Sistemas Recomendadores
IIC-3633

Evaluación de Sistemas Recomendadores
Parte 2

Esta clase

1. Evaluación de sistemas recomendadores (ranking)
2. Evaluación de sistemas recomendadores (diversidad, cobertura, etc..)
3. Evaluación online de sistemas recomendadores.
4. Test estadísticos para demostrar diferencias significativas.

Ejemplo *Precision & Recall*

Total relevantes: x20 Item Recomendado Item Relevante

Recomendador 1

corte = @10

$$\text{Precisión} = \frac{|\text{Recomendados} \cap \text{Relevante}|}{|\text{Recomendados}|} = \frac{5}{10} = 0.5$$

$$\text{Recall} = \frac{|\text{Recomendados} \cap \text{Relevantes}|}{|\text{Relevantes}|} = \frac{5}{20} = 0.25$$

Recomendador 2

corte = @5

$$\text{Precisión} = \frac{|\text{Recomendados} \cap \text{Relevante}|}{|\text{Recomendados}|} = \frac{3}{5} = 0.6$$

$$\text{Recall} = \frac{|\text{Recomendados} \cap \text{Relevantes}|}{|\text{Relevantes}|} = \frac{3}{20} = 0.15$$

Voy a tener **recall máximo si recomiendo todo el catálogo** , pero voy a tener muy baja precisión porque el denominador va a ser muy grande…

¿Qué problemas puede traer precision y recall?

Mean Reciprocal Ranking (MRR)

Mide qué tan bien rankeo los ítems, corresponde al inverso de la posición del primer elemento relevante

$$MRR = \frac{1}{r}$$

r: posición del 1er elemento relevante

Recomendador 1

$$MRR = \frac{1}{r} = \frac{1}{2} = 0.5$$

Recomendador 2

$$MRR = \frac{1}{r} = \frac{1}{2} = 0.5$$

en MRR son iguales.

¿Qué problemas nos puede traer la métrica de MRR?

Precision at N (P@N)

- Corresponde a la precisión en un punto específico de la lista de los ítems recomendados

$$\text{Precision@}n = \frac{1}{n}\sum_{i=1}^{n} Rel(i)$$

$$Rel(i) = 1 \text{ si el item } i \text{ es relevante}$$

Recomendador 1

$$\text{Precision@5} = \frac{1}{n}\sum_{i=1}^{n} Rel(i) = \frac{2}{5} = 0.4$$

Recomendador 2

$$\text{Precision@5} = \frac{1}{n}\sum_{i=1}^{n} Rel(i) = \frac{3}{5} = 0.6$$

Mean Average Precision (MAP)

Average Precision (AP)

- AP se calcula promediando cada vez que encontramos un elemento relevante *(recall point)* sobre una lista única

$$\text{AP} = \frac{\sum_{k=1}^{n} P@i \cdot \text{rel}(k)}{|\text{Relevantes}|}$$

Rel(i) = 0 o Rel(i) = 1
p@k divide por k

Mean Average Precision (MAP)

- Considera el promedio de AP sobre un conjunto de listas recomendadas a todos los usuarios

$$MAP = \frac{1}{N} \sum_{u=1}^{N} AP(u)$$

N: número de usuarios o listas recomendadas

Discounted Cumulative Gain

- **DCG:** Discounted Cumulative Gain, mide la ganancia al ordenar la lista de forma correcta

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

Asume log2 ya que tenemos: rel = 1 o 0

- **nDCG:** Normalized Discounted Cumulative Gain

$$n\mathrm{DCG}_p = \frac{\mathrm{DCG}_p}{i\mathrm{DCG}_p}$$

iDCG : es el DCG ideal donde todos los elementos son relevantes.

Coverage

Coverage: cuánto del dataset puedo recomendar? La idea es acercar a la gente a TODO el contenido.

Si las recomendaciones están sesgadas a una proporción de X ítems , no es un buen signo.

