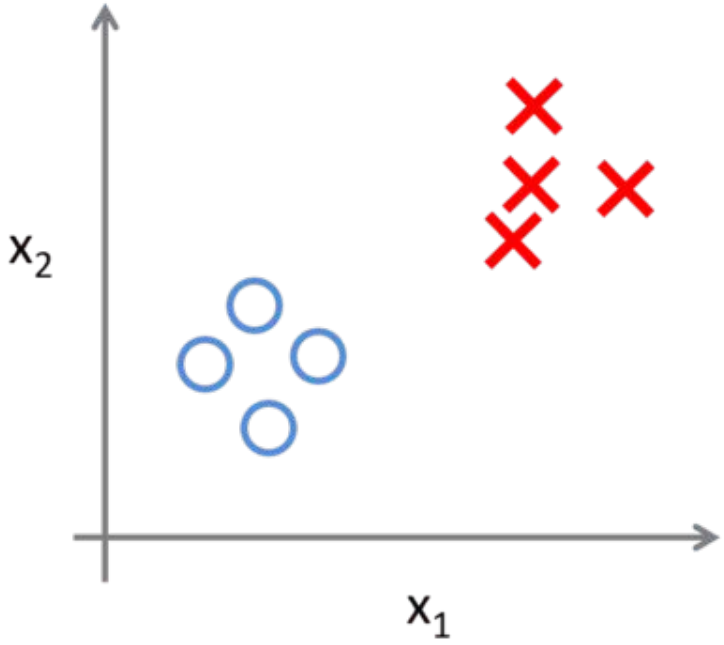
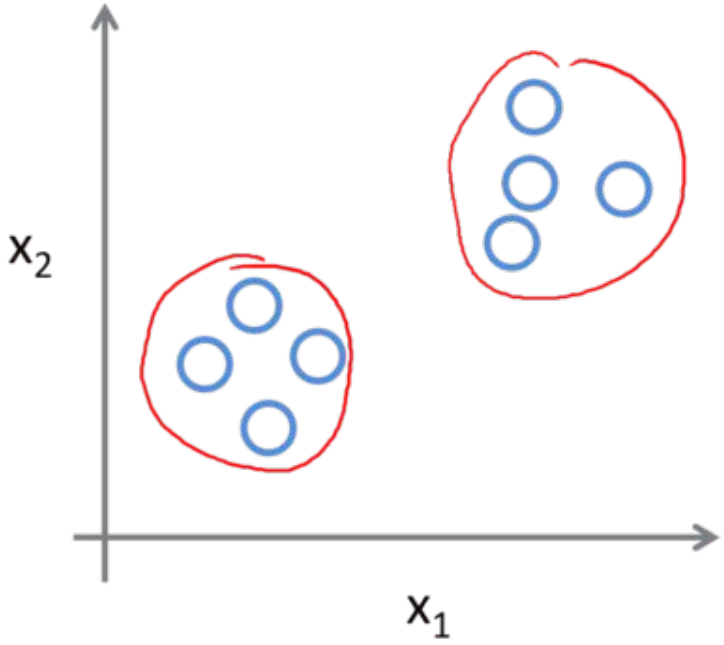
SEMANA 8
UNSUPERVISED LEARNING

- Clustering
  > K-means Algorithm
- Reducción de Dimensionalidad
  > PCA Algorithm

