

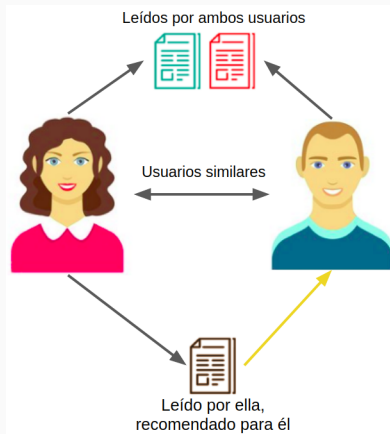
UNIDAD 4: SISTEMAS DE RECOMENDACIÓN

MODELADO DE FILTRADO COLABORATIVO Y DESCOMPOSICIÓN DE MATRICES

Blanca Vázquez y Gibran Fuentes-Pineda

Noviembre 2020

MODELO DE FILTRADO COLABORATIVO



La similitud entre usuarios puede ser usada para inferir las calificaciones no observadas.

- Se basan en emplear las calificaciones obtenidas por los usuarios para la generación de las recomendaciones.
- El principal reto de este modelo es la dispersión de las matrices (valores observados y no observados).

	Usuario 1	Usuario 2	Usuario n
P_1	4	0	0	1
P_2	0	3	0	0
...	0	1	0	0
P_n	2	0	0	3

Para el proceso de recomendación, la mayoría de los modelos de filtrado colaborativo aprovechan las similitudes:

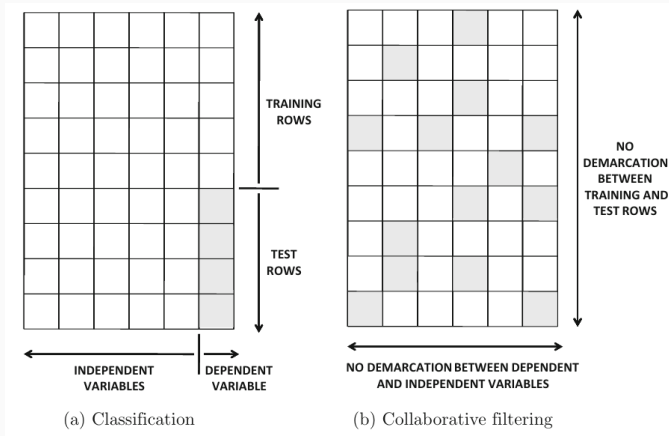
- Entre usuarios, entre productos o ambos.

Para medir la similitud se usa:

- Distancia de Jaccard
- Distancia coseno

PROBLEMA DE CLASIFICACIÓN

El filtrado colaborativo puede ser visto como una generalización del problema de clasificación y regresión.



Comparando el problema de clasificación tradicional con el filtrado colaborativo. Imagen tomada de Aggarwal, 2016.

PROBLEMA DE CLASIFICACIÓN

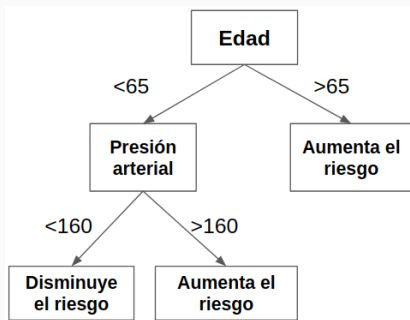
Clasificación	Filtrado
-Separación entre variables independientes y dependiente (clase).	- Cada columna puede ser independiente o dependiente (dependerá sobre lo que se quiere predecir)
- Clara separación entre datos de de entrenamiento y de prueba.	- En el mejor de los casos, los valores observados (serán los de entrenamiento) y los no observados (serán los de prueba).
- Las columnas representan características y las filas instancias.	- Debido a la naturaleza de los datos, se generan los modelos basados en productos o en usuarios.

Debido a la generalidad de los modelos de filtrado colaborativo, diferentes modelos de aprendizaje de máquinas son usados en la generación de recomendaciones:

- Métodos basados en ensambles (árboles de decisión y regresión).
- Naive Bayes
- Redes neuronales
- Modelos de factor latente

ÁRBOLES DE DECISIÓN Y REGRESIÓN

Son una partición jerárquica del espacio de datos a partir de un conjunto de criterios de decisión en las variables independientes.



