Apache Spark
Machine Learning in Scale

Diego J. Bodas Sagi

Disclaimer

MBIT School
Madrid Business Intelligence Technology

- Understanding WHY

Índice - Debate

MBIT School
Madrid Business Intelligence Technology

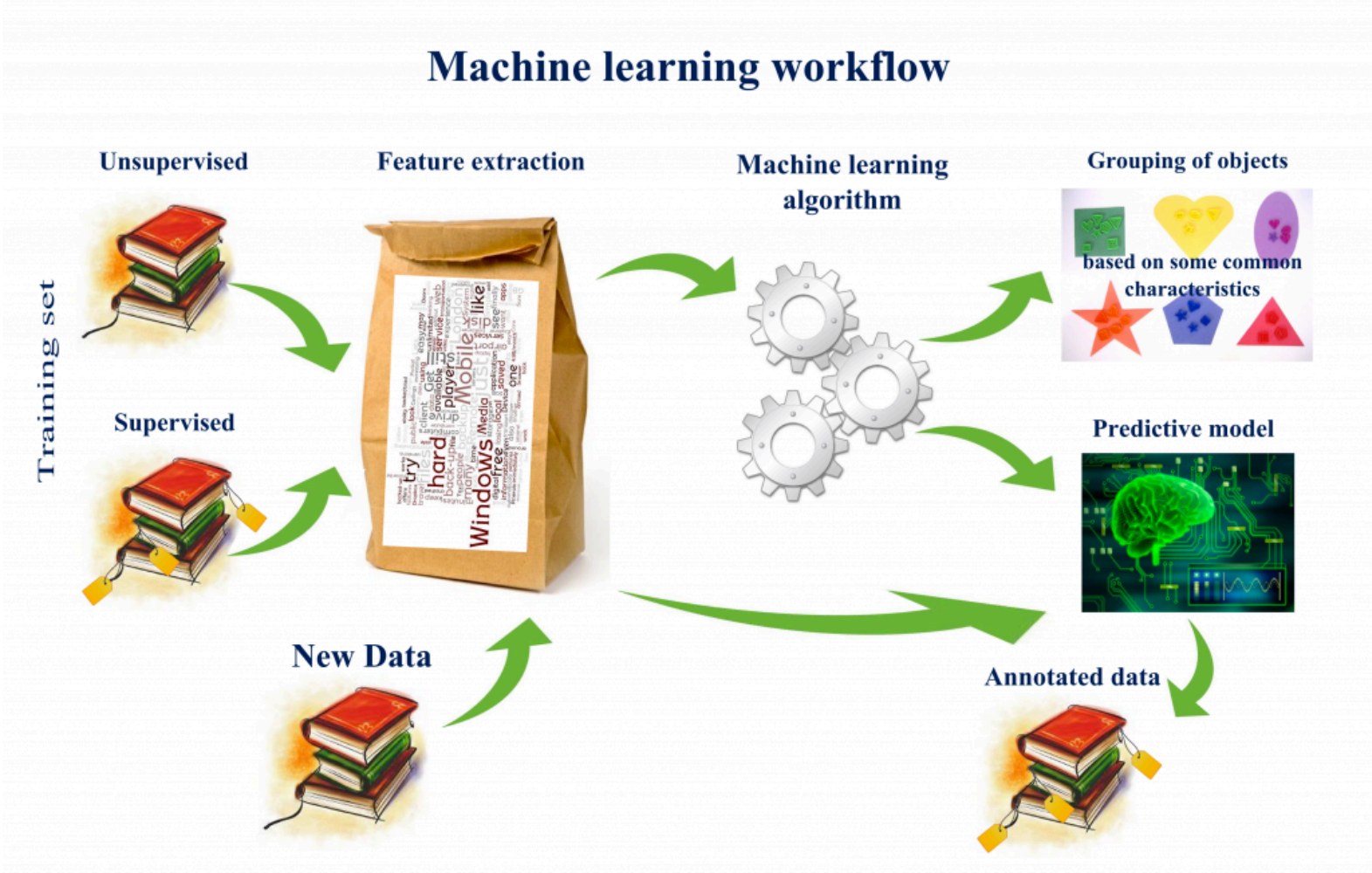
- Machine learning, ¿qué es?
- ¿Qué algoritmos conocéis?
- ¿Qué es la escalabilidad?
- ¿Dónde está el cuello de botella de los algoritmos que conocéis?
- Un cuello de botella típico
- Complejidad

Preguntas para vosotros

MBIT School
Madrid Business Intelligence Technology

- ¿Qué es Machine Learning?
- ¿Qué algoritmos conocemos?
- ¿Qué es la escalabilidad?
- ¿Dónde está el cuello de botella de los algoritmos que conocéis?