Item Coverage: Porcentaje de ítems que son recomendados por lo menos una vez

User Coverage: Porcentaje de usuarios a los cuales se les pudo hacer una recomendación

Diversidad

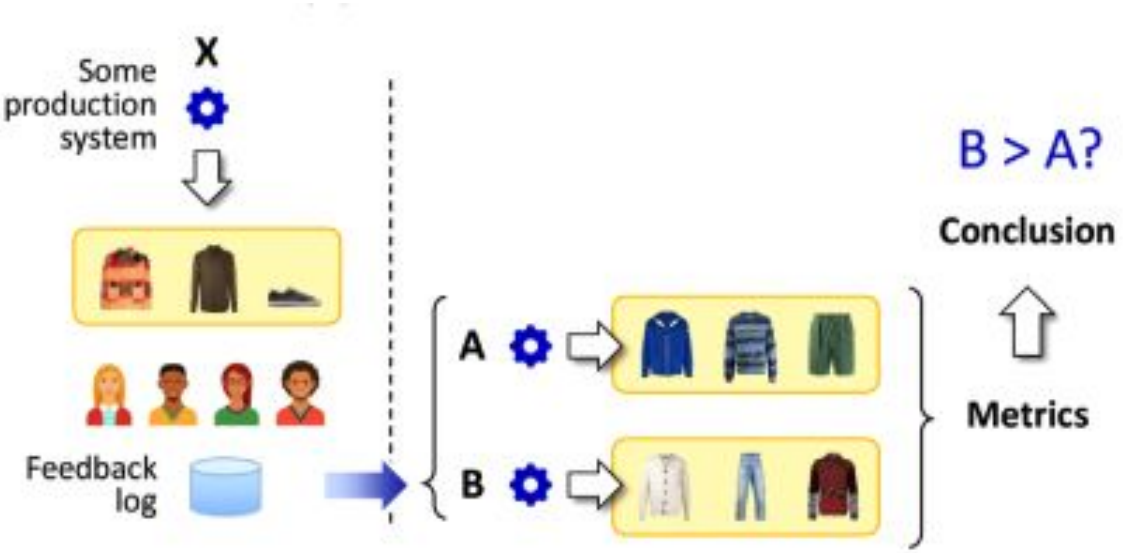
PAIRWISE DIVERSITY: Diversidad promedio que encuentro en el contenido de los items recomendados con respecto a todo el catálogo.

De dos recomendadores igual de buenos voy a preferir el que sea más diverso.

En RRSS si no promuevo la diversidad pueden haber FILTER BUBBLES.

¿Cómo evaluamos nuestro recomendador en tiempo real?

Evaluación offline



Evaluación online

- Hasta ahora hemos utilizado información histórica para simular que usuarios interactúan con un sistema recomendador.

- La evaluación online evalúa en tiempo real si las recomendaciones son buenas.

Métricas de evaluación online

Click Through Rate: división entre las recomendaciones aceptadas y todas las recomendaciones ofrecidas.

Consideramos que el **usuario aceptó la recomendación si hizo clic en al menos uno de los ítems recomendados.**

- **CTR explícito:** Se calcula si el Sistema de Recomendación (RS) tiene evidencia clara de que un usuario hizo clic en un ítem específico como resultado de la recomendación.

- **CTR implícito:** Se puede calcular basado en interacciones si el RS no tiene conocimiento explícito de que el ítem ha sido clickeado como resultado de la recomendación.

¿Por qué es importante el CTR?

Medida Directa del Interés del Usuario: El CTR refleja directamente cuánto interesa una recomendación a los usuarios.

Indicador de Relevancia: Puede señalar la pertinencia de las recomendaciones ofrecidas.

Limitaciones de CTR

No Siempre Indica Satisfacción: El hecho de que un usuario haga clic no garantiza que esté satisfecho con el contenido recomendado.

Sesgo de Posición: Los usuarios suelen hacer clic en ítems que están posicionados en los primeros lugares con más frecuencia.

Puede Favorecer Ítems Populares: En vez de ofrecer recomendaciones personalizadas, el CTR podría favorecer aquellos ítems que ya son populares.

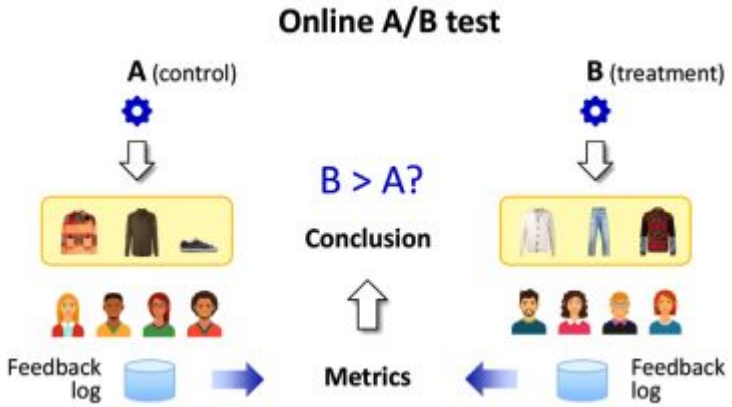
¿Como mejorar CTR ?