Supervised Learning



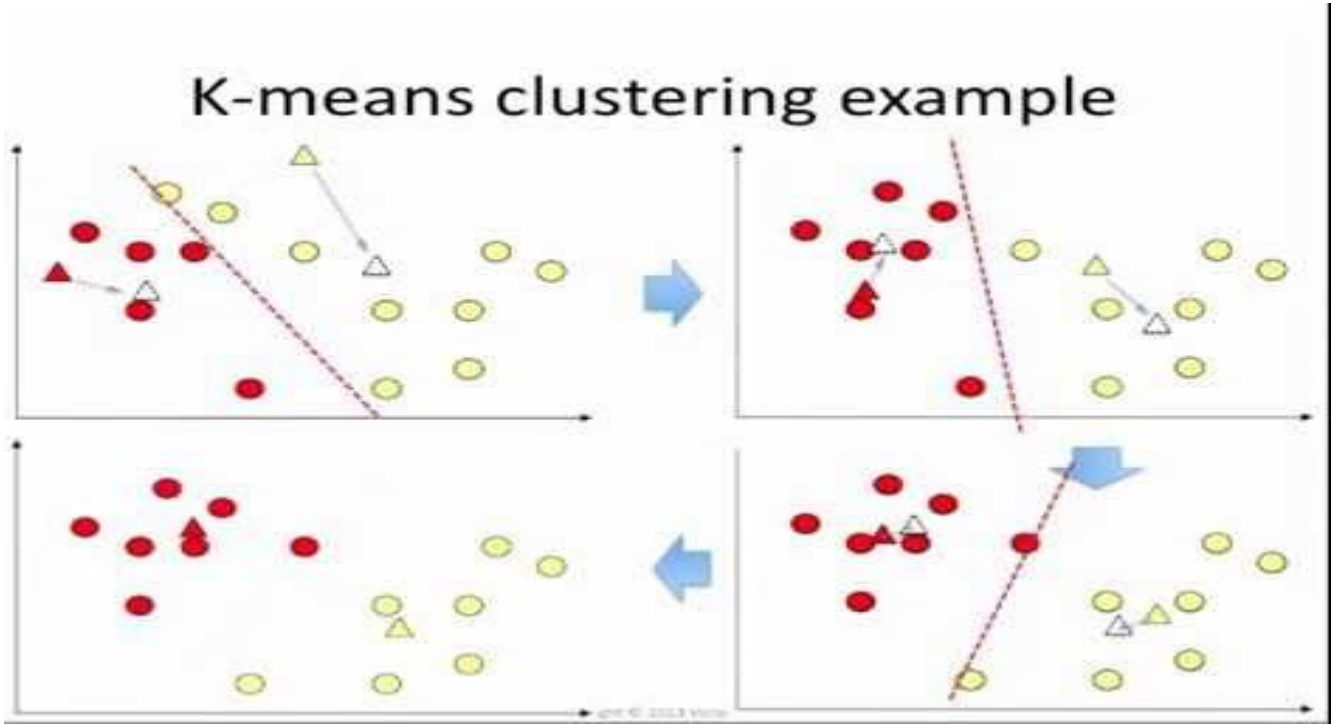
Unsupervised Learning



K-means Algorithm

Tiene 2 fases:

- F. Asignación
- F. Movimiento



Objetivo De Optimización

K-means optimization objective

- $c^{(i)}$ = index of cluster (1,2,...,K) to which example $x^{(i)}$ is currently assigned

μ_k = cluster centroid k ($\mu_k \in \mathbb{R}^n$)

$\mu_{c^{(i)}}$ = cluster centroid of cluster to which example $x^{(i)}$ has been assigned

K $\quad k \in \{1,2,\dots,K\}$

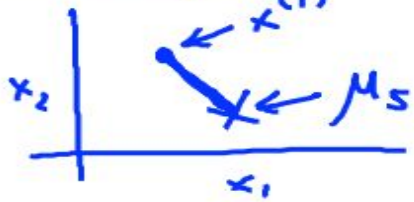
$x^{(i)} \rightarrow 5$ $\quad c^{(i)} = 5$ $\quad \mu_{c^{(i)}} = \mu_5$

Optimization objective:

$$J(c^{(1)},\dots,c^{(m)},\mu_1,\dots,\mu_K) = \frac{1}{m}\sum_{i=1}^{m} ||x^{(i)} - \mu_{c^{(i)}}||^2$$

$$\min_{c^{(1)},\dots,c^{(m)},\ \mu_1,\dots,\mu_K} J(c^{(1)},\dots,c^{(m)},\mu_1,\dots,\mu_K)$$