Machine Learning



Source: http://nkonst.com/machine-learning-explained-simple-words/

Machine Learning Algorithms *(sample)*

	Unsupervised	Supervised

Continuous	• Clustering & Dimensionality Reduction
	  ○ SVD
	  ○ PCA
	  ○ K-means
		• Regression
		  ○ Linear
		  ○ Polynomial
		• Decision Trees
		• Random Forests

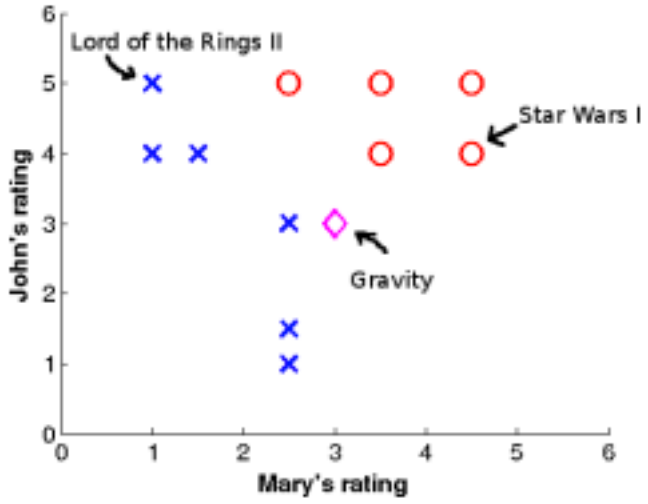
Categorical	• Association Analysis
	  ○ Apriori
	  ○ FP-Growth
	• Hidden Markov Model
		• Classification
		  ○ KNN
		  ○ Trees
		  ○ Logistic Regression
		  ○ Naive-Bayes
		  ○ SVM

Una visión un poco más amplia

MBIT School
Madrid Business Intelligence Technology

- From: https://cs.stanford.edu/~quocle/tutorial1.pdf

Movie name	Mary's rating	John's rating	I like?
Lord of the Rings II	1	5	No
...	...	...	...
Star Wars I	4.5	4	Yes
Gravity	3	3	?



Función de decisión

MBIT School
Madrid Business Intelligence Technology

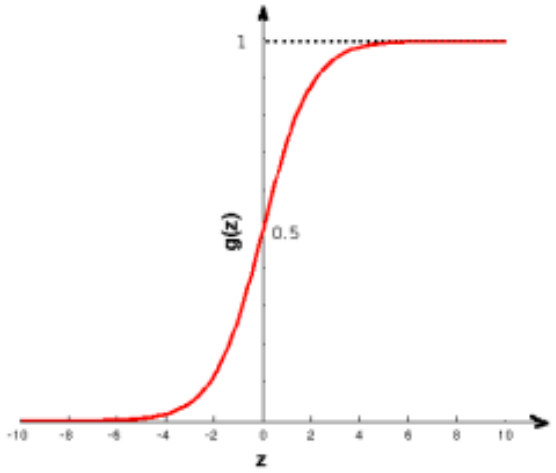
Our decision function can be as simple as a weighted linear combination of Mary's and John's ratings:

$$h(x;\theta,b) = \theta_1 x_1 + \theta_2 x_2 + b, \quad \text{which can also be written as} \quad h(x;\theta,b) = \theta^T x + b \tag{1}$$

A simple way to force h to have values between 0 and 1 is to map it through another function called the sigmoid function, which is bounded between 0 and 1:

$$h(x;\theta,b) = g(\theta^T x + b), \text{ where } g(z) = \frac{1}{1+\exp(-z)}, \tag{2}$$

which graphically should look like this:



The value of function h is now bounded between 0 and 1.

Buscando los coeficientes

MBIT School
Madrid Business Intelligence Technology

To find the values of θ and b we can try to minimize the following *objective function*, which is the sum of differences between the decision function h and the label y:

$$J(\theta, b) = \left(h(x^{(1)}; \theta, b) - y^{(1)}\right)^2 + \left(h(x^{(2)}; \theta, b) - y^{(2)}\right)^2 + \ldots + \left(h(x^{(m)}; \theta, b) - y^{(m)}\right)^2$$
$$= \sum_{i=1}^{m} \left(h(x^{(i)}; \theta, b) - y^{(i)}\right)^2$$