A/B testing para Algoritmos de Recomendación:

- Método de comparación en el cual dos versiones (A y B) se prueban simultáneamente con usuarios reales para determinar cuál es más efectiva.
- Permite analizar y escoger el algoritmo con mejor performance en términos de engagement, relevancia y otros KPIs.

Adaptar las recomendaciones según las preferencias y comportamientos individuales del usuario.

- Va más allá de simplemente mostrar lo más popular, ya que se busca entregar contenido relevante y de interés específico para cada usuario.



¿Cómo validamos que métricas de performance muestran diferencias significativas entre recomendadores?

Procedimiento general para comparar recomendadores:

Recopilación de datos: obtener un conjunto de datos de test que use ambos recomendadores para hacer predicciones.

Calcula las métricas: Usa este conjunto de datos de test para calcular las métricas mencionadas anteriormente para ambos recomendadores para cada uno de los usuarios.

Realiza tests estadísticos: Una vez que tengas las métricas, puedes usar tests estadísticos para comparar las distribuciones de los resultados de ambos recomendadores.

Test estadísticos para comparar recomendadores

Objetivo: Comparar dos sistemas de recomendación usando alguna métrica como RMSE.

Por lo tanto cada recomendador tendrá una lista de RMSEs para todos los usuarios.

Hipótesis Nula (H0): No hay diferencia entre los dos recomendadores.

Hipótesis Alternativa (H1): Hay una diferencia significativa.

Nivel de Significancia (α): Típicamente 0.05.

Elección del Test Estadístico

Criterios de Elección:

¿Los resultados tienen la misma cantidad de ejemplos (están pareados)?

¿La distribución de la diferencia en RMSE es normal?

Test Elegido: t-test pareado (o el que corresponda).

Cálculo del t-test pareado

Paso 1: Calcular la diferencia entre las observaciones pareadas de RMSE.

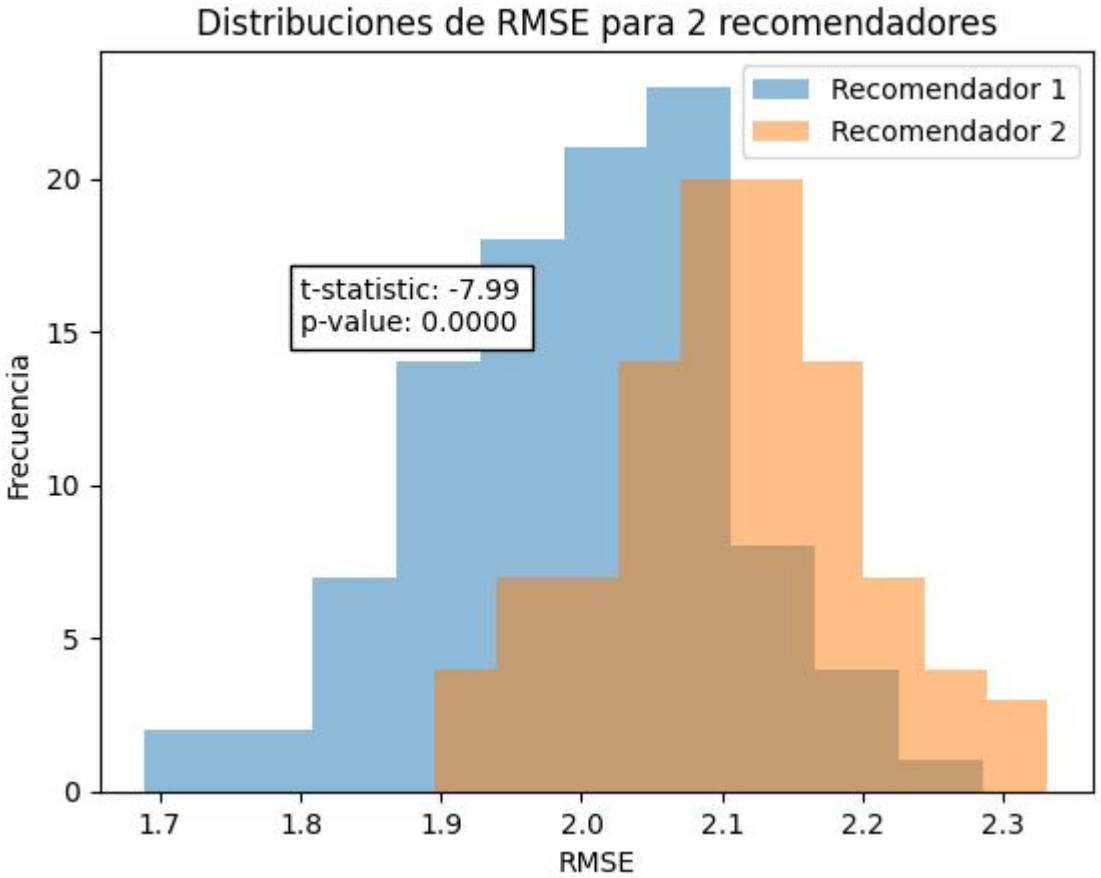
Paso 2: Calcular la media y la desviación estándar de las diferencias.

