EST-297

Fundamentos de Ciencia de Datos

Juan Zamora O.

Junio, 2024.

PONTIFICIA
UNIVERSIDAD
CATÓLICA DE
VALPARAÍSO

Estructura de la Presentación

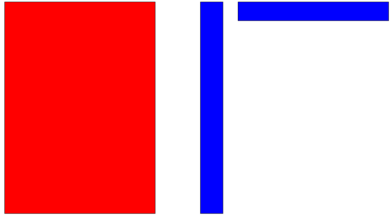
1. Aproximaciones Low-Rank para Clustering
2. (NMF) Factorización de matrices no-negativas
3. NMF
4. Aplicación de NMF para extracción de tópicos en texto

Aproximaciones Low-Rank para Clustering

Una matriz X de rango r admite una factorización de la forma

$$X = BC^T,\ B \in \mathbf{R}^{m \times r},\ C \in \mathbf{R}^{n \times r}$$

X es aproximada con bajo rango (low-rank) cuando $rango(X) << \texttt{min}(m, n)$



(NMF) Factorización de matrices no-negativas

- Grupo de algoritmos de de análisis multivariado y algebra lineal donde una matriz X es factorizada en dos matrices W y H
- Cada columna de X es aproximada por una combinación lineal no-negativa de las columnas de W, donde los coeficientes de mezcla corresponden a las columnas de H
- Las tres matrices tienen elementos no-negativos
- Usado en sistemas recomendadores, procesamiento de audio, agrupamiento de texto.

NMF

- Dada una matriz no-negativa $X \in \mathbf{R}^{m \times n}$ y un $k \in \mathbf{Z} << \min(m, n)$
- Encuentra matrices no-negativas $W \in \mathbf{R}^{m \times k}$ y $H \in \mathbf{R}^{k \times n}$ tales que minimizan

$$\|X - WH\|_F^2 = \sum_i \sum_j (X_{ij} - [WH]_{ij})^2$$

*W: base para un espacio k-dimensional, la i-ésima columna de H: corresponde a representación k-dim de i-ésima columna de X

Método de Lee y Seung (2001)

Lee y Seung propusieron reglas de actualización multiplicativas para minimizar:

$$\min_{W,H \geq 0} \|V - WH\|_F^2$$

Las reglas de actualización son:

$$H \leftarrow H \circ \frac{W^\top V}{W^\top W H}$$
$$W \leftarrow W \circ \frac{V H^\top}{W H H^\top}$$

donde $\circ$ denota el producto elemento a elemento (Hadamard).

Ejemplo numérico: datos iniciales

Matriz original y objetivo

Matriz original V:

$$V = \begin{bmatrix} 5 & 3 \\ 3 & 2 \\ 4 & 1 \end{bmatrix}$$

Factorizar $V \approx WH$, con $W \in \mathbb{R}^{3\times 2}_{\geq 0}$, $H \in \mathbb{R}^{2\times 2}_{\geq 0}$.

Inicialización

Matrices iniciales:

$$W^{(0)} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \\ 1 & 1 \end{bmatrix} \quad , \quad H^{(0)} = \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix}$$

Iteración 1: Actualización de H

$$H_{ij} \leftarrow H_{ij} \times \frac{(W^T V)_{ij}}{(W^T W H)_{ij}}$$

$$W^T V = \begin{bmatrix} 9.5 & 4.0 \\ 9.5 & 4.5 \end{bmatrix} \quad , \quad W^T W H^{(0)} = \begin{bmatrix} 4.25 & 6.5 \\ 4.25 & 6.25 \end{bmatrix}$$

$$H^{(1)} = \begin{bmatrix} 2.235 & 1.231 \\ 2.235 & 0.72 \end{bmatrix}$$

Iteración 1: Actualización de W

$$W_{ij} \leftarrow W_{ij} \times \frac{(VH^T)_{ij}}{(WHH^T)_{ij}}$$

$$VH^{(1)T} = \begin{bmatrix} 14.88 & 13.365 \\ 9.21 & 8.085 \\ 10.171 & 9.16 \end{bmatrix} \quad , \quad W^{(0)}H^{(1)}H^{(1)T} = \begin{bmatrix} 9.54 & 8.49 \\ 8.06 & 8.4 \\ 12.6 & 11.26 \end{bmatrix}$$

$$W^{(1)} = \begin{bmatrix} 1.56 & 0.79 \\ 0.57 & 0.96 \\ 0.81 & 0.81 \end{bmatrix}$$