Stochastic Gradient Descent

MBIT School
Madrid Business Intelligence Technology

To minimize the above function, we can iterate through the examples and slowly update the parameters θ and b in the direction of minimizing each of the small objective $\left(h(x^{(i)};\theta,b) - y^{(i)}\right)^2$. Concretely, we can update the parameters in the following manner:

$$\theta_1 = \theta_1 - \alpha\Delta\theta_1 \quad (3)$$

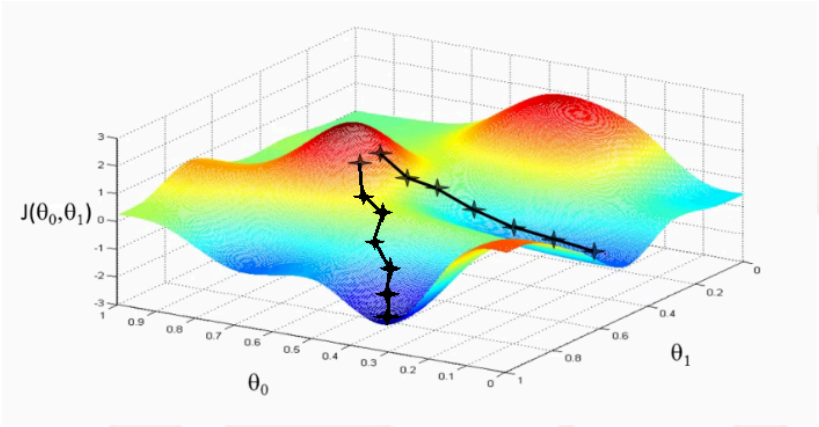
$$\theta_2 = \theta_2 - \alpha\Delta\theta_2 \quad (4)$$

$$b = b - \alpha\Delta b \quad (5)$$

Iterate

Now the question of finding the optimal parameters amounts to finding $\Delta\theta$'s and Δb such that they are in the descent direction. In the following, as our objective function is composed of function of functions, we use the *the chain rule* to compute the derivatives. Remember that the chain rule says that if g is a function of $z(x)$ then its derivative is as follows:

$$\frac{\partial g}{\partial x} = \frac{\partial g}{\partial z}\frac{\partial z}{\partial x}$$



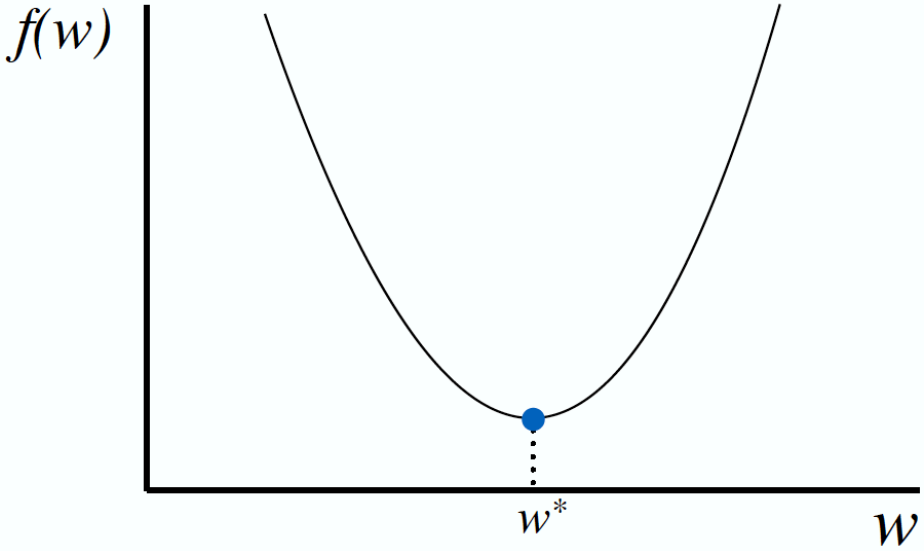
Gráficamente y más simple

MBIT School
Madrid Business Intelligence Technology

Goal: Find $\mathbf{w}^*$ that minimizes

$$f(\mathbf{w}) = ||\mathbf{X}\mathbf{w} - \mathbf{y}||_2^2$$

- Closed form solution exists
- Gradient Descent is iterative (Intuition: go downhill!)