La selección de usar árboles de decisión o de regresión dependerá de la variable dependiente:

- Categórica: árboles de decisión
- Numérica (continua): árboles de regresión

En los árboles de decisión, el criterio de decisión puede ser evaluado usando el promedio pesado del índice de *Gini*.

$$G(S) = 1 - \sum_{i=1}^r p_i^2$$

dónde: $p_1...p_r$ son las fracciones de los datos pertenecientes a r diferentes clases en un nodo S .

- El índice de *Gini* es entre 0 y 1.
- Los valores más pequeños son indicativos de un mayor poder discriminativo.

Los principales desafíos en la extensión de los árboles de decisión en el filtrado colaborativo son:

- Los datos no están separados en forma de variables independientes y clase.
- La matriz de puntuaciones es mayormente dispersa y con datos ausentes.
- Dado que no es claro, las variables independientes y dependientes, ¿qué variable debe predecir el árbol?

¿Qué variable debe predecir el árbol?

- Se construyen árboles de decisión separados para predecir el *rating* por cada producto.
- Supongamos, una matriz de $m \times n$ (m usuarios y n productos), el *rating* del producto que se quiera predecir será la variable **dependiente** y el resto **independiente**.
- El número de árboles a construir será igual al número de productos de la matriz.

Sin embargo, uno de los principales retos de los árboles de decisión en el filtrado colaborativo es: **la ausencia de datos.**

Para atender este problema:

- Se usan métodos de reducción de dimensionalidad
- Posteriormente, se calcula la matriz de covarianza usando los vectores propios.
- La representación reducida se usa para la construcción de los árboles, como si fuera un problema de clasificación.

FILTRADO COLABORATIVO BASADO EN REGLAS

En este tipo de filtrado, se aborda la generación de recomendaciones basado en **reglas de asociación**.

Item \Rightarrow	<i>Bread</i>	<i>Butter</i>	<i>Milk</i>	<i>Fish</i>	<i>Beef</i>	<i>Ham</i>
Customer \Downarrow						
Jack	1	1	1	0	0	0
Mary	0	1	1	0	1	0
Jane	1	1	0	0	0	0
Sayani	1	1	1	1	1	1
John	0	0	0	1	0	1
Tom	0	0	0	1	1	1
Peter	0	1	0	1	1	0

$$\{Butter, Milk\} \Rightarrow \{Bread\}$$

Imagen tomada de Aggarwal, 2016.

Aprovechando las reglas de asociación en el filtrado colaborativo

Cuando el número de productos es pequeño, cada valor de la combinación producto - calificación puede ser tratado como un **pseudo-producto**.

Ejemplo:

- {Producto = leche, calificación = 'me gusta'} \Rightarrow {Producto = 'pan', calificación = 'me gusta'}

Las reglas de asociación son construidas en términos de pseudo-productos.

Ejemplo de reglas basadas en **pseudo-productos**:

- {Producto = mantequilla, calificación = 'me gusta'} AND {Producto = leche, calificación = 'me gusta'} \Rightarrow {Producto = 'pan', calificación = 'me gusta'}

Ejemplo de reglas basadas en **pseudo-usuarios**:

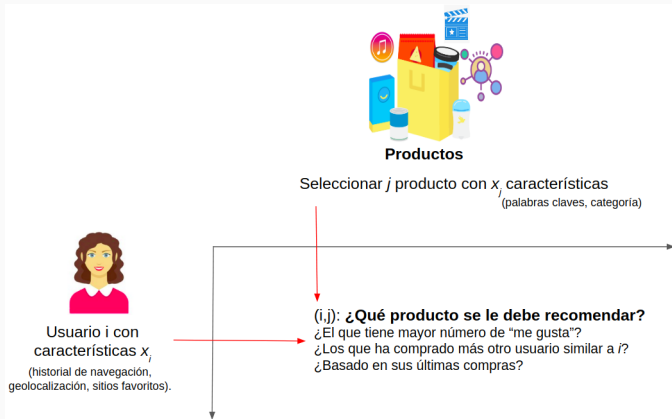
- {Usuario = Pepe, calificación = 'me gusta'} \Rightarrow {Usuario = 'Paco', calificación = 'no me gusta'} AND {Usuario = 'Lola', calificación = 'me gusta'}

Recordemos:

- El objetivo de los sistemas de recomendación es **proporcionar un conjunto de sugerencias de productos a un usuario final.**
- Para generar las recomendaciones, frecuentemente se usa la **similitud** entre productos, usuarios o ambos.