Iteración 2: Actualización de H

$$W^{(1)T}V = \begin{bmatrix} 13.89 & 6.75 \\ 11.37 & 5.04 \end{bmatrix} \quad , \quad W^{(1)T}W^{(1)}H^{(1)} = \begin{bmatrix} 12.98 & 5.76 \\ 9.94 & 4.21 \end{bmatrix}$$

$$H^{(2)} = \begin{bmatrix} 2.39 & 1.44 \\ 2.56 & 0.86 \end{bmatrix}$$

Iteración 2: Actualización de W

$$VH^{(2)T} = \begin{bmatrix} 20.35 & 17.45 \\ 12.87 & 11.12 \\ 11.98 & 10.17 \end{bmatrix} \quad , \quad W^{(1)}H^{(2)}H^{(2)T} = \begin{bmatrix} 18.1 & 15.7 \\ 14.9 & 13.2 \\ 15.9 & 14.2 \end{bmatrix}$$

$$W^{(2)} = \begin{bmatrix} 1.75 & 0.88 \\ 0.49 & 0.81 \\ 0.61 & 0.58 \end{bmatrix}$$

Iteración 3: Actualización de H

$$W^{(2)T}V = \begin{bmatrix} 14.6 & 7.1 \\ 11.4 & 5.2 \end{bmatrix} \quad , \quad W^{(2)T}W^{(2)}H^{(2)} = \begin{bmatrix} 13.9 & 6.1 \\ 10.7 & 4.7 \end{bmatrix}$$

$$H^{(3)} = \begin{bmatrix} 2.51 & 1.68 \\ 2.73 & 0.95 \end{bmatrix}$$

Iteración 3: Actualización de W

$$VH^{(3)T} = \begin{bmatrix} 21.2 & 18.1 \\ 13.2 & 11.5 \\ 12.1 & 10.5 \end{bmatrix} \quad , \quad W^{(2)}H^{(3)}H^{(3)T} = \begin{bmatrix} 19.3 & 16.8 \\ 15.8 & 14.0 \\ 16.8 & 15.0 \end{bmatrix}$$

$$W^{(3)} = \begin{bmatrix} 1.92 & 0.95 \\ 0.41 & 0.67 \\ 0.44 & 0.41 \end{bmatrix}$$

Iteración 4: Actualización de H y resultado final

$$W^{(3)T}V = \begin{bmatrix} 15.1 & 7.3 \\ 11.2 & 5.0 \end{bmatrix} \quad , \quad W^{(3)T}W^{(3)}H^{(3)} = \begin{bmatrix} 14.6 & 6.4 \\ 10.9 & 4.7 \end{bmatrix}$$

$$H^{(4)} = \begin{bmatrix} 2.60 & 1.92 \\ 2.80 & 1.01 \end{bmatrix}$$

Resultado final aproximado:

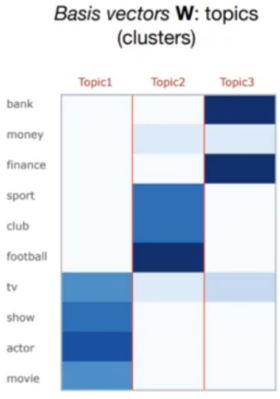
$$W^{(4)} \approx \begin{bmatrix} 1.92 & 0.95 \\ 0.41 & 0.67 \\ 0.44 & 0.41 \end{bmatrix} \quad , \quad H^{(4)} \approx \begin{bmatrix} 2.60 & 1.92 \\ 2.80 & 1.01 \end{bmatrix}$$

y

$$W^{(4)} H^{(4)} \approx V$$

Aplicación de NMF para extracción de tópicos en texto

- Se construye matriz de terminos vs documentos
- Se aplica NMF para obtener W y H



Clustering y Modelos estadísticos de Texto

- Abundante en diversos dominios (redes sociales, medios digitales, registros en salud …)
- Resulta útil poder explorar estas coleciones de alguna manera asistida
- Clustering permite caracterizar de manera *automática* una colección de documentos
- A finales de los 90, aparecieron varios modelos estadístico de texto usando un modelo de mezcla sobre variables aleatorias multinomiales
  - LSI
  - pLSI

¿Qué es LDA?

- LDA aparece a principio del 2000
- Incluye un modelo generativo para los documentos, además de los tópicos
- Cada documento es una mezcla de temas
- Cada tema es una distribución de palabras
- Distribución apriori de tópicos es una Dirichlet

Referencias: Blei et al. 2003

Distribuciones de probabilidad en LDA

- $\theta_d \sim$ Dirichlet(α): distribución de temas en un documento.
- $\phi_k \sim$ Dirichlet(β): distribución de palabras en un tema.
- $z_{d,n} \sim$ Multinomial(θ_d): elección de tema para palabra$n$en documento$d$.
- $w_{d,n} \sim$ Multinomial$(\phi_{z_{d,n}})$: elección de palabra según el tema.