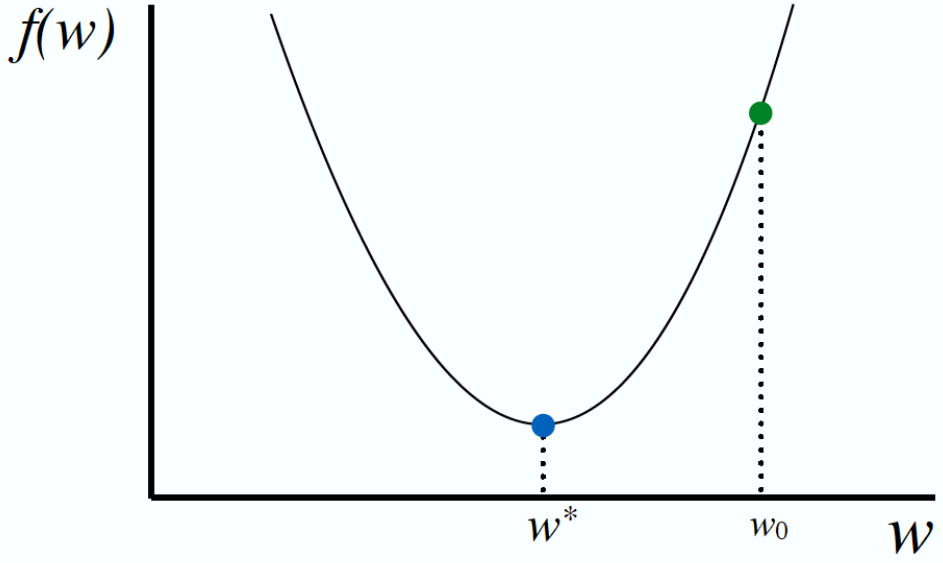


Scalar objective: $f(w) = ||w\mathbf{x} - \mathbf{y}||_2^2 = \sum_{j=1}^{n} (wx^{(j)} - y^{(j)})^2$