DEFINICIÓN DEL PROBLEMA

Recordemos:



- Enfoque basado en características (contenido)
 - Usa las características para generar una recomendación
 - Debilidad: no puede distinguir entre productos / usuarios que tienen vectores de la misma característica.
- Enfoque de filtrado colaborativo
 - Hace recomendaciones basadas en interacciones pasadas entre usuarios / productos.
 - Debilidad: no maneja de forma natural nuevos usuarios o nuevos productos.

Son el estado del arte en los sistemas de recomendación

- Usan los métodos de reducción de dimensionalidad para llenar los valores ausentes en la matriz de *ratings*.
- La idea básica es explotar que filas y columnas en una matriz están altamente correlacionadas (matriz de bajo rango).
- Objetivo: completar los valores ausentes usando la matriz de bajo rango.

INTRODUCCIÓN A LOS MODELOS DE FACTOR LATENTE

Intuición de los modelos de factor latente

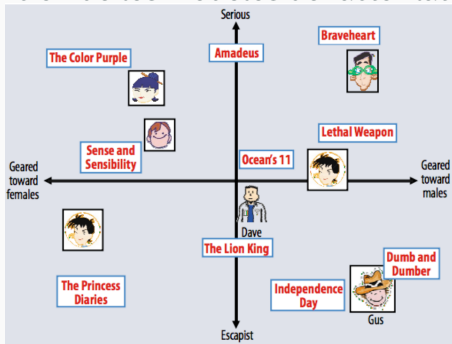


Imagen tomada de Koren, 2009.

Intenta explicar los *ratings* de los usuarios a partir de inferir patrones (factores latentes) de productos y usuarios.

Koren, Y., Bell, R., Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8), 30-37.

¿Cómo identificar patrones de productos y usuarios?

- A partir de calcular los vectores latentes, usando factorización de matrices.

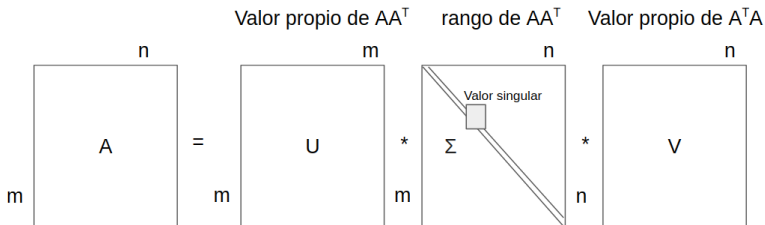
Características de la factorización de matrices:

- Tiene buena escalabilidad
- Se obtienen predicciones más precisas
- Flexibilidad

¿Cómo generar los vectores latentes?

- Usar descomposición de valores singulares (SVD)
- Análisis de componentes principales (PCA)

$$A = U\Sigma V^T$$



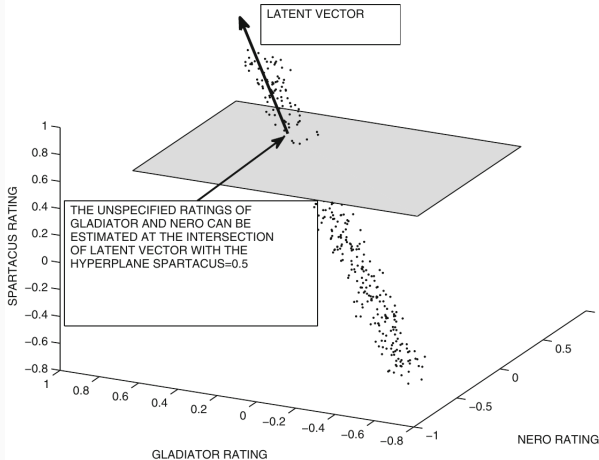
Desventaja de la descomposición de valores singulares

- La SVD generalmente se emplea en matrices sin datos ausentes.
- Enfoques previos, usaban el *rating* promedio de los usuarios y productos para llenar las matrices.

Vectores latentes

- La clave de los modelos de factor latente es la habilidad para estimar los vectores latentes con datos ausentes.
- Recordemos que partimos del supuesto que los elementos en la matriz están altamente correlacionados, por lo tanto, estas redundancias ayudan a reconstruir los valores ausentes.

MODELOS DE FACTOR LATENTE



Aprovechar las redundancias basadas en la correlación para estimar los datos faltantes para un usuario cuya única calificación especificada es un valor de 0.5 para la película Spartacus.

Imagen tomada de Aggarwal, 2016.

