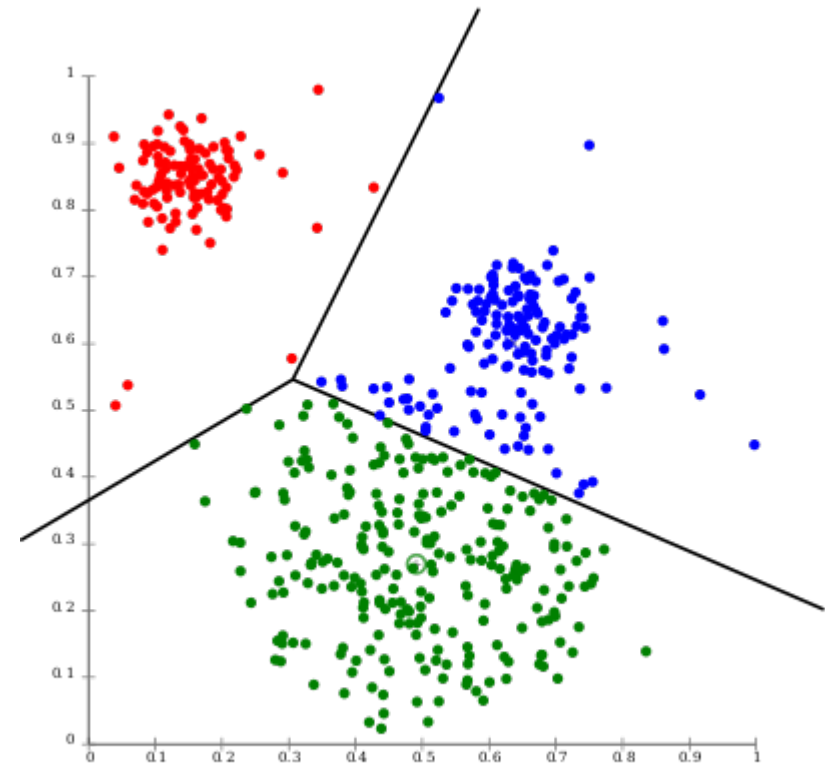
- Tiene como resultado un número específico. Si las etiquetas suelen ser un valor numérico, mediante las variables de las características, se pueden obtener dígitos como dato resultante.



Dos tipos de aprendizaje supervisado

- **Clasificación:**

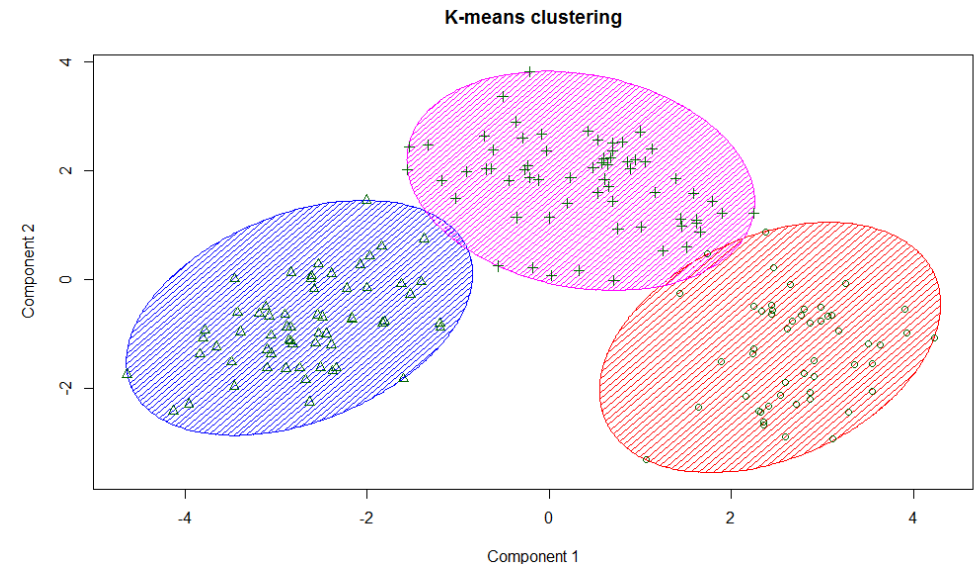
- En este tipo, el algoritmo encuentra diferentes patrones y tiene por objetivo clasificar los elementos en diferentes grupos.



