Métricas de Evaluación

IIC 3633 - Sistemas Recomendadores

Denis Parra
Profesor Asistente, DCC, PUC CHile

TOC

En esta clase

1. Prediccion de Ratings: MAE, MSE, RMSE
2. Evaluacion via Precision-Recall
3. Metricas P@n, MAP,
4. Metricas de Ranking: DCG, nDCG,
5. Metricas en Tarea 1

Con respecto al paper sobre CF de Resnick et al. (1994)

- Ver Video de "re-presentación" del paper por P. Resnick y John Riedl en CSCW 2013, conmemorando que ha sido el paper más citado de dicha conferencia:
  Video CF paper re-presented at CSCW2013

3/20

Evaluación Tradicional: Predicción de Ratings

MAE: Mean Absolute Error

$$MAE = \frac{\sum_{i=1}^{n} |\hat{r}_{ui} - r_{ui}|}{n}$$

MSE: Mean Squared Error

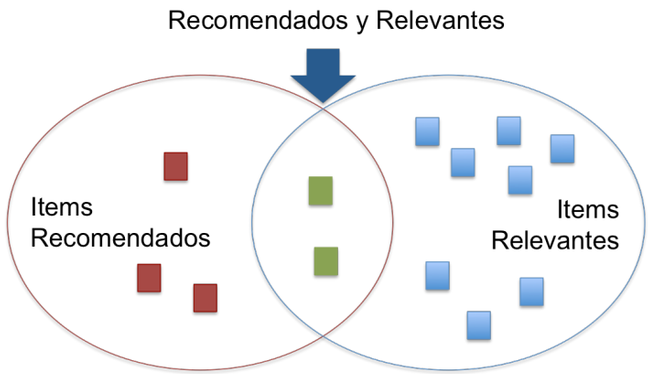
$$MSE = \frac{\sum_{i=1}^{n} (\hat{r}_{ui} - r_{ui})^2}{n}$$

RMSE: Root Mean Squared Error

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{r}_{ui} - r_{ui})^2}{n}}$$

Evaluación de una Lista de Recomendaciones

Si consideramos los elementos recomendados como un conjunto S y los elementos relevantes como el conjunto R, tenemos:



Luego, Precision es:

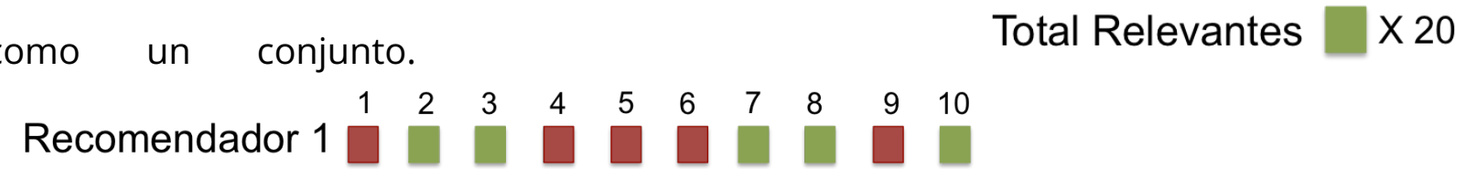
$$Precision = \frac{|Recomendados \cap Relevantes|}{|Recomendados|}, y$$

$$Recall = \frac{|Recomendados \cap Relevantes|}{|Relevantes|}$$

5/20

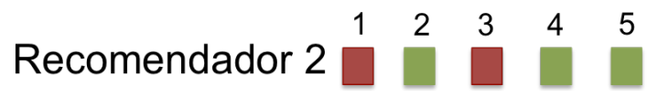
Ejemplo 1: Precision y Recall

Si bien la lista de recomendaciones está rankeada, para estas métricas la lista se entiende más bien como un conjunto.



