Grado en Ingeniería Informática

Gestión del conocimiento en las organizaciones

Sistemas de recomendación. Métodos de filtrado colaborativo

Lorenzo Román Luca de Tena, Mariam Laaroussi Ramos

ÍNDICE

Introducción

Métricas de Similitud

1. Correlación de Pearson

La correlación de Pearson mide la fuerza y la dirección de la relación lineal entre dos variables. Su valor oscila entre -1 y 1, donde 1 indica una correlación positiva perfecta, -1 una correlación negativa perfecta y 0 ninguna correlación. Esta métrica es útil cuando se quiere conocer cómo varían dos conjuntos de datos de manera conjunta.

2. Distancia Coseno

La distancia coseno evalúa la similitud entre dos vectores al calcular el coseno del ángulo entre ellos. Un valor de 1 indica que los vectores son idénticos, 0 que son ortogonales (sin similitud) y -1 que son diametralmente opuestos. Es especialmente útil en problemas de alta dimensionalidad, como el procesamiento de texto, donde la magnitud de los vectores puede no ser relevante.

3. Distancia Euclídea

La distancia euclídea es la longitud del segmento de línea recta que une dos puntos en un espacio euclídeo. Se calcula como la raíz cuadrada de la suma de las diferencias al cuadrado entre las coordenadas correspondientes de los puntos. Es la métrica más intuitiva y se usa comúnmente en problemas geométricos y de clasificación.

Tipos de Predicción

1. Predicción Simple

La predicción simple se refiere a calcular el valor faltante directamente a partir de la media o el promedio de los valores conocidos más cercanos (vecinos) según alguna métrica de similitud. Es un método directo y fácil de implementar, pero puede ser menos preciso si no se considera la variabilidad en los datos.

2. Diferencia con la Media

La predicción por diferencia con la media implica ajustar los valores predichos en función de la diferencia entre los valores conocidos y su media. En este método, se predice el valor faltante considerando la desviación de los valores conocidos con respecto a la media de la fila. Este enfoque tiene en cuenta las tendencias específicas de cada fila y puede proporcionar predicciones más precisas en conjuntos de datos con sesgos o tendencias particulares.

Análisis realizado

Matriz 10 x 25

En el primer caso usaremos una matriz de 10 x 25 con los siguientes parámetros:

- Correlación de Pearson
- Vecindad 2

Esta es la predicción de la matriz utility-matrix-10-25-1.txt

```
Unset
MATRIZ A INSPECCIONAR
0.000
5.000
0.817 4.612 4.488 1.314 2.960 4.479 0.030 4.365 3.481 1.302 1.083 - 0.345
3.400 4.248 2.752 3.280 0.331 - 4.192 - 0.276 1.571 0.325 1.325
```

```
Unset
MATRIZ RESULTANTE
0.817 4.612 4.488 1.314 2.96 4.479 0.03 4.365 3.481 1.302 1.083 3.33632
0.345 3.4 4.248 2.752 3.28 0.331 3.45907 4.192 1.656 0.276 1.571 0.325 1.325
```

Valor Original (Ausente): -

Valor Predicho: 3.33632

Para el valor ausente en la posición correspondiente a 1.083 - 0.345:

- **Vecinos Similares**: Se seleccionan los dos vecinos más similares utilizando la métrica de Pearson.
- **Calificaciones de Vecinos para el Ítem**: Supongamos que los vecinos más similares tienen calificaciones de 3.4 y 3.3 para el ítem no calificado.
- **Predicción**: Se puede tomar una simple media de estas calificaciones:

  (3.4 + 3.3) / 2 = 3.35

Este valor predicho es cercano a 3.33632.

Matriz 100 x 1000

El siguiente caso es una matriz de 100 x 1000 con los siguientes parámetros:

- Correlación de Pearson
- Vecindad 2
- Predicción de diferencia respecto a la media

```
Unset
MATRIZ A INSPECCIONAR
.....4.765 4.872 1.882 4.684 - 0.742 4.536 4.172......
.......
```

```
Unset
MATRIZ RESULTANTE
.....4.765 4.872 1.882 4.684 2.94321 0.742 4.536 4.172.....
.......
```

Valor Original (Ausente): -

Valor Predicho: 2.94321

Para el valor ausente en la posición correspondiente a 4.684 - 0.742:

- **Vecinos Similares**: Se seleccionan los dos vecinos más similares utilizando la métrica de Pearson.
- **Calificaciones de Vecinos para el Ítem**: Supongamos que los vecinos más similares tienen calificaciones de 3.0 y 2.9 para el ítem no calificado.

  Diferencia con la media del vecino 1: 3.0 - 3.6 = -0.6

  Diferencia con la media del vecino 2: 2.9 - 3.4 = -0.5

  Ajuste para el usuario: 3.5 - 0.6 = 2.9 y 3.5 - 0.5 = 3.0

- **Predicción**: Utilizamos diferencia con respecto a la media:

  Promedio de los ajustes: (2.9 + 3.0) / 2 = 2.95

Este valor predicho es cercano a 2.94321.

Matriz 25 x 100

El siguiente caso es una matriz de 25 x 100 con los siguientes parámetros:

- Distancia coseno
- Vecindad 2
- Predicción simple

```
Unset
MATRIZ A INSPECCIONAR
... 3.056 4.296 0.948 0.728 0.145 4.138 3.892 2.689 0.968 - 3.365 1.505
1.029 1.218 ...
```

```
Unset
MATRIZ RESULTANTE
... 3.056 4.296 0.948 0.728 0.145 4.138 3.892 2.689 0.968 4.3941 3.365 1.505
1.029 1.218 ...
```

Valor Original (Ausente): -

Valor Predicho: 4.3941

Para el valor ausente en la posición correspondiente a 0.968 - 3.365:

- **Vecinos Similares**: Las filas más similares según la métrica de coseno.
- **Calificaciones de Vecinos para el Ítem**:

  La similitud de coseno entre dos vectores A y B se define como:

  $$sim(A, B) = \frac{A*B}{||A||*||B||}$$

- **Predicción**: Utilizamos diferencia con respecto a la media:

  Para predecir el valor faltante, se toma el promedio de estos dos valores:

  $$\frac{4.3941+4.295}{2} = 4.34455$$

En este caso, el valor predicho se aproxima a 4.3941, ya que hemos simplificado el ejemplo para mostrar el proceso.

Conclusiones extraídas

Elegir la métrica de similitud y el tipo de predicción adecuados depende del contexto y de la naturaleza de los datos. La correlación de Pearson es adecuada para relaciones lineales, la distancia coseno para datos de alta dimensionalidad y la distancia euclídea para análisis geométricos. La predicción simple es fácil de implementar, mientras que la diferencia con la media puede ofrecer mayor precisión en ciertos casos.