

MINERÍA DE DATOS.

Hugo Andrés Dorado B.



Contenido

- Definiciones en minería de datos.
- Tipo de aprendizaje.
- Algoritmo de predicción.
- Tipos error.
- Sobre parametrización.
- Diseño del estudio.
- Validación cruzada.

Definiciones


- **Big data:** es una tendencia que hace referencia al **almacenamiento de grandes volúmenes de datos** y a los procedimientos usados para encontrar patrones repetitivos dentro de esos datos.
- **Minería de datos:** Es un campo de las ciencias de la computación que tiene como propósito descubrir **patrones en grandes volúmenes** de conjuntos de datos.
Utiliza los métodos de la inteligencia artificial, aprendizaje automático y estadística.

Definiciones

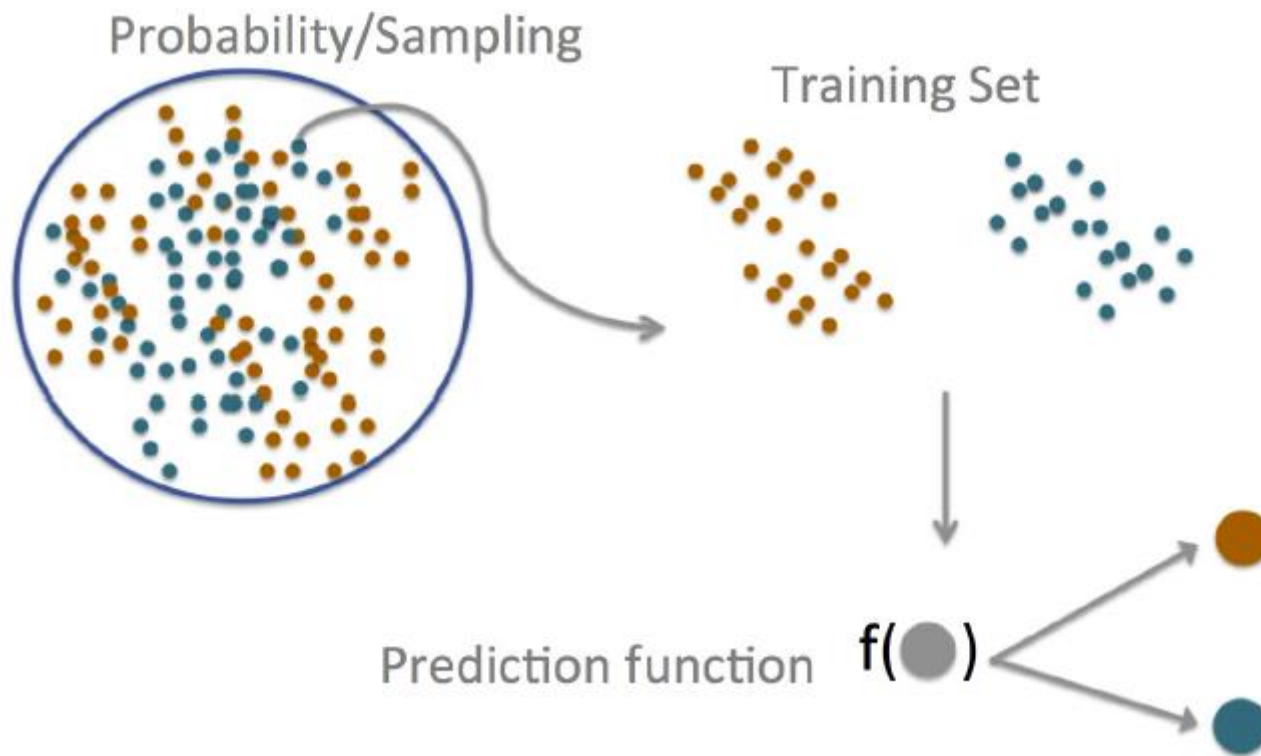
- **Características, variables de entrada (Features):** Variables medidas sobre las observaciones que se asocian luego a un variable salida.
- **Variable de salida:** Variable a explicar de interés.
- **Función costo:** Es una función que permite aproximar un conjunto de variables de entrada para generar una respuesta aproximada según la variable de salida.

$$Y = f(X) + \varepsilon$$

Tipo de aprendizaje

- **Aprendizaje supervisado:** se deduce una función, de acuerdo a un conjunto de variables de salida para la reducción de un error.
 - Clasificación.
 - Regresión.  Predicción o interpretación
- **Aprendizaje no supervisado:** No hay una variable de salida, se busca compresión de los datos tratando un conjunto de variables de entrada.
 - Clustering.
 - Componentes principales.