Paso 1

MBIT School
Madrid Business Intelligence Technology

Start at a random point

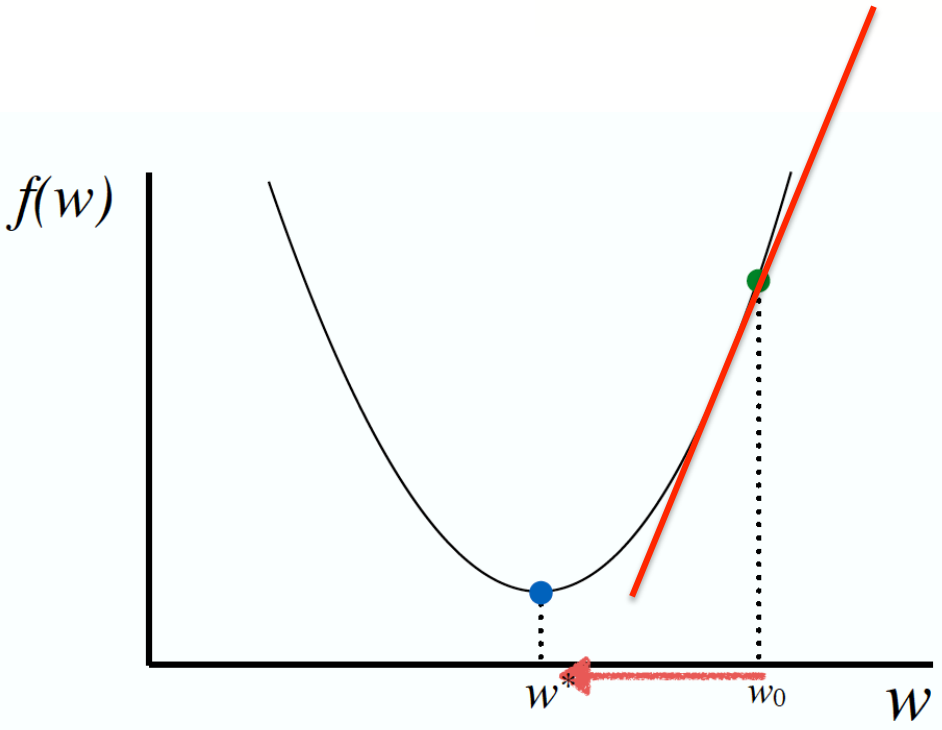


Paso 2

MBIT School
Madrid Business Intelligence Technology

Start at a random point

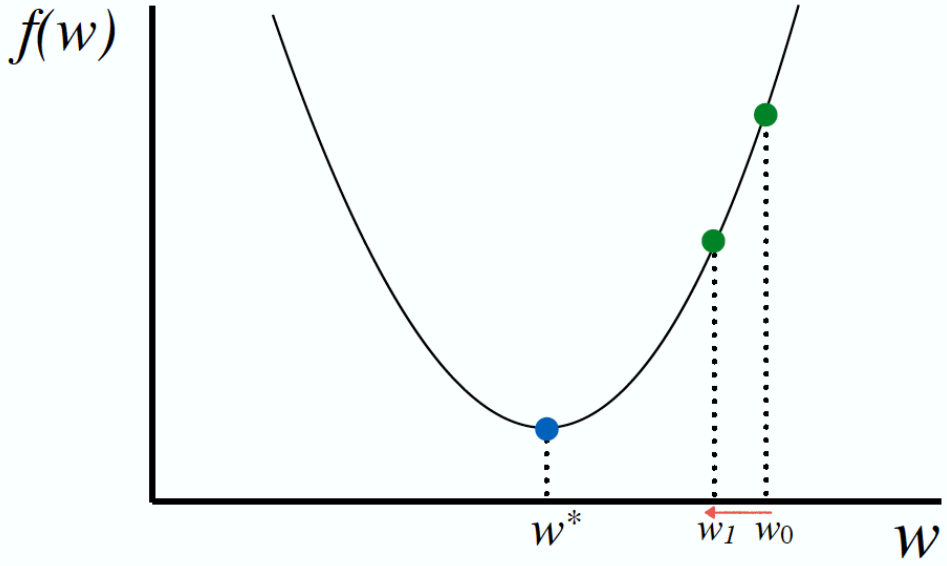
Determine a descent direction



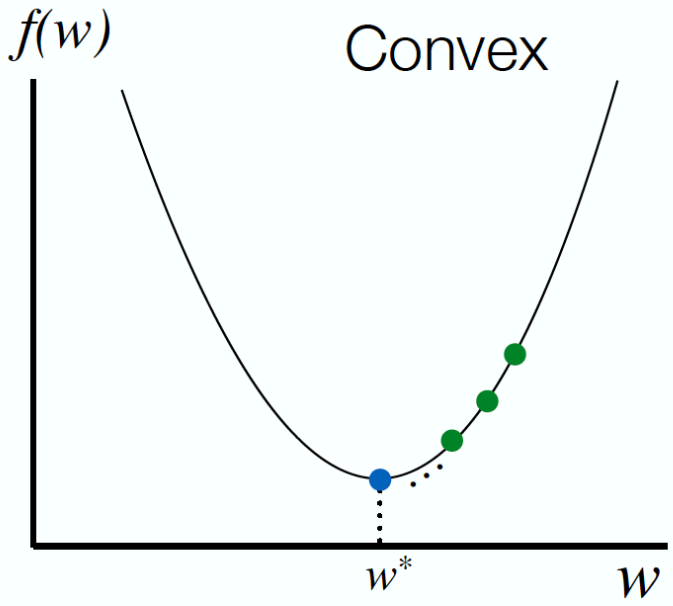
Paso 3

MBIT School
Madrid Business Intelligence Technology

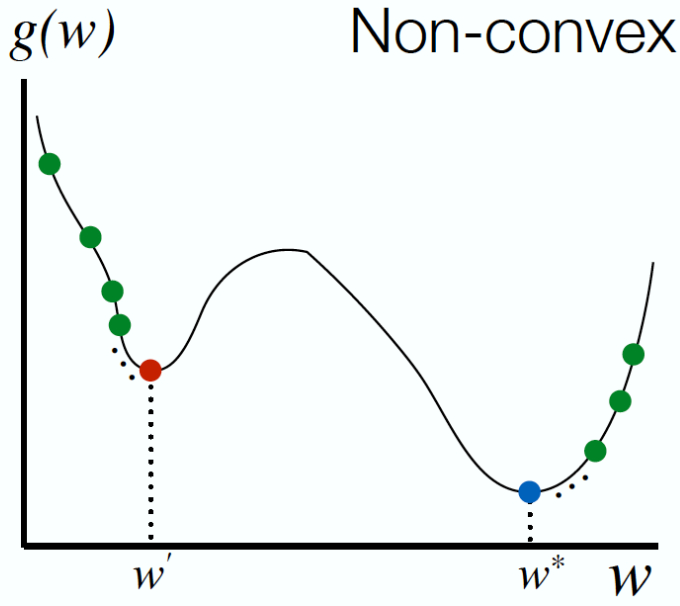
Start at a random point
Repeat
    Determine a descent direction
    Choose a step size
    Update
Until stopping criterion is satisfied



[Nota] Convergencia



Any local minimum is a global minimum



Multiple local minima may exist

Least Squares, Ridge Regression and Logistic Regression are all convex!

Un cuello de botella típico

MBIT School
Madrid Business Intelligence Technology

- Independientemente del algoritmo que emplees, habrás visto MATRICES, MATRICES, MATRICES
  - Tipos de datos matriciales
  - Productos de matrices
  - Matrices inversas

Multiplying a matrix by its inverse results in the identity matrix

- For an $n \times n$ matrix $\mathbf{A}$, $\mathbf{A}^{-1}$ denotes its inverse (when it exists)
- $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$

Only a square matrix (when $n = m$) can have an inverse

Repaso

- ¿Cómo se calcula la matriz inversa?

$$A^{-1} = \frac{1}{|A|} \begin{bmatrix} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} & \begin{vmatrix} a_{13} & a_{12} \\ a_{33} & a_{32} \end{vmatrix} & \begin{vmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{vmatrix} \\ \begin{vmatrix} a_{23} & a_{21} \\ a_{33} & a_{31} \end{vmatrix} & \begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix} & \begin{vmatrix} a_{13} & a_{11} \\ a_{23} & a_{21} \end{vmatrix} \\ \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix} & \begin{vmatrix} a_{12} & a_{11} \\ a_{32} & a_{31} \end{vmatrix} & \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \end{bmatrix}.$$