$$Precision = ??$$

$$Recall = ??$$



$$Precision = ??$$

$$Recall = ??$$

6/20

Ejemplo 1: Precision y Recall

$$Precision = \frac{5}{10} = 0,5$$

$$Recall = \frac{5}{20} = 0,25$$

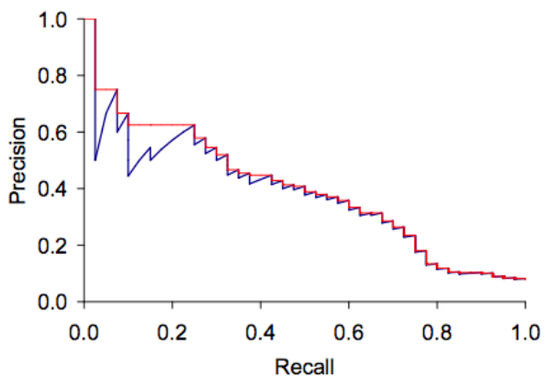
$$Precision = \frac{3}{5} = 0,6$$

$$Recall = \frac{3}{20} = 0,15$$

7/20

Compromiso entre Precision y Recall

Al aumentar el Recall (la proporción de elementos relevantes) disminuimos la precision, por lo cual hay un compromiso entre ambas métricas.



► **Figure 8.2** Precision/recall graph.

Por ello, generalmente reportamos la media harmónica entre ambas métricas:

$$F_{\beta=1} = \frac{2 * Precision * Recall}{P + R}$$

- Ref: http://nlp.stanford.edu/IR-book/pdf/08eval.pdf

8/20

De evaluación de Conjuntos a Ranking

- Mean Recicropal Rank (MRR)
- Precision@N
- MAP
- Rank score
- DCG
- nDCG

Mean Reciprocal Rank (MRR)

Consideramos la posición en la lista del primer elemento relevante.

$$MRR = \frac{1}{r}\text{, donde r: ranking del 1er elemento relevante}$$

$$MRR_1 = ??$$

$$MRR_2 = ??$$

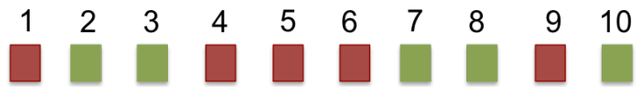
Problema: Usualmente tenemos más de un elemento relevante!!

Mean Reciprocal Rank (MRR)

Consideramos la posición en la lista del primer elemento relevante.

$$MRR = \frac{1}{r}\text{, } donde\ r\text{: } ranking\ del\ 1er\ elemento\ relevante$$

Recomendador 1



$$MRR_1 = \frac{1}{2} = 0,5$$

Recomendador 2

$$MRR_2 = \frac{1}{2} = 0,5$$

Problema: Usualmente tenemos más de un elemento relevante!!

11/20

Precision at N (P@N)

Corresponde a la *precision* en puntos específicos de la lista de items recomendados. En otras palabras, dado un ranking específica en la lista de recomendaciones, qué proporción de elementos relevantes hay hasta ese punto

$$Precision@n = \frac{\sum_{i=1}^{n} Rel(i)}{n}, donde\ Rel(i) = 1 si\ elemento\ es\ relevante$$

$$Precision@5 = ??$$

$$Precision@5 = ??$$

12/20

Precision at N (P@N)

Corresponde a la *precision* en puntos específicos de la lista de items recomendados. En otras palabras, dado un ranking específica en la lista de recomendaciones, qué proporción de elementos relevantes hay hasta ese punto

$$Precision@n = \frac{\sum_{i=1}^{n} Rel(i)}{n}, donde\ Rel(i) = 1 si\ elemento\ es\ relevante$$

$$Precision@5 = \frac{2}{5} = 0,4$$

$$Precision@5 = \frac{3}{5} = 0,6$$

Pro: permite evaluar topN; Problema: aún no permite una evalución orgánica del los items con $ranking < n$.

13/20

Mean Average Precision (MAP)

Average Precision (AP)

- El AP se calcula sobre una lista única de recomendaciones, al promediar la precision cada vez que encontramos un elemento relevante, es decir, en cada recall point.

$$AP = \frac{\sum_{k \in K} P@k \times rel(k)}{|relevantes|}$$

donde $P@k$ es la precision en el recall point k, $rel(k)$ es una función que indica 1 si el ítem en el ranking j es relevante (0 si no lo es), y K son posiciones de ranking con elementos relevantes.

