Ingeniería Industrial
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

Tarea 3

Curso: Web Intelligence IN5526.
Semestre: Primavera 2017.
Fecha: 25/11/2017.

Profesor: Juan D. Velásquez.
Andrés Córdova.
Auxiliar: Valeria Lobos-Ossandón.

1. Contexto

Considere un servicio en línea de *streaming* de películas. Para mantener el interés de los usuarios es necesario ayudarlos a explorar la vasta colección de películas disponibles, de forma tal que se mantengan satisfechos en el proceso. Una forma posible es un sistema automatizado de recomendaciones, que genere recomendaciones personalizadas a cada usuario.

2. Problema a resolver

Se quiere construír un sistema de recomendación de películas a usuarios basada en filtrado colaborativo. Se le entregará una matriz de utilidad de 943 usuarios y 1682 películas. Esta matriz tiene evaluaciones del 1 al 5, y ceros para los casos sin evaluación. Con esto, se pide lo siguiente:

1. Dividir la matriz de utilidad en entrenamiento y test. Para ello, seleccione al azar y sin reemplazo entre los puntos con evaluaciones de la matriz de utilidad y ponga esos puntos en matrices vacías, de forma que se cumpla que

$$Y = Y_{\text{entrenamiento}} + Y_{\text{test}} \tag{1}$$

   Reserve $\frac{3}{4}$ de los datos como juego de entrenamiento y el cuarto restante para test.

2. Implementar y comparar por RMSE los siguientes métodos de recomendación en el juego de validación:

   Promedio Se predice la evaluación de un usuario j a una película i como el promedio entre la evaluación promedio del usuario j a las películas que ha evaluado y la evaluación promedio de todos los usuarios que han evaluado i. Formalmente:

$$\hat{y}(i,j) = \frac{1}{2}\left\{\frac{1}{|U|}\sum_{u\in U} y(i,u) + \frac{1}{|K|}\sum_{k\in K} y(k,j)\right\} \tag{2}$$

   Donde U es el conjunto de usuarios que han evaluado el item i y K es el conjunto de ítem que han sido evaluados por el usuario j.

   Vecinos más cercanos Se predice la evaluación de un usuario j a una película i como:

$$\hat{y}(i,j) = \frac{\sum_{k\in K} y(k,j)\cdot sim(i,k)}{\sum_{k\in K} sim(i,k)} \tag{3}$$

Donde K es el conjunto de ítem más similares al item i. La similitud entre películas está dada por el coeficiente de correlación de Pearson de los vectores de usuarios que han evaluado el item i:

$$sim(i,j) = \frac{\sum_{k \in U}(y(i,k) - \bar{y}(k))(y(j,k) - \bar{y}(k))}{\sqrt{\sum_{k \in U}(y(i,k) - \bar{y}(k))^2}\sqrt{\sum_{k \in U}(y(j,k) - \bar{y}(k))^2}} \quad (4)$$

Donde U es el conjunto de usuarios que $\bar{y}(k)$ es la evaluación promedio del usuario k a todos los ítem que ha evaluado.

Este método está parametrizado por el tamaño de K, de forma que deberá encontrar el mejor $|K|$. Para ello use validación cruzada para evaluar el método con distintos $|K|$ entre 1 y 100 (se recomienda ir de 20 en 20: $|K| \in \{1, 20, 40, 60, 80, 100\}$). Reporte el desempeño obtenido con los distintos $|K|$ para justificar la decisión. Para los métodos implementados, especifique qué es lo que hace el método cuando se quiere hacer una evaluación cuando el ítem no ha sido evaluado previamente o el usuario no ha evaluado ningún ítem previamente.

Recomendación basada en modelo Se predice la evaluación de un usuario j a una película i como:

$$\hat{y}(i,j) = \theta^{(j)} \cdot x^{(i)} \quad (5)$$

donde $\theta^{(j)}$ y $x^{(i)}$ son estimados como las soluciones de

$$\min_{\{\theta^{(j)}\}_{j \in U}, \{x^{(i)}\}_{i \in K}} \frac{1}{2} \sum_{(i,j):r(i,j)=1} (\theta^{(j)} \cdot x^{(i)} - y(i,j))^2 + \frac{\lambda}{2} \sum_{i=1}^{n_I} ||x^{(i)}||^2 + \frac{\lambda}{2} \sum_{j=1}^{n_u} ||\theta^{(j)}||^2 \quad (6)$$

Se le entregará una implementación básica de este método, que viene los hiperparámetros ya definidos. Se recomienda que el método basado en vecinos más cercanos sea implementado siguiendo la misma interfaz que este método. Así la comparación en el juego de validación se volverá más simple y comprensible.

3. Reporte el desempeño del mejor método de recomendación en el juego de test.

4. Se quiere ver cómo mejora el desempeño del método de vecinos más cercanos a medida que se van agregando usuarios al sistema. Para ello, modifique la implementación de los métodos para que puedan trabajar con matrices de utilidad vacías, y permitan agregar evaluaciones al sistema, de forma de hacer lo siguiente:

   - Agregar evaluaciones al sistema, obtenidas del juego de entrenamiento
   - Calcular el RMSE con las evaluaciones del juego de test

   Comience con una matriz de utilidad con un cuarto de las evaluaciones del juego de entrenamiento, comience a agregar evaluaciones, y reporte el RMSE para el juego de test a medida que se agregan evaluaciones al sistema.

3. Herramientas a usar

Para efectuar el trabajo no existe restricción en las herramientas, sin embargo se recomienda el uso de Python para la programación, análisis de datos y generación de contenido gráfico. Para el informe se puede usar cualquier editor de texto considerando que el informe final debe ser entregado en PDF.

4. Entregables

El trabajo a entregar deberá tener lo siguiente:

- Informe que muestre claramente el trabajo realizado. Este informe deberá tener por lo menos las siguientes secciones:
  1. Resumen ejecutivo
  2. Introducción
  3. Desarrollo
  4. Análisis de Resultados
  5. Conclusiones
- Scripts, modelos, etc. que permitan reproducir los resultados del informe.