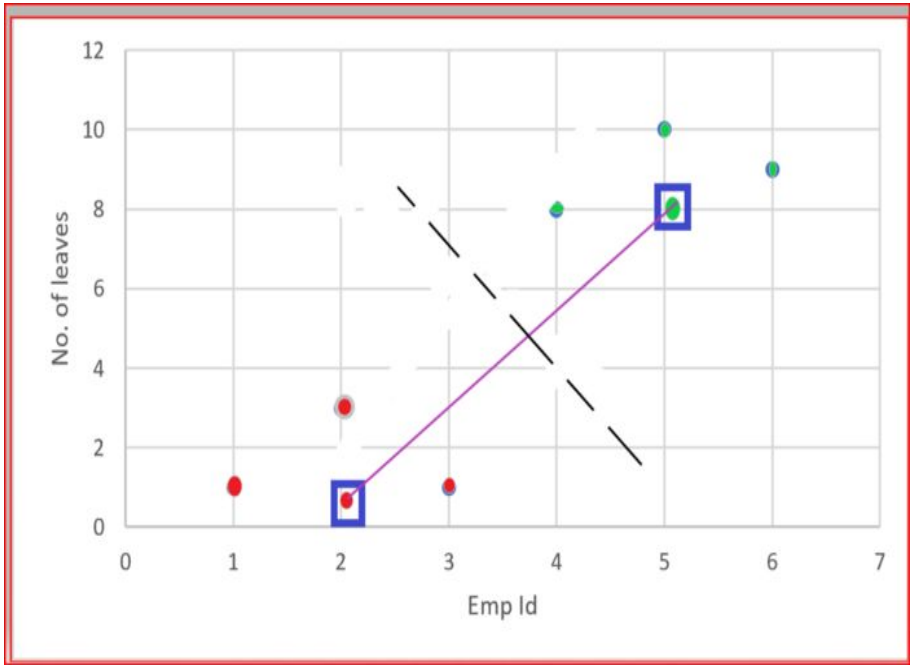
Distortion



Random Initialization

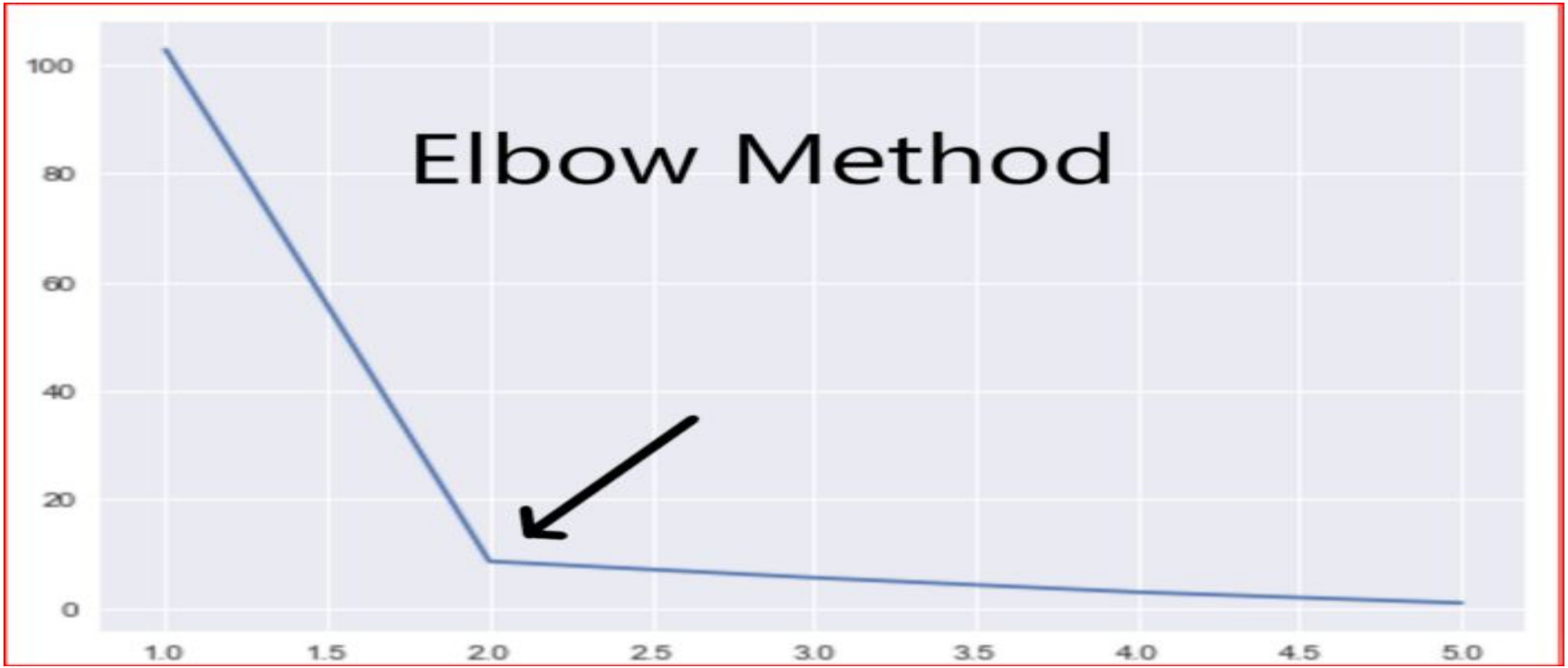
Paso 1: elegimos no. de grupos K

Paso 2: elegimos aleatoriamente centroides



Paso 3: Realizar las 2 fases del algoritmo explicados anteriormente.