Predicción en modelos de machine learning



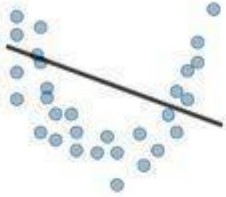


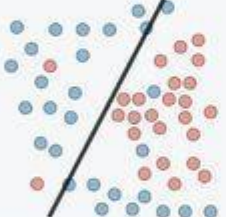
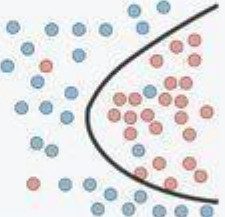
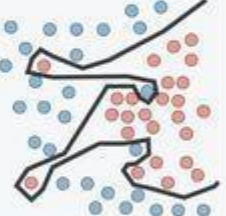

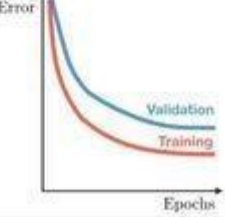

Algoritmos de predicción

- **Pregunta:** ¿Que mails son spam?, ¿Que zonas son bosque?, ¿Que clientes serán morosos?
- **Entrada de datos:** conjuntos de e-mail, Imágenes satelitales, información de clientes. (Ya clasificados)
- **Variables de entrada:** Frecuencia de ciertas palabras, índices espectrales por color, variables seleccionadas
- **Algoritmo:** Redes neuronales artificiales, support vector machine, J46
- **Parámetros:** (Tasa de decaimiento, neuronas ocultas) ,(costo), (umbral de confianza)
- **Evaluación.** (Precisión, exactitud, concordancia)

Tipos de error.

- **Error dentro de la muestra:** La tasa de error que se obtiene en los mismos datos para construir el modelo.
- **Error fuera de la muestra:** La tasa de error que se obtiene al traer nuevos datos no mostrados, también conocido como error de generalización.