Aprendizaje supervisado versus aprendizaje no supervisado

- **Algoritmos no supervisados**

- Solo se le otorgan las características, sin proporcionarle al algoritmo ninguna etiqueta. Su función es la **agrupación**, por lo que el algoritmo debería catalogar por similitud y poder crear grupos, sin tener la capacidad de definir cómo es cada individualidad de cada uno de los integrantes del grupo.



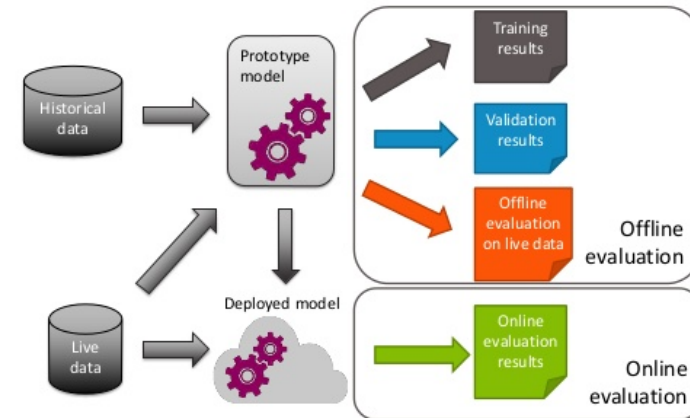
Evaluación de la precisión del modelo

Ningún método domina a los demás, o domina sobre un conjunto de datos particular.

Por lo tanto, es una tarea importante decidir sobre cualquier conjunto de datos qué método produce los mejores resultados.

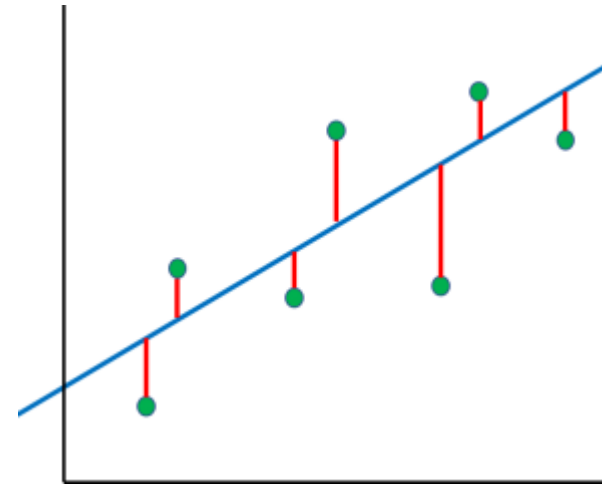
Seleccionar el mejor enfoque puede ser una de las partes más difíciles de realizar el aprendizaje automático en práctica.

When to evaluate



Medición de la calidad del ajuste - MSE

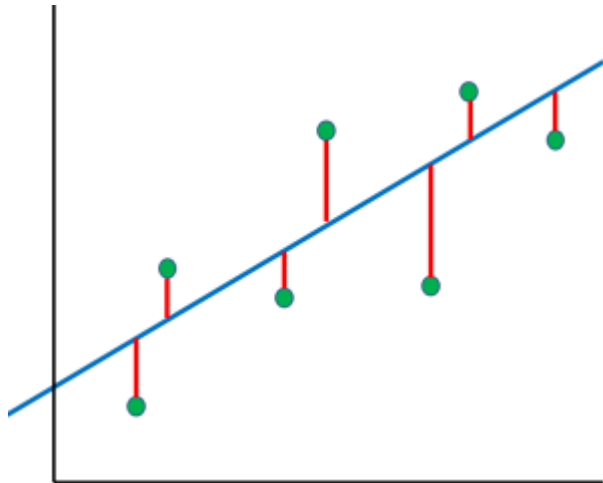
- El Error Cuadrático Medio es el criterio de evaluación más usado para problemas de regresión. Se usa sobre todo cuando usamos aprendizaje automático supervisado.