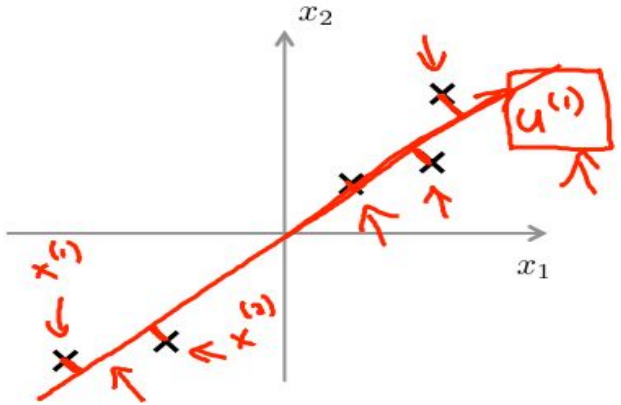
¿Cómo elegir el numero de clusters?



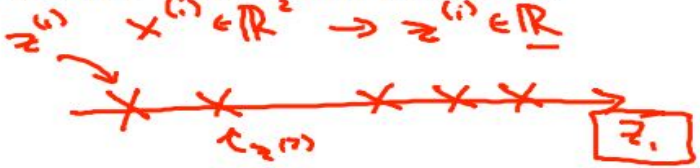
PCA Algorithm

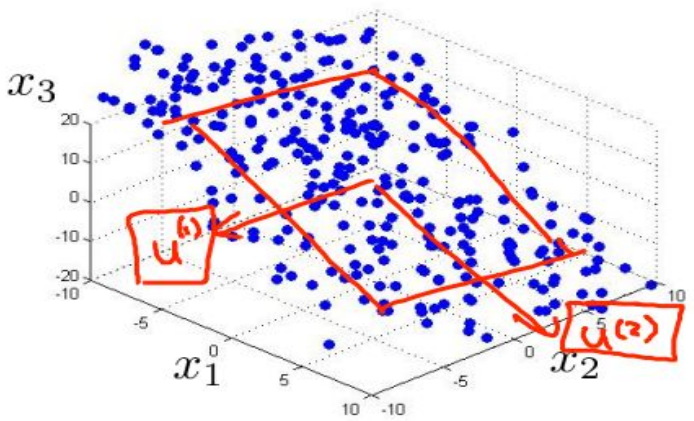
- Se obtiene rapidez en la ejecución del algoritmo.
- Los datos se pueden visualizar más fácilmente