Paso 3: Calcular la estadística t usando la fórmula:

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

donde d es la media de las diferencias, sd es la desviación estándar de las diferencias, y n es el número de pares.

Paso 4: Calcular el valor p utilizando la distribución t con n−1 grados de libertad. Obteniendo el área de las colas.



En este ejemplo vemos las distribuciones de RMSE de dos recomendadores que son significativamente diferentes.

Interpretación del valor p

Interpretación: Un valor p menor que **α** (e.g., 0.05) sugiere que podemos rechazar H0.

Esto significa que **hay diferencias significativas entre los dos sistemas recomendadores**. Por lo tanto podemos escoger uno por sobre otro dependiendo de los resultados obtenidos.

Test estadísticos no paramétricos.

Permutación: permuta los datos entre los dos grupos y observar cuántas veces obtenemos una estadística tan extrema como la observada.

Bootstrap: remuestreo con reemplazo muchas veces, calcular la estadística en cada muestra y luego comparar estas estadísticas con la que se obtuvo con los datos reales para calcular el p.

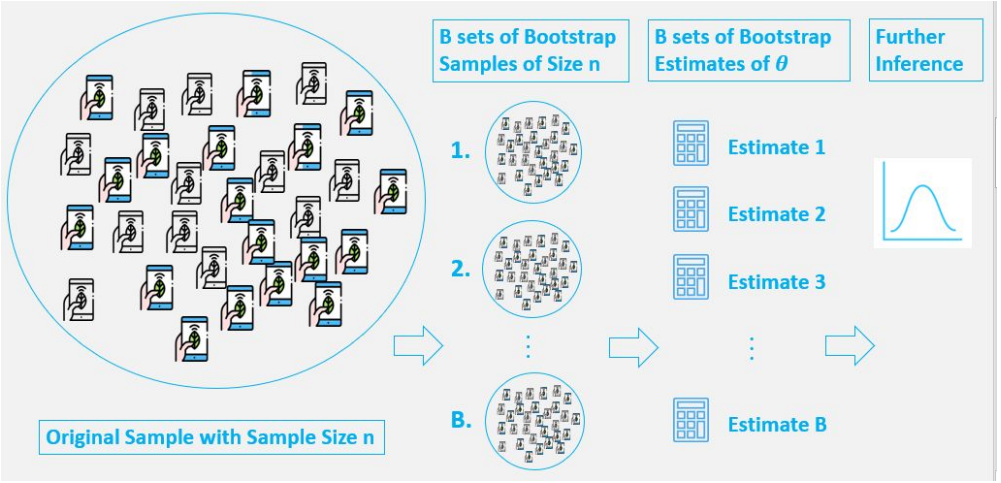


figura 1. bootstrap sampling

Test de Wilcoxon de rangos con signos

- Se basa en calcular diferencias entre pares
- El test estadístico corresponde al numero de diferencias positivas o negativas.
- Ho: la mediana de las diferencias entre pares es igual a zero.

Otras alternativas

- Test de Mann-Whitney U (o Wilcoxon Rank-Sum Test)
- Test de Kruskal-Wallis H (alternativa no-paramétrica a ANOVA)

Esta clase

1. Evaluación de sistemas recomendadores (ranking)
2. Evaluación de sistemas recomendadores (diversidad, cobertura, etc..)
3. Evaluación online de sistemas recomendadores.
4. Test estadísticos para demostrar diferencias significativas entre dos o más recomendadores.