Overfitting – sobre parametrización

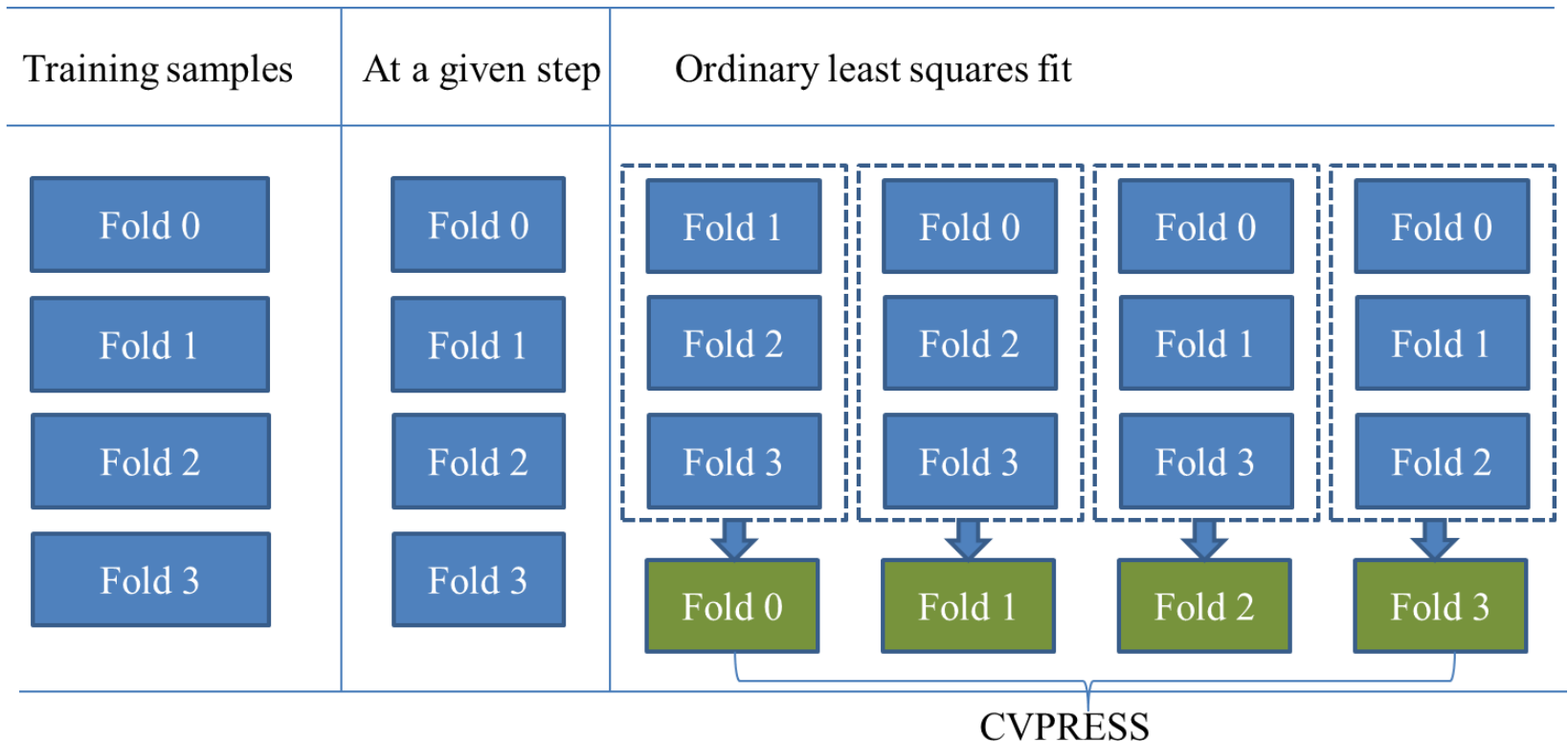
	Underfitting	Just right	Overfitting
Symptoms	<ul style="list-style-type: none"> • High training error • Training error close to test error • High bias 	<ul style="list-style-type: none"> • Training error slightly lower than test error 	<ul style="list-style-type: none"> • Very low training error • Training error much lower than test error • High variance
Regression illustration			
Classification illustration			
Deep learning illustration			
Possible remedies	<ul style="list-style-type: none"> • Complexify model • Add more features • Train longer 		<ul style="list-style-type: none"> • Perform regularization • Get more data

Generalización
Nuevos datos

Diseño de estudio para el conjunto de datos

1. Definir una tasa de error.
2. Partir el conjunto de datos en:
Entrenamiento, prueba y validación (opcional)
(60,20,20) Grandes; (60,40) medianos
3. Sobre el conjunto de entrenamiento hacer selección de variables de entradas
4. Sobre el conjunto de entrenamiento realizar optimización de parámetros. (Utilizar cross validation)
5. Validar de acuerdo a la tasa de error.

K - fold



Medir el desempeño

		DISEASE	
		+	-
TEST	+	TP	FP
	-	FN	TN

TP: Verdadero positivo.

FP: Falso positivo.

FN: Falso negativo.

TN: Verdadero negativo.

Mean squared error (MSE):

$$\frac{1}{n} \sum_{i=1}^n (\text{Prediction}_i - \text{Truth}_i)^2$$

Root mean squared error (RMSE):

$$\sqrt{\frac{1}{n} \sum_{i=1}^n (\text{Prediction}_i - \text{Truth}_i)^2}$$

Sensitivity

$$\rightarrow \text{TP} / (\text{TP} + \text{FN})$$

Specificity

$$\rightarrow \text{TN} / (\text{FP} + \text{TN})$$

Positive Predictive Value

$$\rightarrow \text{TP} / (\text{TP} + \text{FP})$$

Negative Predictive Value

$$\rightarrow \text{TN} / (\text{FN} + \text{TN})$$

Accuracy

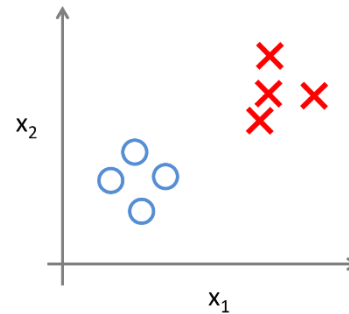
$$\rightarrow (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{FN} + \text{TN})$$

Métodos en machine learning para implementar

Modelos supervisados.