Intuición

- La intuición geométrica vista en el ejemplo anterior es útil, cuando los vectores latentes son mutuamente ortogonales.
- Sin embargo, eso no siempre sucede.

Factorización de matrices

- ¿Cómo calcular los vectores latentes cuando no son mutuamente ortogonales?
- La **factorización** es una manera general de aproximación de una matriz cuando es propensa a la reducción de dimensionalidad (correlaciones entre columnas o filas).

Principios básicos

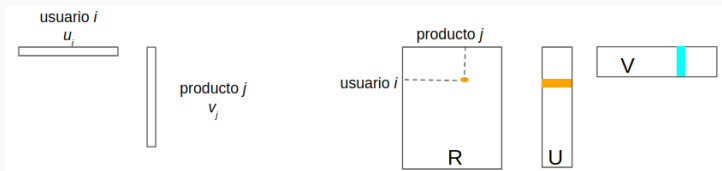
Dada una matriz de *ratings* R de $m \times n$, esta matriz puede ser factorizada en 2 matrices:

- Matriz U : $m \times k$
- Matriz V : $n \times k$

$$R \approx UV^T$$

- Cada columna de U o V se refiere como vector latente o componente latente
- Cada fila de U o V se refiere como factor latente.

Cálculo de los vectores latentes



$$R \approx UV^T$$

donde R_{ij} es la calificación que el usuario i le da al producto j , U es del tamaño $m \times n$, k es el rango, u_i es el vector del usuario i y v_j es el vector del producto j .

$$R \approx UV^T$$

- Cada fila \bar{u}_i en la matriz U se denomina **factor de usuario**.
- Cada fila \bar{v}_j de la matriz V se denomina **factor de producto**.
- Objetivo final: modelar cada producto y usuario como un vector de factores.

Por lo tanto, para predecir el *rating* r_{ij} en R puede ser expresado como:

$$r_{ij} \approx \bar{u}_i \cdot \bar{v}_j$$

FACTORIZACIÓN DE MATRICES

Dado que los factores latentes $\bar{u}_i = (u_{i1}, \dots, u_{ik})$ y $\bar{v}_j = (v_{j1}, \dots, v_{jk})$ pueden verse como las afinidades de los usuarios por k conceptos diferentes, la ecuación:

$$r_{ij} \approx \bar{u}_i \cdot \bar{v}_j$$

Puede ser expresada como:

$$r_{ij} \approx \sum_{s=1}^k u_{is} \cdot v_{js}$$

$= \sum_{s=1}^k$ (afinidad del usuario i para el concepto s) * (afinidad del producto j para el concepto s)

Ejemplo

		NERO	JULIUS CAESAR	CLEOPATRA	SLEEPLESS IN SEATTLE	PRETTY WOMAN	CASABLANCA
HISTORY	1	1	1	1	0	0	0
	2	1	1	1	0	0	0
	3	1	1	1	0	0	0
BOTH	4	1	1	1	1	1	1
ROMANCE	5	-1	-1	-1	1	1	1
	6	-1	-1	1	1	1	1
	7	-1	-1	-1	1	1	1
		R					

Imagen tomada de Aggarwal, 2016.

Ejemplo

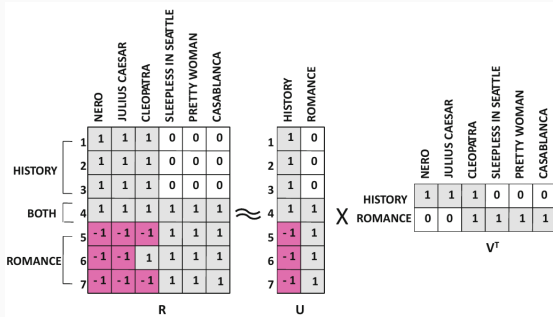
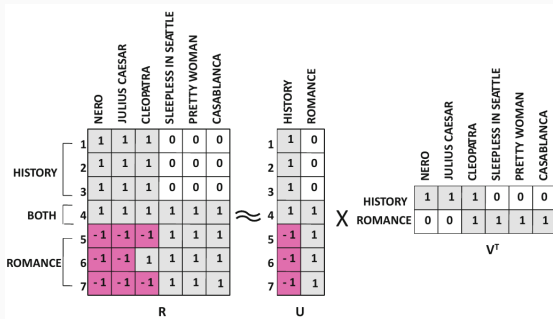


Imagen tomada de Aggarwal, 2016.

FACTORIZACIÓN DE MATRICES

En nuestro ejemplo, los conceptos se refieren a los géneros romántico e histórico.



