

# Hoja de Referencia VIP: Aprendizaje no Supervisado

Afshine AMIDI y Shervine AMIDI

6 de octubre de 2018

Traducido por Jaime Noel Alvarez Luna. Revisado por Alonso Melgar López y Fernando Diaz.

## Introducción al Aprendizaje no Supervisado

□ **Motivación** – El objetivo del aprendizaje no supervisado es encontrar patrones ocultos en datos no etiquetados  $\{x^{(1)}, \dots, x^{(m)}\}$ .

□ **Desigualdad de Jensen** – Sea  $f$  una función convexa y  $X$  una variable aleatoria. Tenemos la siguiente desigualdad:

$$E[f(X)] \geq f(E[X])$$

## Expectativa-Maximización

□ **Variables latentes** – Las variables latentes son variables ocultas/no observadas que dificultan los problemas de estimación y a menudo son denotadas como  $z$ . Estos son los ajustes más comunes en los que hay variables latentes:

Ajustes	Variance latente $z$	$x z$	Comentarios
Mezcla de $k$ gaussianos	Multinomial( $\phi$ )	$\mathcal{N}(\mu_j, \Sigma_j)$	$\mu_j \in \mathbb{R}^n, \phi \in \mathbb{R}^k$
Análisis factorial	$\mathcal{N}(0, I)$	$\mathcal{N}(\mu + \Lambda z, \psi)$	$\mu_j \in \mathbb{R}^n$

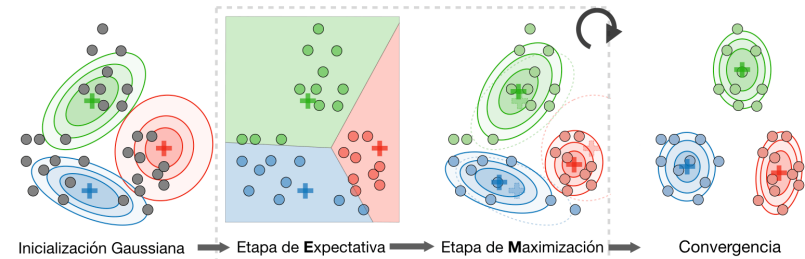
□ **Algoritmo** – El algoritmo Expectativa-Maximización (EM) proporciona un método eficiente para estimar el parámetro  $\theta$  a través de la estimación por máxima verosimilitud construyendo repetidamente un límite inferior en la probabilidad (E-step) y optimizando ese límite inferior (M-step) de la siguiente manera:

- **E-step:** Evalúa la probabilidad posterior  $Q_i(z^{(i)})$  de que cada punto de datos  $x^{(i)}$  provenga de un determinado clúster  $z^{(i)}$  de la siguiente manera:

$$Q_i(z^{(i)}) = P(z^{(i)}|x^{(i)}; \theta)$$

- **M-step:** Usa las probabilidades posteriores  $Q_i(z^{(i)})$  como pesos específicos del clúster en los puntos de datos  $x^{(i)}$  para re-estimar por separado cada modelo de clúster de la siguiente manera:

$$\theta_i = \operatorname{argmax}_{\theta} \sum_i \int_{z^{(i)}} Q_i(z^{(i)}) \log \left( \frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \right) dz^{(i)}$$

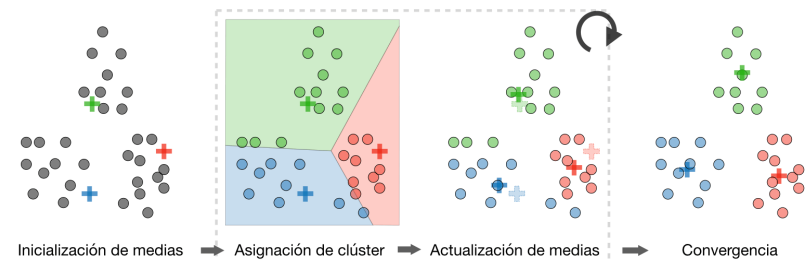


## Agrupamiento $k$ -means

Denotamos  $c^{(i)}$  al clúster de puntos de datos  $i$ , y  $\mu_j$  al centro del clúster  $j$ .