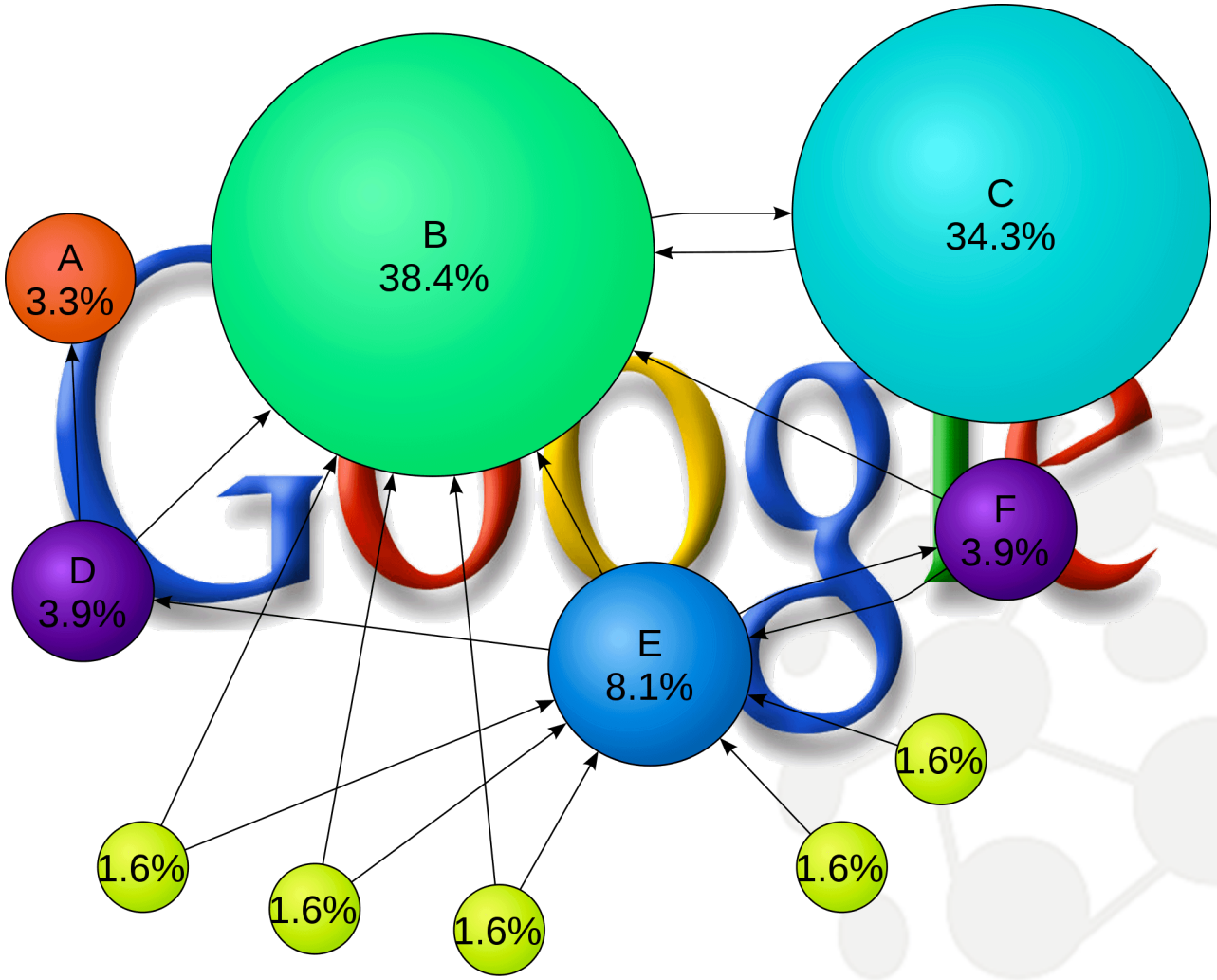
¿Dónde aparecen matrices?

MBIT School
Madrid Business Intelligence Technology

- En cualquier sitio, como una regresión

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^m \\ 1 & x_2 & x_2^2 & \dots & x_2^m \\ 1 & x_3 & x_3^2 & \dots & x_3^m \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^m \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