Medición de la calidad del ajuste - MSE

- En la figura vemos que estamos usando una regresión lineal (en azul) para estimar los datos que tenemos (los puntos verdes). El modelo lineal tiene un error (en rojo) que podemos definir con la siguiente fórmula:

$$\text{error cuadrático} = (\text{real} - \text{estimado})^2$$



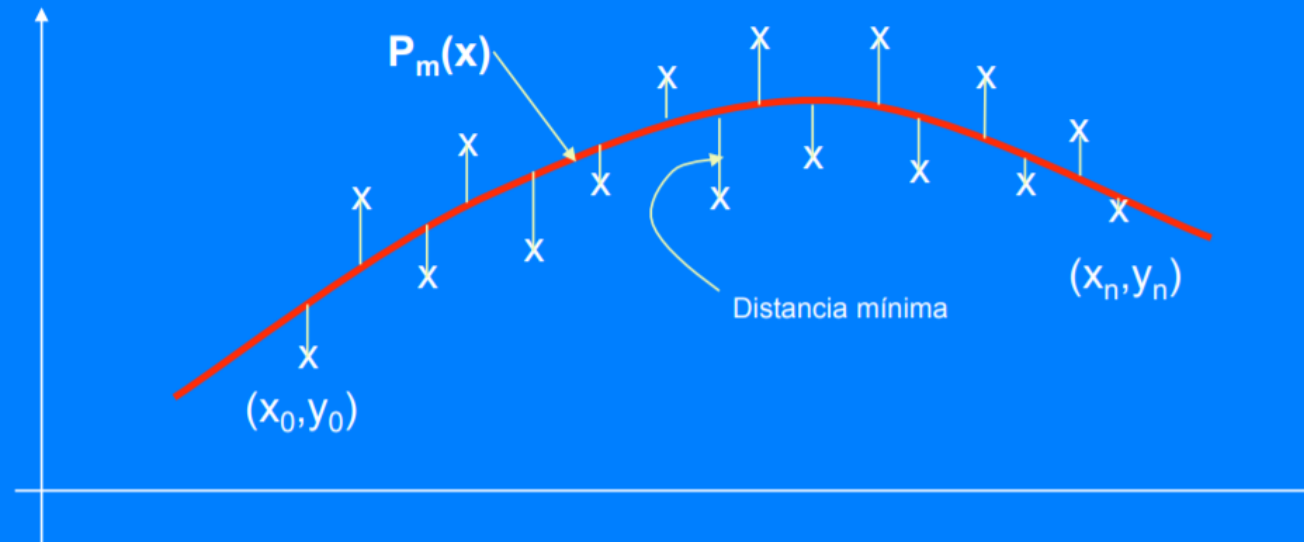
El valor estimado es el valor que nos da el modelo. En este caso, la línea azul.

Calculamos el error al cuadrado, en lugar del error simple, para que el error siempre sea positivo. De esta forma sabemos que el error perfecto es 0. Si no elevásemos el error al cuadrado, unas veces el error sería positivo y otras negativo. Otra posibilidad sería usar el valor absoluto, en lugar de elevarlo al cuadrado. Sin embargo, si usamos el valor absoluto, obtendremos una función no-derivable.

Ahora que sabemos cómo calcular el error en cada punto, podemos calcular cual es el error medio. Para ello, sumamos todos los errores y los dividimos entre el número total de puntos. Si llamamos M al número total de puntos nos queda la fórmula del Error Cuadrático Medio (MSE, por sus siglas en inglés, Mean Squared Error):

Ajuste polinomial de curvas.

