Aprendizaje automático

Practica 8: Sistemas de recomendación

Universidad de Zaragoza, curso 2022/2023

Juan Eizaguerri Serrano 816079

Introducción y función de coste

En esta práctica se va a implementar un sistema de recomendación de películas basado en las valoraciones de los usuarios. Se cuenta con el dataset MovieLens 100k, que contiene calificaciones de 943 usuarios a 1682 películas. Para implementar el sistema de recomendaciones se utilizarán los siguientes datos en forma de matrices:

- **Y(nUsuarios x nPeliculas):** Ranking proporcionado por cada usuario para cada película.
- **R(nUsuarios x nPeliculas):** Indica si cada película i ha sido calificada por un usuario j.
- **X(nPeliculas x nAtributos):** Características de cada película.
- **Theta(nUsuarios x nAtributos):** Atributos para cada usuario.

En todas las pruebas se utilizarán 100 atributos.

En primer lugar se implementa la función de coste, que estima el error de predicción dadas las X y las Thetas. Para cada usuario sólo se hace este cálculo con las películas que ha calificado.

Se va a utilizar un filtrado colaborativo que trate de minimizar el coste en X y Theta al mismo tiempo, por lo que también es necesario calcular el gradiente del coste respecto a estos dos datos.

Por último, se implementa regularización para penalizar los valores altos de X y Θ.

Entrenamiento y generación de predicciones

El objetivo del entrenamiento es minimizar la función de coste de estas predicciones modificando los valores de X y Θ. Esto se hace utilizando la función fmincg, que realiza un descenso de gradiente aplicando filtrado colaborativo a lo largo de 100 iteraciones.

Una vez acabado el entrenamiento, la predicción de los ratings de los usuarios se puede calcular como $X * \Theta^T$, dando lugar a una matriz de tamaño nPeliculas x nUsuarios.

A modo de ejemplo, se ha creado un nuevo usuario y se le han asignado calificaciones a distintas películas.

```
Original ratings provided:
Rated 4 for Toy Story (1995)
Rated 3 for Twelve Monkeys (1995)
Rated 5 for Usual Suspects, The (1995)
Rated 4 for Outbreak (1995)
Rated 5 for Shawshank Redemption, The (1994)
Rated 3 for While You Were Sleeping (1995)
Rated 5 for Forrest Gump (1994)
```

```
Rated 2 for Silence of the Lambs, The (1991)
Rated 4 for Alien (1979)
Rated 5 for Die Hard 2 (1990)
Rated 5 for Sphere (1998)
```

Después del entrenamiento, el modelo es capaz de predecir una calificación para cada película, que una vez ordenadas en orden descendente, se pueden utilizar como recomendaciones.

Tras 100 iteraciones, usando un factor regularizador $\lambda=10$ se llega a un coste $J=66742.46$, y las siguientes recomendaciones.

```
Top recommendations for you:
Predicting rating 4.0 for movie Star Wars (1977)
Predicting rating 4.0 for movie Forrest Gump (1994)
Predicting rating 4.0 for movie Shawshank Redemption, The (1994)
Predicting rating 3.9 for movie Titanic (1997)
Predicting rating 3.9 for movie Raiders of the Lost Ark (1981)
Predicting rating 3.8 for movie Usual Suspects, The (1995)
Predicting rating 3.8 for movie Schindler's List (1993)
Predicting rating 3.7 for movie Return of the Jedi (1983)
Predicting rating 3.7 for movie Godfather, The (1972)
Predicting rating 3.7 for movie Toy Story (1995)
```

Introducción de nuevos usuarios

Cuando se añade un usuario al sistema de recomendación, al no haber introducido ninguna calificación, no se dispone de información suficiente como para realizar predicciones. Para solucionar este problema se puede utilizar la normalización de la media, que consiste en mostrar las predicciones como la diferencia respecto a la media global.

Por un lado, esto hará que las recomendaciones para los nuevos usuarios sean los de mayor calificación media, y además, reducirá el impacto de los valores altos en el sistema.

Aplicando la normalización de la media al ejemplo del apartado anterior se obtiene el siguiente resultado.

```
Top recommendations for you:
Predicting rating 0.5 for movie Forrest Gump (1994)
Predicting rating 0.5 for movie Die Hard 2 (1990)
Predicting rating 0.4 for movie Saint, The (1997)
Predicting rating 0.4 for movie Die Hard: With a Vengeance (1995)
Predicting rating 0.4 for movie Independence Day (ID4) (1996)
Predicting rating 0.3 for movie Broken Arrow (1996)
Predicting rating 0.3 for movie Return of the Jedi (1983)
Predicting rating 0.3 for movie Spawn (1997)
Predicting rating 0.3 for movie True Lies (1994)
```

```
Predicting rating 0.3 for movie Rock, The (1996)
```

Varias de las películas como Forrest Gump y Return of the Jedi coinciden, y el coste es J=34035.23, menor que sin aplicar normalización.

Si al nuevo usuario no se le asigna ninguna calificación, las películas recomendadas serán las de mayor valoración media.

```
Top recommendations for you:
Predicting rating 0.0 for movie Lost World: Jurassic Park, The (1997)
Predicting rating 0.0 for movie E.T. the Extra-Terrestrial (1982)
Predicting rating 0.0 for movie Evita (1996)
Predicting rating 0.0 for movie Wag the Dog (1997)
Predicting rating 0.0 for movie Game, The (1997)
Predicting rating 0.0 for movie Jackal, The (1997)
Predicting rating 0.0 for movie Tin Cup (1996)
Predicting rating 0.0 for movie Volcano (1997)
Predicting rating 0.0 for movie Murder at 1600 (1997)
Predicting rating 0.0 for movie Client, The (1994)
```

Impacto de la regularización

Se observa que aunque los resultados son buenos, la escala de los ratings no es la deseada. Se puede jugar con el parámetro de regularización λ para penalizar en mayor o menor medida los valores altos de X y Θ, dando lugar a distintos resultados.

Con los datos sin normalizar, y λ=10, la máxima predicción calculada es de 5. Si se aumenta hasta λ=100, se penalizan demasiado los atributos y da lugar a predicciones entre el 0 y el 1.5 aproximadamente. Si por el contrario, se reduce este parámetro en un órden de magnitud hasta λ=1, las predicciones llegan a tomar valores de hasta 20.

Se puede utilizar esta técnica para recuperar la escala entre el 0 y el 5 para el modelo entrenado con los datos normalizados. Probando distintos valores para este modelo, se observa que utilizando λ=0.25 se obtienen los siguientes resultados:

```
Top recommendations for you:
Predicting rating 4.6 for movie Benny & Joon (1993)
Predicting rating 2.8 for movie Dear God (1996)
Predicting rating 2.6 for movie Double vie de Véronique, La (Double Life of Veronique,
The) (1991)
Predicting rating 2.6 for movie Sneakers (1992)
Predicting rating 2.5 for movie Across the Sea of Time (1995)
Predicting rating 2.5 for movie Warriors of Virtue (1997)
Predicting rating 2.5 for movie Run of the Country, The (1995)
Predicting rating 2.4 for movie Legends of the Fall (1994)
Predicting rating 2.4 for movie Nobody's Fool (1994)
Predicting rating 2.4 for movie Killing Fields, The (1984)
```

Se ha cumplido el objetivo de recuperar la escala entre 0 y 5, y además se ha disminuido la función de coste, siendo J=3780.75 al final del entrenamiento.