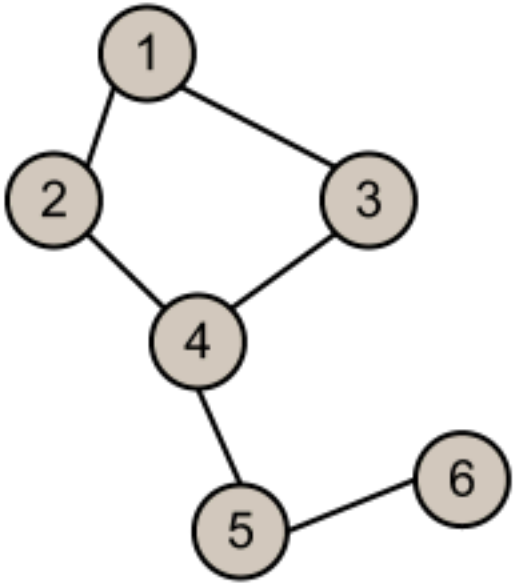
Super-Matrices



Y matrices de adyacencia

MBIT School
Madrid Business Intelligence Technology

Undirected Graph & Adjacency Matrix



Undirected Graph

	1	2	3	4	5	6
1	0	1	1	0	0	0
2	1	0	0	1	0	0
3	1	0	0	1	0	0
4	0	1	1	0	1	0
5	0	0	0	1	0	1
6	0	0	0	0	1	0

Adjacency Matrix

¿A cuánto de qué viene este rollo?

MBIT School
Madrid Business Intelligence Technology

- Tenemos muchos algoritmos que implican un gran número de cálculos e iteraciones
- Empleando un conjunto de datos muy grandes
- Con abstracción matemática en forma matricial
- Y ciertas operaciones con matrices no son nada triviales dependiendo del tamaño de la matrix

Midiendo la complejidad

MBIT School
Madrid Business Intelligence Technology

Describes how algorithms respond to changes in input size

- Both in terms of processing time and space requirements
- We refer to complexity and Big O notation synonymously

Notation: $f(x) = O(g(x))$

- Can describe an algorithm's time or space complexity

Informal definition: f does not grow faster than g

Formal definition: $|f(x)| \leq C|g(x)| \quad \forall x > N$

Complejidades conocidas

MBIT School
Madrid Business Intelligence Technology

Matrix inversion of an $n \times n$ matrix

- $O(n^3)$ time complexity to perform inversion
- $O(n^2)$ space complexity to store result

Y ahora sí, la solución…

MBIT School
Madrid Business Intelligence Technology

Pero no es oro todo lo que reluce

Esto es lo que hay (feb-2017)

- Data types
- Basic statistics
  - summary statistics
  - correlations
  - stratified sampling
  - hypothesis testing
  - streaming significance testing
  - random data generation
- Classification and regression
  - linear models (SVMs, logistic regression, linear regression)
  - naive Bayes
  - decision trees
  - ensembles of trees (Random Forests and Gradient-Boosted Trees)
  - isotonic regression
- Collaborative filtering
  - alternating least squares (ALS)
- Clustering
  - k-means
  - Gaussian mixture
  - power iteration clustering (PIC)
  - latent Dirichlet allocation (LDA)
  - bisecting k-means
  - streaming k-means
- Dimensionality reduction
  - singular value decomposition (SVD)
  - principal component analysis (PCA)
- Feature extraction and transformation
- Frequent pattern mining
  - FP-growth
  - association rules
  - PrefixSpan
- Evaluation metrics
- PMML model export
- Optimization (developer)
  - stochastic gradient descent
  - limited-memory BFGS (L-BFGS)

Lo importante no es la herramienta

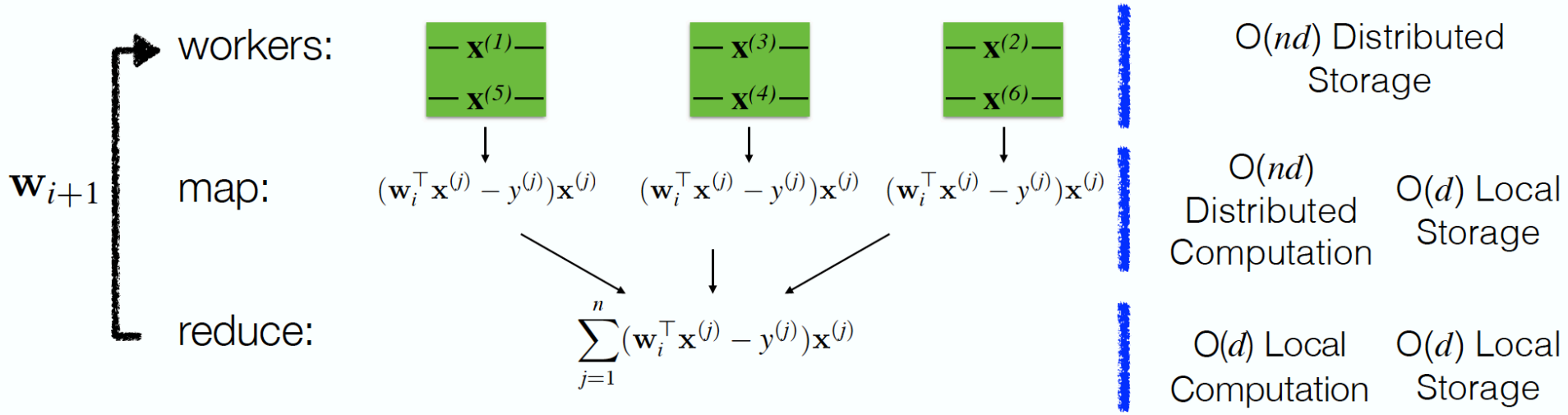
MBIT School
Madrid Business Intelligence Technology