Aproximación polinomial por mínimos cuadrados

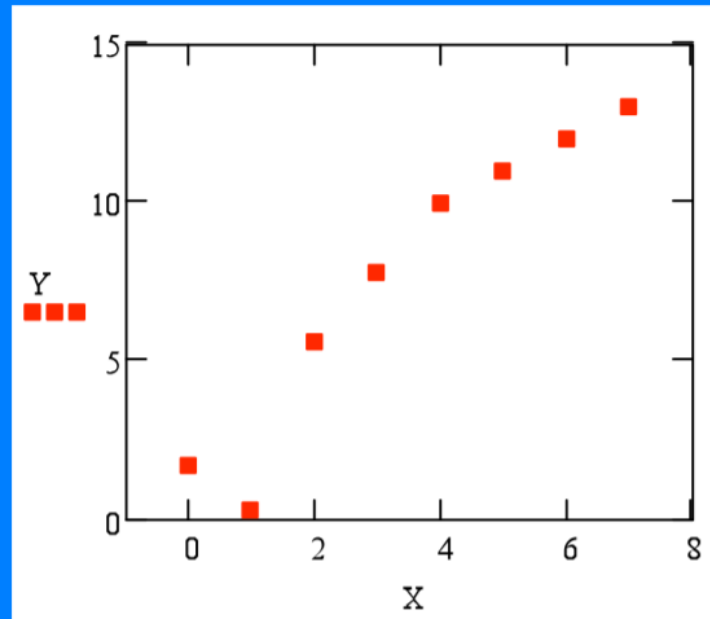


Objetivo: Obtener un polinomio o función que relaciones x e y

Ajuste polinomial de curvas

Ejemplo:
Se tiene la siguiente secuencia de datos:

X	0.0	1.0	2.0	3.0	4.0	5.0	6.0	7.0
Y	1.7	0.3	5.6	7.8	10.	11.	12.	14.



Descenso de gradiente

- Ejercicios demostrativos (CLICK)

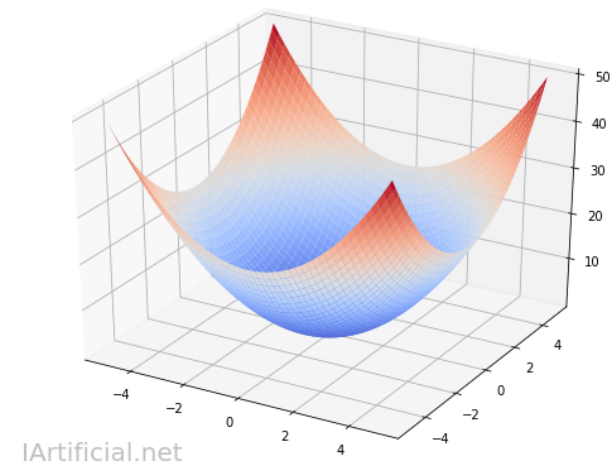
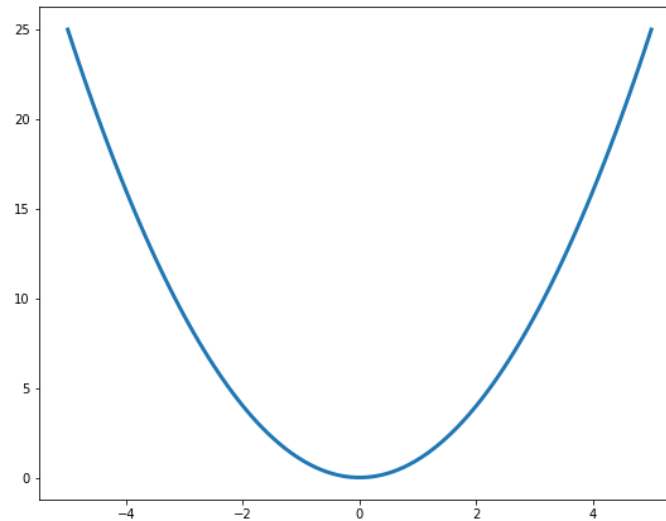


decenso_de_gradiente.r



decenso_de_gradiente_2.Rmd

$$RMSE = \sqrt{\frac{1}{M} \sum_{i=1}^M (real_i - estimado_i)^2}$$

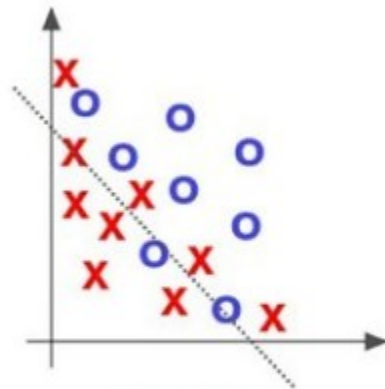


IArtificial.net

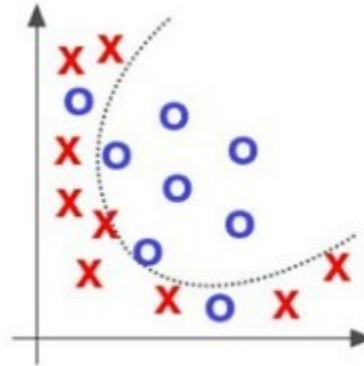
¿Cómo reconocer el sobreajuste?