- Redes neuronales artificiales
- Árboles de clasificación y regresión.
- Random forest.
- Support vector machine.

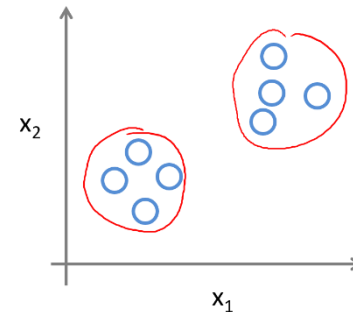
Supervised Learning



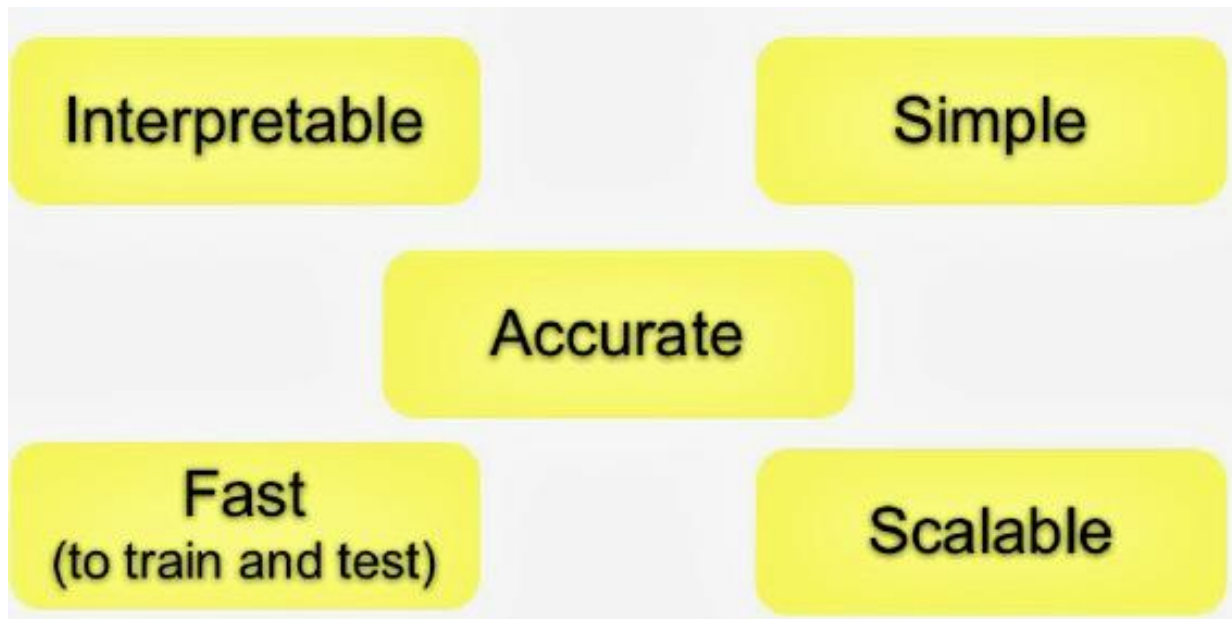
Modelos no supervisados.

- Cluster jerárquico.
- Kmeans
- PCA

Unsupervised Learning



El mejor método de aprendizaje de máquina



Resumen

Buscar datos:

- <https://archive.ics.uci.edu/ml/datasets.html>

Filtrar datos:

- Funciones básicas en R, desde la lectura.

Análisis exploratorio

Transformar datos:

- Merge, dcast, plyr

Determinar estrategia de partición de datos de entrenamiento y validación.

- Seleccionar atributos.
- Optimizar parámetros.

Escoger el modelo final, resultados y pruebas de tasa de error.

Transferir resultados a usuario.

Bibliografía

- <http://caret.r-forge.r-project.org/>
- <http://www.rdatamining.com/>
- <http://ucanalytics.com/blogs/learn-r-12-books-and-online-resources/>
- <https://www.coursera.org/specializations/jhu-data-science>

GRACIAS!!!!