MAP es la media de varias "Average Precision"

- Considerando n usuarios en nuestro dataset y que a cada uno de dimos una lista de recomendaciones,

$$MAP = \frac{\sum_{u=1}^{n} AP(u)}{m}, donde\ m\ es\ el\ numero\ de\ usuarios.$$

14/20

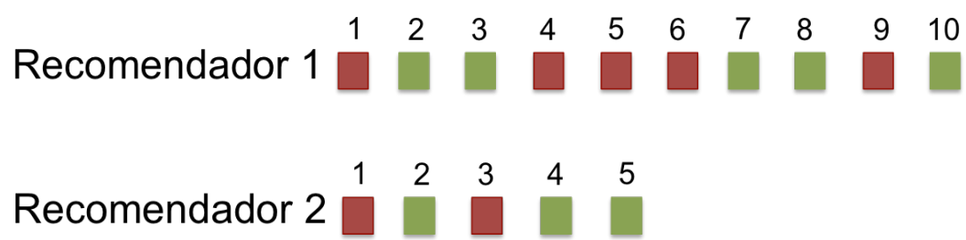
Mean Average Precision (MAP) - II

Como no siempre sabemos de antemano el número de relevantes o puede que hagamos una lista que no alcanza a encontrar todos los elementos relevantes, podemos usar una formulación alternativa** para **Average Precision (AP@n)**

$$AP@n = \frac{\sum_{k \in K} P@k \times rel(k)}{min(m, n)}$$

donde n es el máximo número de recomendaciones que estoy entregando en la lista, y m es el número de elementos relevantes.

- Ejericio: calcule $AP@n$ y luego $MAP@n$, con $n = 10$, y $m = 20$ de:



** https://www.kaggle.com/wiki/MeanAveragePrecision

15/20

Rankscore

- Rank Score se define como la tasa entre el Rank Score de los items correctos respecto al mejor Rank Score alcanzable por el usuario en teoría.

PARAMETROS	FORMULA
h el conjunto de items correctamente recomendados, i.e. hits **rank retorna la posición (rank) de un item** **T es el conjunto de items de interés** **α es el ranking half life, i.e. un factor de reducción exponencial**	$rankscore = \frac{rankscore_p}{rankscore_{\max}}$ $rankscore_p = \sum_{i \in h} 2^{-\frac{rank(i)-1}{\alpha}}$ $rankscore_{\max} = \sum_{i=1}^{\lvert T \rvert} 2^{-\frac{i-1}{\alpha}}$

16/20

DCG y nDCG

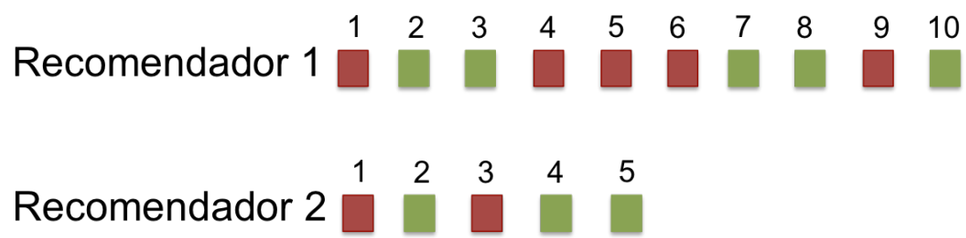
- DCG: Discounted cummulative Gain

$$DCG = \sum_{i}^{p} \frac{2^{rel_i} - 1}{log_2(1+i)}$$

- nDCG: normalized Discounted cummulative Gain, para poder comparar listas de distinto largo

$$nDCG = \frac{DCG}{iDCG}$$

Ejercicio: Calcular nDCG para



17/20

Coverage

- Como no a todos los usuarios se logran hacer recomendaciones, consideramos en la evaluación el **User Coverage**, el porcentaje de usuarios a los cuales se les pudo hacer recomendaciones.
- Como no a todos los items pueden ser recomendaciones, consideramos en la evaluación el **Item Coverage**, el porcentaje de items que fueron recomendados al menos una vez.

18/20

Métricas para Tarea 1

- Para predicción de ratings: RMSE, MAE
- Para ranking: MAP (en realidad, será MAP@10), nDCG
- Recall

Referencias

- Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval (Vol. 1, p. 6). Cambridge: Cambridge university press.
- Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern information retrieval (Vol. 463). New York: ACM press.
- Slides "Evaluating Recommender Systems" http://www.math.uci.edu/icamp/courses/math77b/lecture_12w/pdfs/Chapter%2007%20-%20Evaluating%20recommender%20systems.pdf

20/20