- El sobreajuste va a estar relacionado con la complejidad del modelo, mientras más complejidad le agreguemos, mayor va a ser la tendencia a sobreajustarse a los datos, ya que va a contar con mayor flexibilidad para realizar las predicciones y puede ser que los patrones que encuentre estén relacionados con el ruido (pequeños errores aleatorios) en los datos y no con la verdadera señal o relación subyacente.
- No existe una regla general para establecer cual es el nivel ideal de complejidad que le podemos otorgar a nuestro modelo sin caer en el sobreajuste; pero podemos valernos de algunas herramientas analíticas para intentar entender como el modelo se ajusta a los datos y reconocer el sobreajuste.

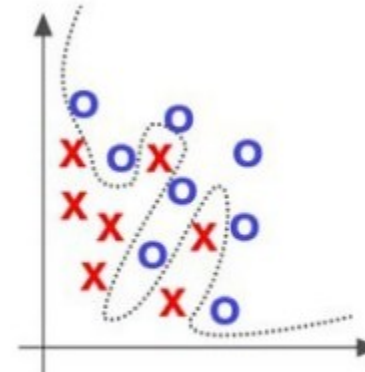
¿Cómo reconocer el sobreajuste? **TEC** | Tecnológico de Costa Rica



Subajuste



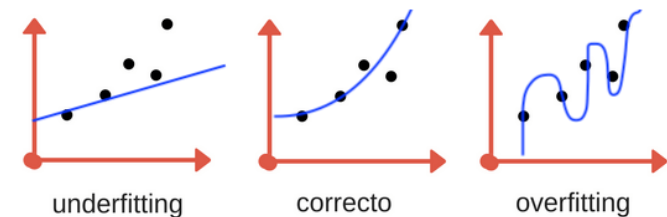
Apropiado



Sobreajuste

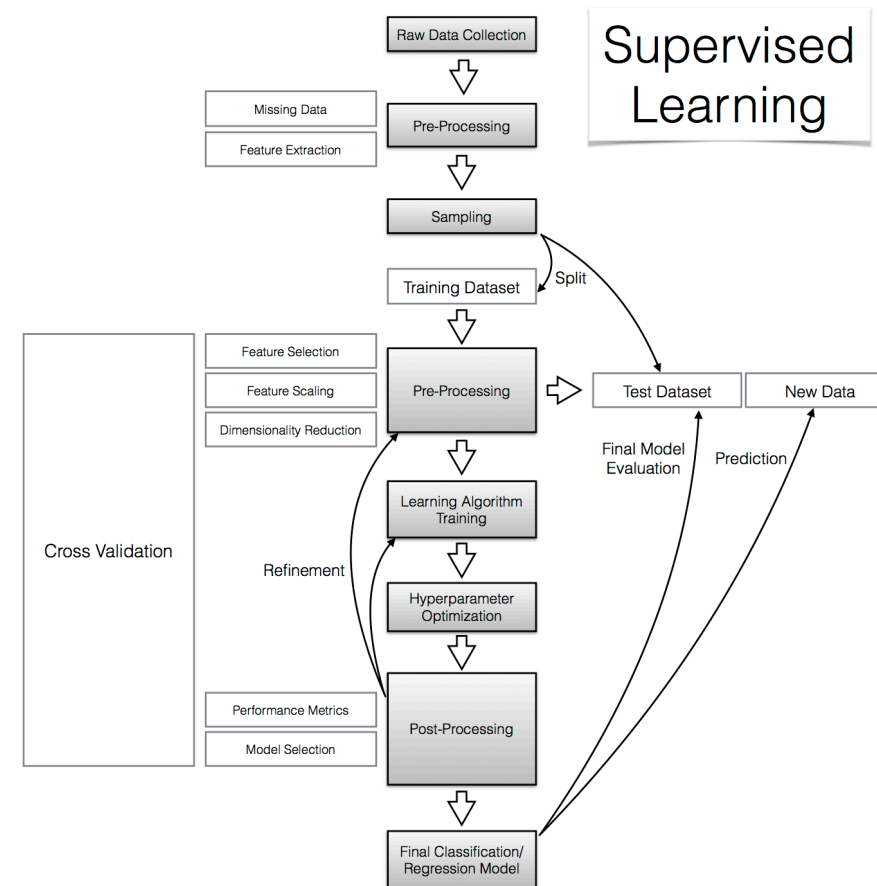
Sobreajuste

- **El sobreajuste se refiere a un modelo que modela los datos de entrenamiento demasiado bien.** Esto ocurre cuando un modelo aprende el detalle, incluyendo el ruido en los datos de entrenamiento en la medida en que tiene un impacto negativo en el rendimiento del modelo en datos nuevos. Esto significa que el ruido o las fluctuaciones aleatorias en los datos de entrenamiento son recogidos y aprendidos por el modelo. El problema es que estos conceptos no se aplican a los datos nuevos y afectan negativamente a la capacidad de los modelos para generalizar.