Principal Component Analysis (PCA) algorithm

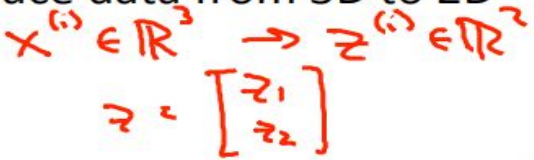


Reduce data from 2D to 1D



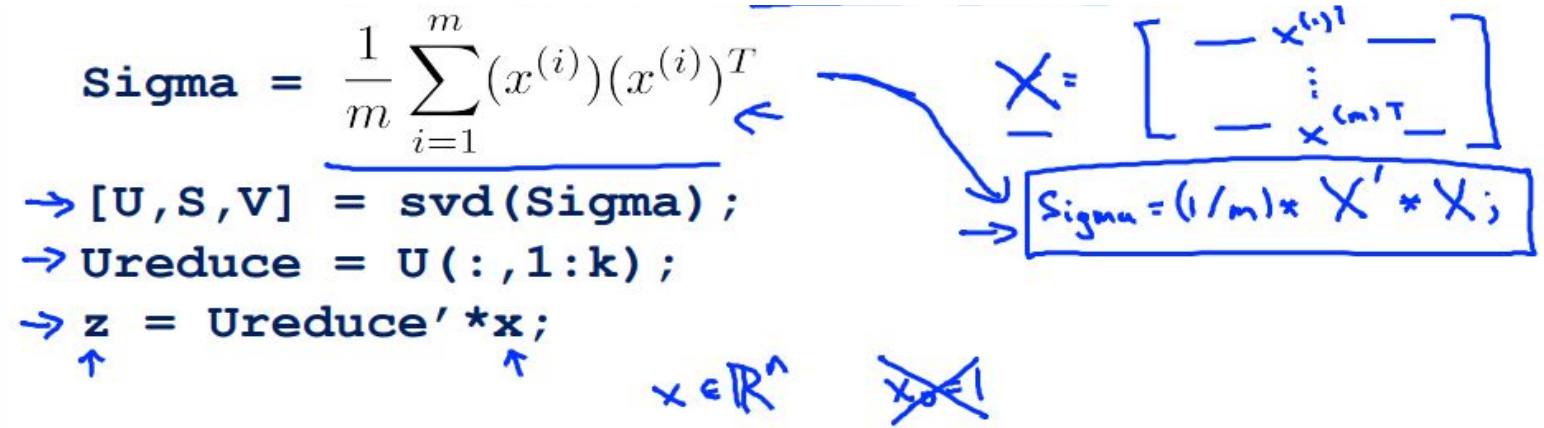


Reduce data from 3D to 2D

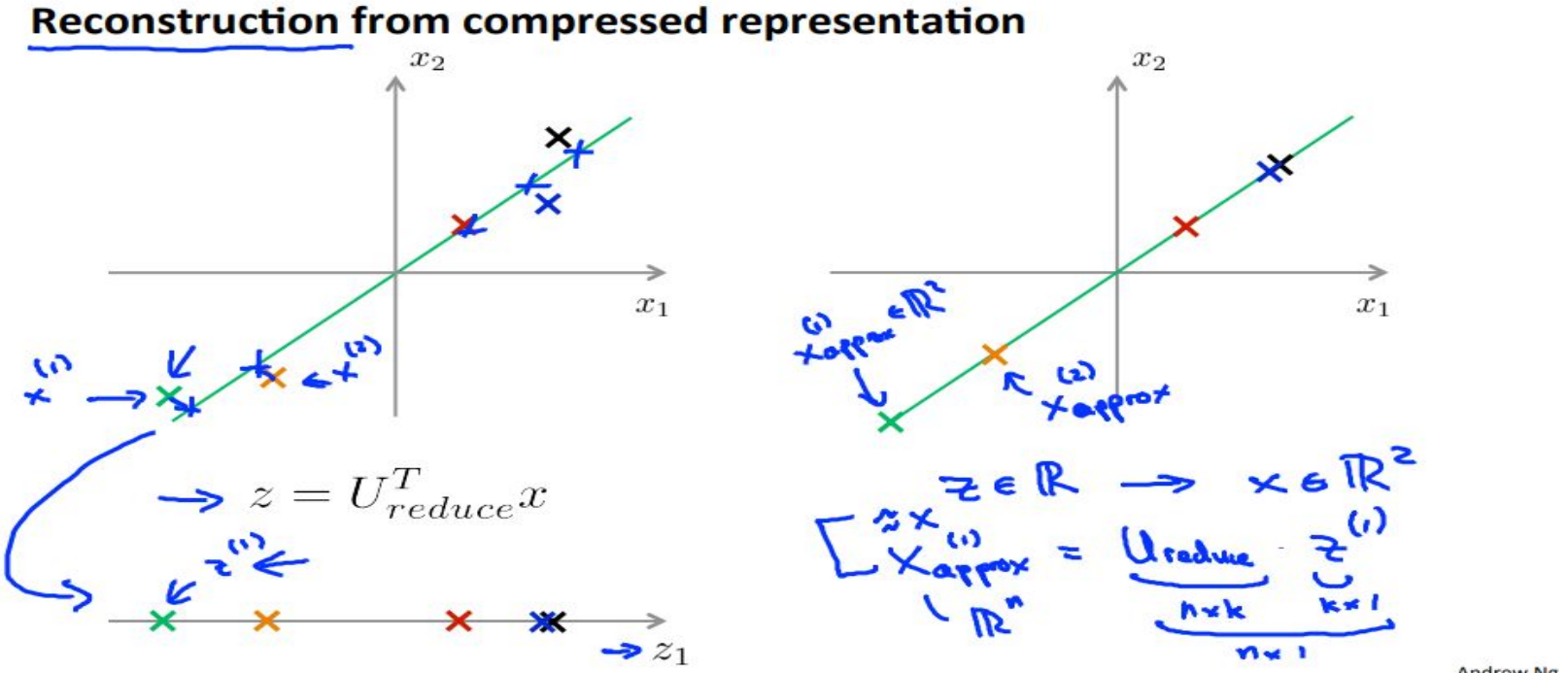


APLICACIÓN DE PCA

1. Etapa de Pre-Procesamiento de datos.
2. Calculamos la Matriz de Covarianzas.
3. Con la función **svd** hallamos la matriz **U**
4. De este matriz **U** tomamos nuestras primeras vectores K .
5. Finalmente hallamos nuestra matriz reducida de variables Z.



Reconstrucción a partir de la Representación Comprimida



Elegir el número de componentes principales

```
[U,S,V] = svd(Sigma)
```

Pick smallest value of k for which

$$\frac{\sum_{i=1}^{k} S_{ii}}{\sum_{i=1}^{m} S_{ii}} \geq 0.99$$

(99% of variance retained)

k=100

TIPS

- Debe definirse ejecutando PCA solo en el set de entrenamiento.
- No es recomendado usar PCA para prevenir Overfitting.

PCA is not linear regression