PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
Departamento de Ciencias de la Computación
IIC3633 - Sistemas Recomendadores
Segundo Semestre 2018

Review 1
How Not To Sort By Average Rating

Genaro Laymuns

Problema: Hay usuarios que indican que un objeto les gusta o no en un sitio web. El desafío es cómo ordenar dichos objetos de manera que los con un mejor "score"se encuentren primeros en la lista.

Al inicio del paper se indican dos maneras incorrectas de atacar este problema. Estos dos ejemplos nos permiten ver los dos factores que debe considerar un sistema de asignación de puntajes para ordenar los items: en primer lugar está el **rating promedio**, el cual se aprende en el primer ejemplo que es más representativo del score de un objeto que la diferencia absoluta de likes; y en segundo lugar se observa que se debe incluir alguna **medida de confianza** que nos diga qué tan probable es que el rating promedio real se encuentre efectivamente cerca del valor obtenido según las evaluaciones que los usuarios han realizado.

El proceso de poner like o dislike puede ser modelado como un proceso de Bernoulli de parámetro p ($\mathbb{P}(like) = p$). Si un objeto ha recibido un total de n ratings, la suma de todos los ratings positivos (consideramos obtener un 1 como like, y obtener un 0 como dislike) sigue una distribución Binomial. Así, el rating promedio estará dado por una suma de n realizaciones independientes de un proceso de Bernoulli, dividido por el número total de ratings. El hecho de que estas realizaciones sean independientes, y que por lo general se consideren valores de n elevados, nos permite utilizar el Teorema Central del Límite para aproximar el rating promedio p como un valor dentro de una distribución Normal.

En el límite $n \to \infty$, el rating promedio p sigue la siguiente distribución

$$p \dot{\sim} N(\mathbb{E}[p], \mathbb{V}ar[p]).$$

Remplazando la esperanza y varianza por los valores observados ($\hat{p}$ corresponde al rating promedio observado), se obtiene que el rating promedio sigue una distribución normal con parámetros conocidos

$$p \dot{\sim} N\left(\hat{p}, \frac{\hat{p}(1-\hat{p})}{n}\right).$$

Es posible entonces utilizar la aproximación normal para estimar con una confianza del $(1-\alpha)\,\%$ que el rating promedio real se encuentra dentro del intervalo cuyos límites son

$$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Donde $z = z_{\alpha/2}$ corresponde al cuantil $1-\alpha/2$ de la distribución normal estándar. El punto que se debe utilizar es el de la izquierda, de manera que se penalice más a los ratings cuya varianza sea elevada. Esto es una buena mejora en comparación a utilizar simplemente el rating promedio, pero sigue siendo insuficiente, pues un item con 99 likes y 1 dislike tendrá un rating menor que un item con solo un rating positivo. Los problemas que presenta esta aproximación son los siguientes:

- Solo sirve para muestras grandes. No sirve para comparar items con una diferencia considerable de ratings
- La aproximación es buena para items con $\hat{p}$ lejos de 0 y 1. En el caso de items que sean objetivamente buenos ($p \approx 1$) u objetivamente malos ($p \approx 0$), su score verdadero no será representativo mediante este modelo.

Para solucionar este problema, se utiliza en vez de el intervalo anterior, que proviene del Teorema Central del Límite, el intervalo del score de Wilson, que está dado por

$$\frac{\hat{p} + \frac{z^2}{2n}}{1 + \frac{z^2}{n}} \pm \frac{z}{1 + \frac{z^2}{n}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}$$

El score de Wilson se caracteriza por funcionar bien tanto para muestras pequeñas como para valores de $\hat{p}$ cercanos a 0 o 1. Nuevamente, debemos tomar el signo negativo para obtener el límite inferior de nuestra estimación de p.

El intervalo del score de Wilson podría ser mejorado considerando la familia de estimadores de la forma

$$\frac{\hat{p} + \omega_n p^*}{1 + \omega_n} \pm \frac{z}{1 + \omega_n}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \omega_n \cdot \frac{p^*(1-p^*)}{n}}.$$

Estos estimadores corresponden a ponderar el estimador observado $\hat{p}$ con un rating promedio fijo p^* (en el caso del score de Wilson, $p^* = 1/2$), donde el peso que se le da a $\hat{p}$ es 1 y el peso que se le da a p^* corresponde a ω_n, donde ω_n es una sucesión positiva y decreciente que converge a cero. Claramente, para valores de n elevados, el intervalo de confianza de este modelo aproxima al que entrega el Teorema Central del Límite. Este método podría presentar dos mejoras al score de Wilson:

- En el score de Wilson, los objetos con pocos ratings le dan más peso a $1/2$ que a $\hat{p}$. Sin embargo, esto podría ser una gran mejora para dicho item si en promedio los usuarios de dicho sitio web evaluaran mal a los productos. El hecho de incluir p^* como parámetro nos permite, por ejemplo, elegir p^* como el rating promedio observado de todos los productos de nuestro sitio web, de manera que los items con pocas evaluaciones sean llevados al promedio del sitio web, en vez de alejarlos de lo que los usuarios de dicho sitio web generalmente opinan de los items.

- El hecho de poder elegir la secuencia de los ω_n nos permite elegir qué tanta importancia se le quiere dar al rating observado de los items con pocas evaluaciones. Si el sitio web quiere rankear items que en general son consumidos pocas veces (como podría ser el caso de lugares para alojar), podría preferirse elegir un $\omega_n = O(1/n^2)$, de manera que se le pueda dar mayor importancia al rating promedio observado incluso para items con pocas evaluaciones.

Finalmente, es importante destacar que estos modelos solo sirven para rankear items en base a evaluaciones 0-1 (dislike-like). Si bien esto puede ser extendido al caso en que el sistema de ratings sea de 0-5 estrellas (utilizando por ejemplo que 0-2 estrellas corresponde a dislike y 3-5 a like) estos modelos nunca serán los óptimos para rankear items, pues en este caso el rating es más parecido a una distribución continua que la Bernoulli, y por lo tanto contiene mucha más información de la que los modelos recién mencionados utilizan.