Parallel Gradient Descent for Least Squares

Vector Update: $\mathbf{w}_{i+1} = \mathbf{w}_i - \alpha_i \sum_{j=1}^{n} (\mathbf{w}_i^\top \mathbf{x}^{(j)} - y^{(j)}) \mathbf{x}^{(j)}$

Compute summands in parallel!
note: workers must all have $\mathbf{w}_i$

Example: n = 6; 3 workers



Tampoco es tan difícil

MBIT School
Madrid Business Intelligence Technology

```
for i in range(numIters):
    alpha_i = alpha / (n * np.sqrt(i+1))
    gradient = train.map(lambda lp: gradientSummand(w, lp))
                    .sum()
    w -= alpha_i * gradient
return w
```

A tener en cuenta

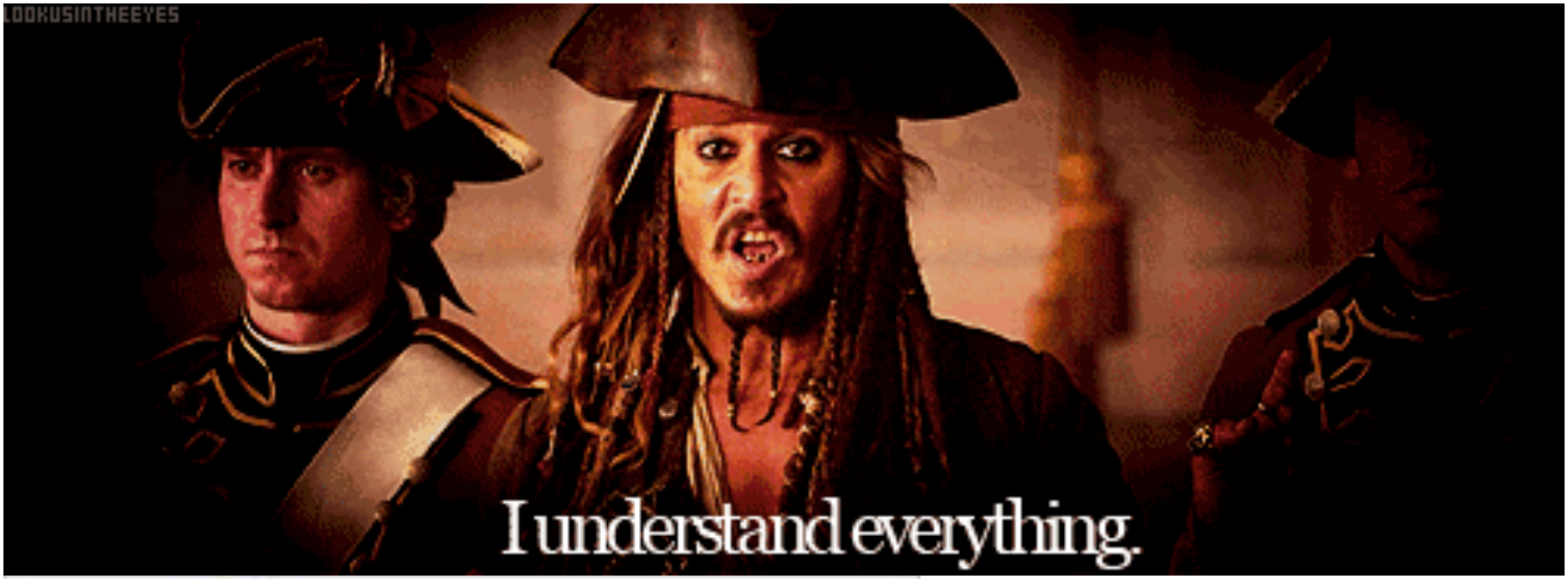
MBIT School
Madrid Business Intelligence Technology

1. Combina cálculo en “memoria” con cálculo paralelo para reducir el coste de comunicaciones
2. Minimiza las comunicaciones
3. Reduce las iteraciones

Resumen

MBIT School
Madrid Business Intelligence Technology

- Casi todos los algoritmos constan de fases interativas
- Además, suelen necesitar cálculo matricial
- Todo ello hace que, con volúmenes de datos medianamente grandes, tengamos problemas de procesamiento
- Pero Spark MLlib está aquí para ayudarnos
- La programación en paralelo no es cuestión trivial y debe considerarse ciertas precauciones


I understand everything.


im so sick of not being understood!
what do you mean?


"HAVE I MADE MYSELF UNDERSTOOD?"
May