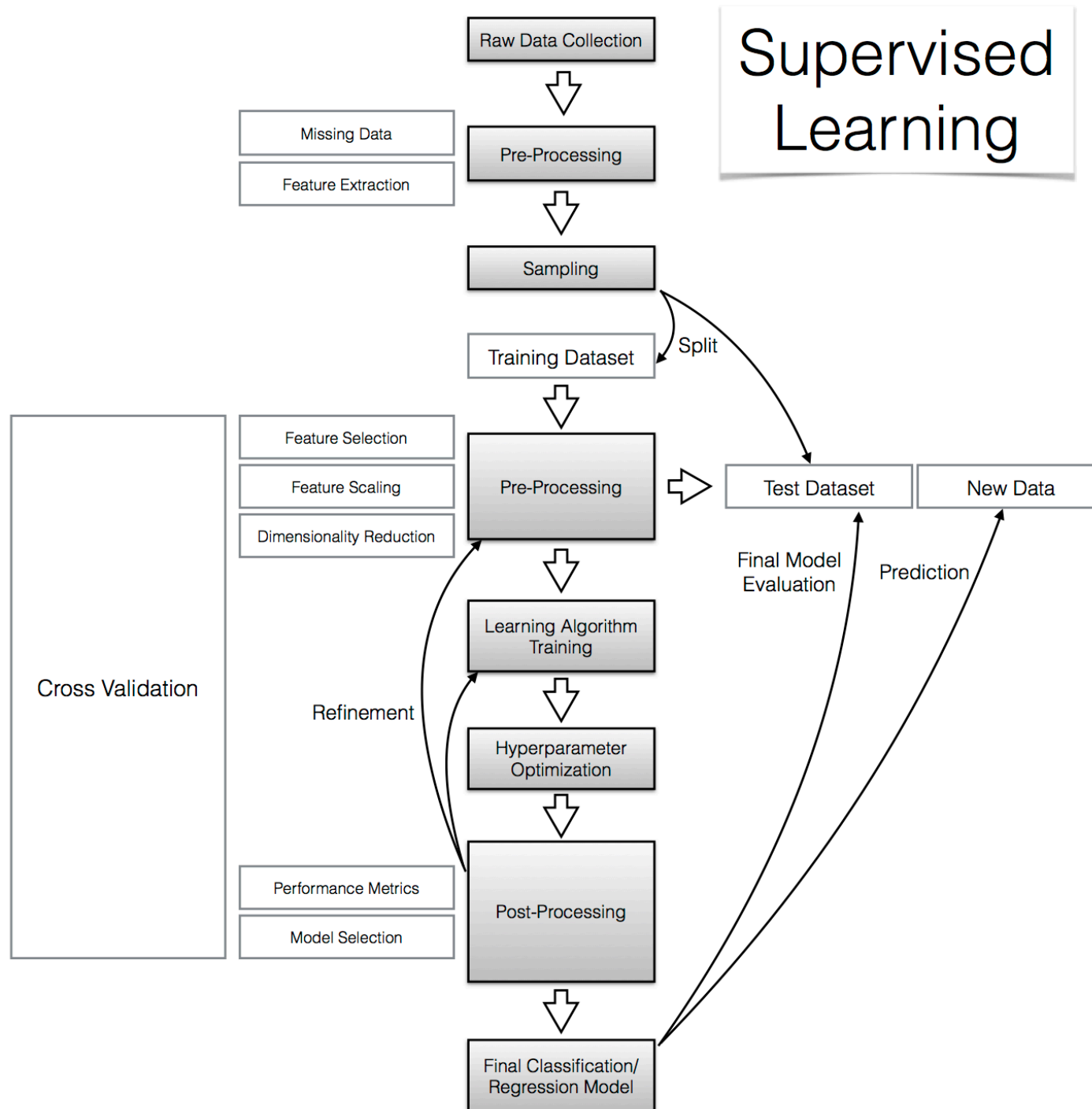
Testing data

http://sebastianraschka.com/Articles/2014_intro_supervised_learning.html

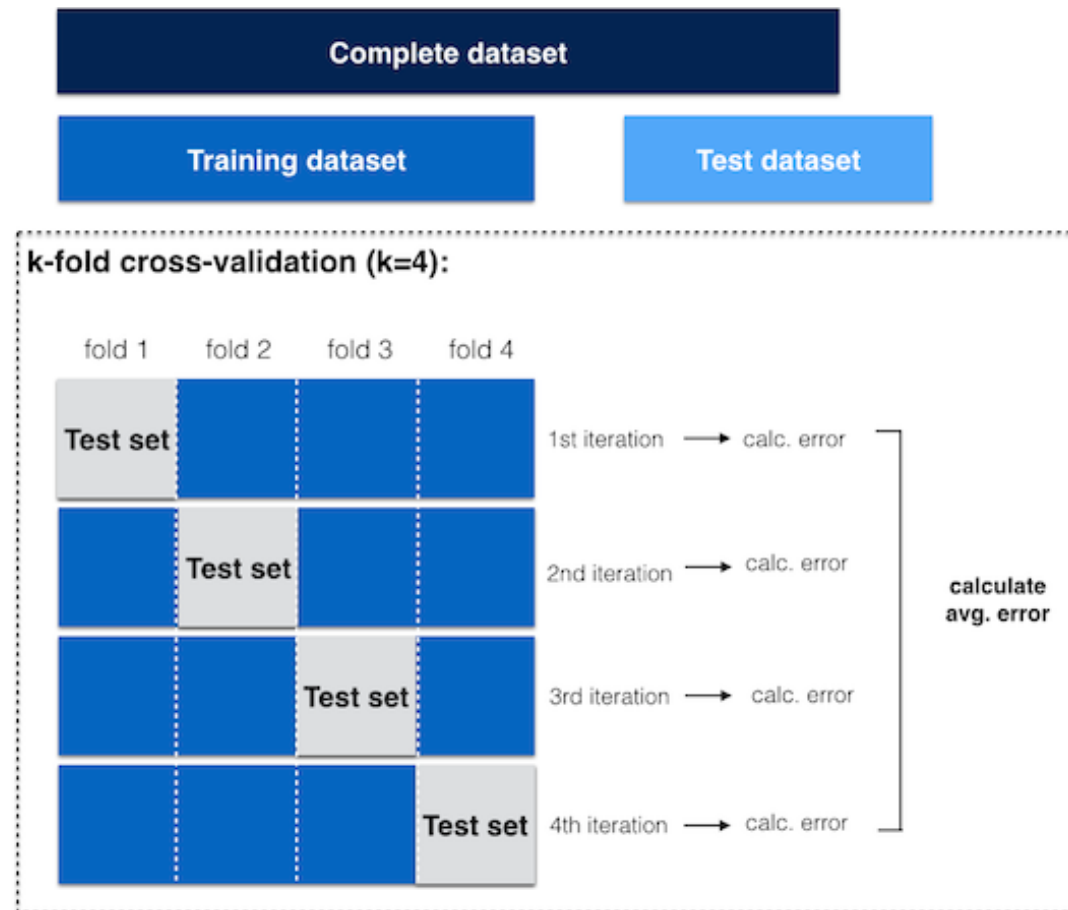


Tes

<http://sekom2015.com>



Ejemplo de validación cruzada

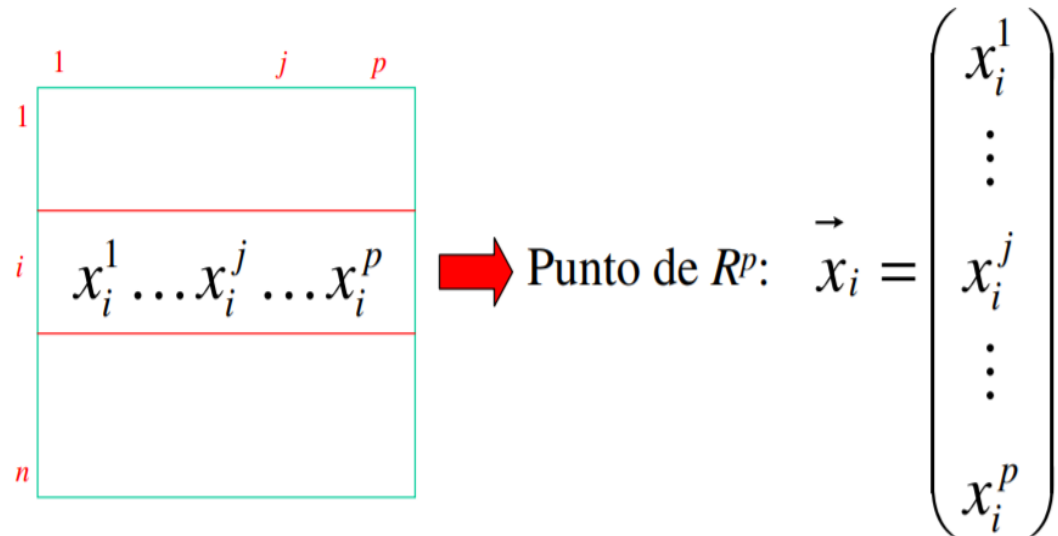


$N > P$ and $N < P$

- Inferencia
- Colineación de las variables

X : n individuos descritos por p variables cuantitativas.

X : matriz $n \times p$

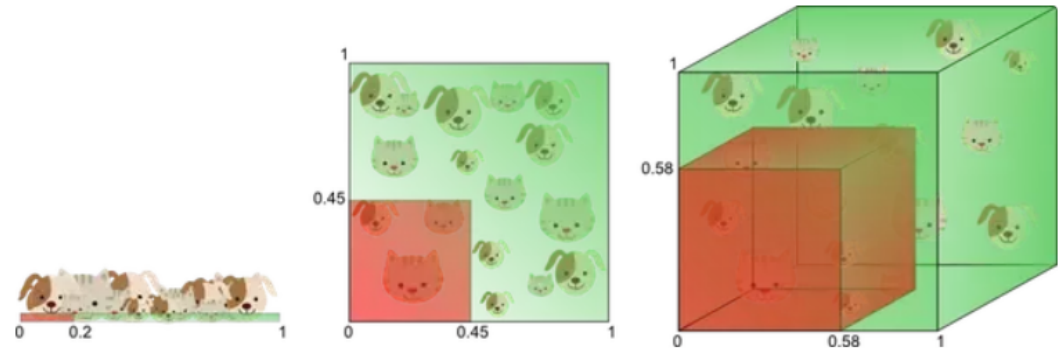


	1		j		p
1					
i	x_i^1	\dots	x_i^j	\dots	x_i^p
n					

➡ Punto de R^p : $\vec{x}_i = \begin{pmatrix} x_i^1 \\ \vdots \\ x_i^j \\ \vdots \\ x_i^p \end{pmatrix}$

La maldición de la dimensionalidad

- Se trata de una serie de problemas que emergen cuando los datos tiene gran dimensionalidad (muchas variables), que causan gran dispersión de los datos.



Bibliografía

- http://www.stat.rice.edu/~jrojo/PASI/lectures/Costa%20rica/1_Introduccion.pdf
- <https://dlegorreta.wordpress.com/tag/gradiente-descendente/>
- http://sebastianraschka.com/Articles/2014_intro_supervised_learning.html
- <http://www.diegocalvo.es/validacion-cruzada-en-r/>
- <http://ligdigonzalez.com/sobreaajuste-y-subajuste-en-machine-learning/>
- <https://relopezbriega.github.io/blog/2016/05/29/machine-learning-con-python-sobreaajuste/>
- <https://ccc.inaoep.mx/~emorales/Cursos/NvoAprend/Acetatos/EvaluacionSobreAjuste.pdf>
- <http://apuntes-r.blogspot.com/2014/11/validacion-cruzada.html>