Por lo tanto, la ecuación quedaría así:

$r_{ij} \approx (\text{afinidad del usuario } i \text{ para historia}) * (\text{afinidad del producto } j \text{ para historia}) + (\text{afinidad del usuario } i \text{ para romance}) * (\text{afinidad del producto } j \text{ para romance}).$

Importante:

- La noción de concepto (ej., romance e historia) frecuentemente no tienen una interpretación semántica.
- Un vector latente comúnmente es más parecido a esto:

$$\begin{bmatrix} 1 \\ -2.5 \\ 3 \\ -0.2 \end{bmatrix}$$

- Lo importante es: un vector latente describe los patrones de correlación en la matriz R .

Principales diferencias entre métodos

Existen muchos métodos de factorización de matrices, las principales diferentes son:

- Las restricciones de las matriz U y V
 - Ortogonalidad o no negatividad de los vectores latentes.
- La naturaleza de la función objetivo
 - Minimizar la norma de Frobenius o maximizar la estimación de probabilidad.

Calculando las matrices U y V

Como se ha mencionado previamente, cualquier matriz R puede descomponerse en 2 matrices (U,V) , tal que UV^T puede aproximar la matriz original de R (sin valores ausentes).

¿Cómo obtener las matrices U y V que aproximen a R?

Se puede formular como un problema de optimización:

$$\text{minimizar } J = \frac{1}{2} \|R - UV^T\|^2$$

Esta función objetivo se puede ver como una **función de pérdida cuadrática**. Asimismo, métodos del descenso del gradiente pueden dar una solución óptima para esta factorización.

¿Cómo obtener las matrices U y V que aproximen a R con datos ausentes?

Recordemos: las matrices de *ratings* se caracterizan tener **¡datos ausentes!**, es decir, solo un conjunto de entradas tiene valores observados.

- Por lo tanto, la función objetivo anterior, no podría ser aplicada
 - No es posible calcular la norma de Frobenius con entradas faltantes.

¿Cómo obtener las matrices U y V que aproximen a R con datos ausentes?

Solución: la función objetivo debe re-escribirse usando solo las entradas observadas para aprender U y V .

- *Nice part*: una vez aprendidos los factores latentes U y V es posible reconstruir la matriz completa usando UV^T .

¿Cómo obtener las matrices U y V que aproximen a R con datos ausentes?

- Supongamos todos los pares observados (i, j) en R los cuales denotaremos como S .
- Sea $i \in (1, \dots, m)$ el índice de un usuario y $j \in (1, \dots, n)$ el índice de un producto.

Por lo tanto, el conjunto S de pares observados (usuario-producto) se define:

$$S = \{(i, j) : r_{ij} \text{ par observado}\}$$

¿Cómo obtener las matrices U y V que aproximen a R con datos ausentes?

Una vez calculadas las matrices U y V usando las entradas observadas, es posible predecir las entradas faltantes:

$$\hat{r}_{ij} = \sum_{s=1}^k u_{is} \cdot v_{js}$$

La diferencias entre las entradas observadas y las entradas predichas de una entrada específica (i, j) es:

$$e_{ij} = (r_{ij} - \hat{r}_{ij}) = r_{ij} - \sum_{s=1}^k u_{is} \cdot v_{js}$$

Actualización de la función objetivo

Dada una matriz R con entradas faltantes, la función objetivo se calcula sobre las entradas observadas:

$$\text{minimizar } J = \frac{1}{2} \sum_{(i,j) \in S} e_{ij}^2 = \frac{1}{2} \sum_{(i,j) \in S} \left(r_{ij} - \sum_{s=1}^k u_{is} \cdot v_{js} \right)^2$$

Nota: la función objetivo suma el error únicamente sobre las entradas observadas.

Actualización de la función objetivo

Cabe señalar que cada uno de los términos:

$$(r_{ij} - \sum_{s=1}^k u_{is} \cdot v_{js})^2$$

Es el error cuadrado e_{ij}^2 entre la entrada observada y la entrada predicha (i, j) .

Recuerda, u_{is} y v_{js} son las variables desconocidas las cuales necesitan ser aprendidas para minimizar la función objetivo.

- El algoritmo del descenso del gradiente es el método más usado para minimizar la función objetivo.
 - Fácil de implementar
 - El tiempo de ejecución es relativamente rápido
- El método de mínimos cuadrados alternos también se ha usado para minimizar la función.
 - Este método se puede paralelizar.