□ **Algoritmo** – Después de haber iniciado aleatoriamente los centroides del clúster  $\mu_1, \mu_2, \dots, \mu_k \in \mathbb{R}^n$ , el algoritmo  $k$ -means repite el siguiente paso hasta la convergencia:

$$c^{(i)} = \operatorname{argmin}_j \|x^{(i)} - \mu_j\|^2 \quad \text{y} \quad \mu_j = \frac{\sum_{i=1}^m 1_{\{c^{(i)}=j\}} x^{(i)}}{\sum_{i=1}^m 1_{\{c^{(i)}=j\}}}$$



□ **Función de distorsión** – Para ver si el algoritmo converge, observamos la función de distorsión definida de la siguiente manera:

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

## Agrupación jerárquica

□ **Algoritmo** – Es un algoritmo de agrupamiento con un enfoque de aglomeramiento jerárquico que construye clústeres anidados de forma sucesiva.

□ **Tipos** – Hay diferentes tipos de algoritmos de agrupamiento jerárquico que tienen por objetivo optimizar diferentes funciones objetivo, que se resumen en la tabla a continuación:

Enlace de Ward	Enlace promedio	Enlace completo
Minimizar dentro de la distancia del clúster	Minimizar la distancia promedio entre pares de clúster	Minimizar la distancia máxima entre pares de clúster

## Métricas de evaluación de agrupamiento

En un entorno de aprendizaje no supervisado, a menudo es difícil evaluar el rendimiento de un modelo ya que no contamos con las etiquetas verdaderas, como en el caso del aprendizaje supervisado.

□ **Coefficiente de silueta** – Sea  $a$  y  $b$  la distancia media entre una muestra y todos los demás puntos en la misma clase, y entre una muestra y todos los demás puntos en el siguiente grupo más cercano, el coeficiente de silueta  $s$  para una muestra individual se define de la siguiente manera:

$$s = \frac{b - a}{\max(a, b)}$$

□ **Índice de Calinski-Harabaz** – Sea  $k$  el número de conglomerados,  $B_k$  y  $W_k$  las matrices de dispersión entre y dentro de la agrupación, respectivamente definidas como:

$$B_k = \sum_{j=1}^k n_{c(i)} (\mu_{c(i)} - \mu)(\mu_{c(i)} - \mu)^T, \quad W_k = \sum_{i=1}^m (x^{(i)} - \mu_{c(i)})(x^{(i)} - \mu_{c(i)})^T$$

el índice de Calinski-Harabaz  $s(k)$  indica qué tan bien un modelo de agrupamiento define sus grupos, de tal manera que cuanto mayor sea la puntuación, más denso y bien separados estarán los conglomerados. Se define de la siguiente manera:

$$s(k) = \frac{\text{Tr}(B_k)}{\text{Tr}(W_k)} \times \frac{N - k}{k - 1}$$

## Análisis de componentes principales

Análisis de componentes principales (en inglés, *Principal Component Analysis*) es una técnica de reducción de la dimensionalidad que encuentra la varianza maximizando las direcciones sobre las cuales se proyectan los datos.

□ **Autovalor, Autovector** – Dada una matriz  $A \in \mathbb{R}^{n \times n}$ , se dice que  $\lambda$  es un autovalor (en inglés, *Eigenvalue*) de  $A$  si existe un vector  $z \in \mathbb{R}^n \setminus \{0\}$ , llamado autovector (en inglés, *Eigenvector*), de tal manera que tenemos:

$$Az = \lambda z$$

□ **Teorema espectral** – Sea  $A \in \mathbb{R}^{n \times n}$ . Si  $A$  es simétrica, entonces  $A$  es diagonalizable a través de una matriz ortogonal real  $U \in \mathbb{R}^{n \times n}$ . Al observar  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ , tenemos:

$$\exists \Lambda \text{ diagonal, } A = U \Lambda U^T$$

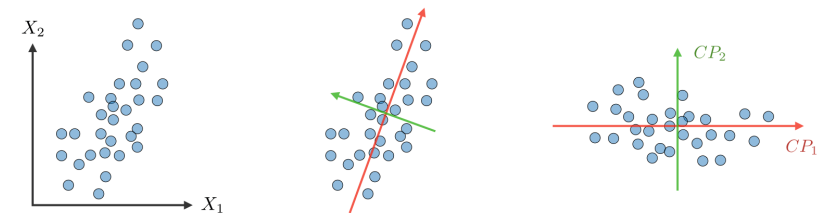
*Observación: el autovector asociado con el autovalor más grande se denomina autovector principal de la matriz  $A$ .*

□ **Algoritmo** – El procedimiento de Análisis de Componentes Principales (ACP) es una técnica de reducción de la dimensionalidad que proyecta los datos en  $k$  dimensiones maximizando la varianza de los datos de la siguiente manera:

- **Paso 1:** Normalizar los datos para obtener una media de 0 y una desviación estándar de 1.

$$x_j^{(i)} \leftarrow \frac{x_j^{(i)} - \mu_j}{\sigma_j} \quad \text{donde} \quad \mu_j = \frac{1}{m} \sum_{i=1}^m x_j^{(i)} \quad \text{y} \quad \sigma_j^2 = \frac{1}{m} \sum_{i=1}^m (x_j^{(i)} - \mu_j)^2$$

- **Paso 2:** Calcular  $\Sigma = \frac{1}{m} \sum_{i=1}^m x^{(i)} x^{(i)T} \in \mathbb{R}^{n \times n}$ , que es simétrico con autovalores reales.
- **Paso 3:** Calcular  $u_1, \dots, u_k \in \mathbb{R}^n$  los  $k$  autovectores ortogonales principales de  $\Sigma$ , es decir, los autovectores ortogonales de los  $k$  mayores autovalores.
- **Paso 4:** Proyectar los datos en  $\text{span}_{\mathbb{R}}(u_1, \dots, u_k)$ . Este procedimiento maximiza la varianza entre todos los espacios  $k$ -dimensionales.



Datos en el espacio de funciones  $\rightarrow$  Buscar componentes principales  $\rightarrow$  Datos en el espacio de CP

## Análisis de componentes independientes

Es una técnica destinada a encontrar las fuentes generadoras subyacentes.

□ **Suposiciones** – Suponemos que nuestros datos  $x$  han sido generados por el vector fuente  $n$ -dimensional  $s = (s_1, \dots, s_n)$ , donde  $s_i$  son variables aleatorias independientes; a través de una matriz  $A$  de mezcla y no singular, de la siguiente manera:

$$x = As$$

El objetivo es encontrar la matriz separadora  $W = A^{-1}$ .

□ **Algoritmo ICA de Bell y Sejnowski** – Este algoritmo encuentra la matriz separadora  $W$  siguiendo los siguientes pasos:

- Escribir la probabilidad de  $x = As = W^{-1}s$  como:

$$p(x) = \prod_{i=1}^n p_s(w_i^T x) \cdot |W|$$

- Escriba la probabilidad dado nuestros datos de entrenamiento  $\{x^{(i)}, i \in \llbracket 1, m \rrbracket\}$  y denotando  $g$ , la función sigmoide, como:

$$l(W) = \sum_{i=1}^m \left( \sum_{j=1}^n \log \left( g'(w_j^T x^{(i)}) \right) + \log |W| \right)$$

Por lo tanto, la regla de aprendizaje de ascenso de gradiente estocástica es tal que para cada ejemplo de entrenamiento  $x^{(i)}$ , actualizamos  $W$  de la siguiente manera:

$$W \leftarrow W + \alpha \left( \begin{pmatrix} 1 - 2g(w_1^T x^{(i)}) \\ 1 - 2g(w_2^T x^{(i)}) \\ \vdots \\ 1 - 2g(w_n^T x^{(i)}) \end{pmatrix} x^{(i)T} + (W^T)^{-1} \right)$$