Estimación de parámetros

- El modelo observa solo las palabras. Los temas son variables latentes.
- Se busca inferir:
  - θ_d: proporción de temas en cada documento.
  - ϕ_k: distribución de palabras por tema.
  - $z_{d,n}$: asignación de temas a palabras.
- Métodos comunes:
  - Muestreo de Gibbs (Gibbs Sampling)
  - Inferencia variacional (Variational Bayes)

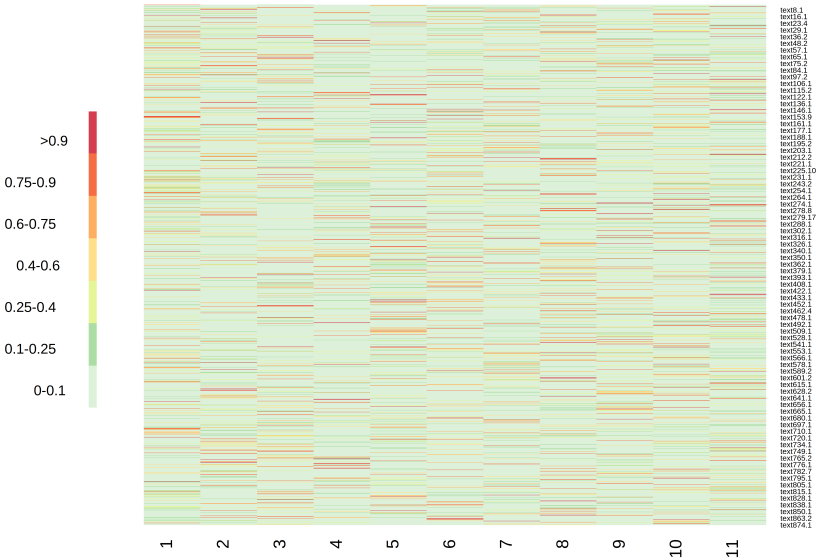
¿Qué es Gibbs Sampling?

- Método de Monte Carlo para estimar distribuciones condicionales.
- En LDA:
  - Se fija el tema de todas las palabras excepto una.
  - Se estima la probabilidad condicional de cada posible tema para esa palabra.
  - Se repite este proceso para todas las palabras, muchas veces.
- El resultado converge a una estimación de la distribución posterior conjunta.

Ejemplo de las palabras más representativas en 11 tópicos

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11
<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>
internacionalización	región	desarrollo	importante	prestigio	problemas	mejoras	tiempo	inclusiva	formación	académicos
nacional	valparaíso	medio	áreas	futuro	sociales	universidades	profesores	innovadora	profesional	comunidad
investigación	institución	vinculación	vanguardia	educación	desarrollo	aporte	profesional	estudiantes	calidad	estudiantes
región	comunidad	investigación	personas	estudiantes	temas	país	cambio	aprendizaje	nuevas_tecnologías	personas
liderazgo	entorno	estudiantes	estudios	programas	excelente	calidad	mejoras	abierta	procesos	funcionarios
reconocida	compromiso	sostenible	compleja	calidad	institución	desarrollo	mundo	personas	valores	conocimientos
referente	local	innovadora	espacios	carreras	resolver	conocimientos	tiempos	desarrollar	continua	oportunidad
proyectos	tradición	ambiente	nuevas_tecnologías	chile	público	infraestructura	puedan	investigación	institución	preocupada
áreas	ciudad	permitan	carreras	mejoras	país	enseñanza	gestión	más_inclusiva	trabajo	espacios
manteniendo	nacional	institución	territorio	mundo	comprometida	tres	investigación	calidad	estudiantes	servicio

Asociación texto y tópico



¿De qué sirve esta perspectiva generadora de documentos?

- Existen técnicas estadísticas y computacionales para invertir este procedimiento a partir de documentos existentes (…nuestros documentos), pudiendo así inferir la composición *más probable* de los tópicos que permitieron generar esta colección de documentos.
- Los tópicos estimados tienen un significado identificado por el/la analista
- Para encontrar la cantidad de tópicos se utiliza una medida denominada *Perplexity*
  - Se calcula tomando la log-verosimilitud de los documentos con los tópicos resultantes
  - Que tanto es posible reproducir la composición de los documentos dados los tópicos
  - El bjetivo es escoger el número de tópicos que minimiza la Perplexity