

RJ-CHAPTER 8- EVALUATION - Notes

Pooling en TTEC

- los docs del pool representan a los asesores en orden aleatorio (por docID). Intento enterarlos juntos con todas las runs
 - TTEC Common Core Track 2017 1^{er} vez TTEC usó un alg. de ordenación del pool por DocID. Nuestro alg. de Multi-Armad Bayesian Bandits para pooling → Campus Virtual
 - Alg. Pooling intentan:
 - Reducir esfuerzo asesor
 - Fair con todos los runs que contribuyen al pool.
 - Neutras, sin tener pre sistemas que no contribuyen al pool
- Ver nuestro paper JAIST 2019 When to Stop...
Poco vor como se evalua toda esa o. O IPM 2017,
SAC 2016 para ver Multi-Armad Bandit y solo
evaluamos effort.

Query logs

Rank bias: Por la posición en el ranking las páginas más arriba tienen más posibilidades de ser "clicked" aunque no sean relevantes y viceversa.

Click generated preferences

Una vez evitado el bias y generadas preferencias para una query, esas preferencias se pueden usar como benchmark de evaluación.

Dado un ranking para esa query la calidad de este ranking se puede medir contando el número de preferencias violadas con respecto a las preferencias generadas en el benchmark. Esto es lo que mide el Tau Estimante de Correlación entre los 2 rankings de forma general.

Las páginas con high CD(d,p) Click Deviations son las que se usan para generar preferencias.

Kendall Z correlation coefficient

$$n \text{ items} \quad \binom{n}{2} = \frac{n(n-1)}{2} = n^2 \text{ pairs possible}$$

A, B, C, D, E \rightsquigarrow 10 pairs possible

System 1 rank

A
C
B
E
D

System 2 rank

A
B
C
E
D

$$Z = \frac{n^2 \text{ pairs concordants} - n^2 \text{ pairs discordants}}{n^2 \text{ pairs totals}} =$$

$$= \frac{9-1}{10} = 0.8$$

A: Rel
B: Rec

$$\text{Recall} = \frac{|A \cap B|}{|A|} = \frac{|Rel \cap Rec|}{|Rel|}$$

$$\text{Prec} = \frac{|A \cap B|}{|B|} = \frac{|Rel \cap Rec|}{|Rec|}$$

• False Positivo (Type I error) (Repara un no relevante)

$$\text{Fallout} = \frac{|\bar{A} \cap B|}{|\bar{A}|} = \frac{|\text{NoRel} \cap Rec|}{|\text{NoRel}|}$$

• False Negativo (Type II Error) (un relevante no se respondió)

1 - Recall

• Accuracy = $\frac{\# \text{ Decisions Correctas}}{\# \text{ Decisions Totales}}$

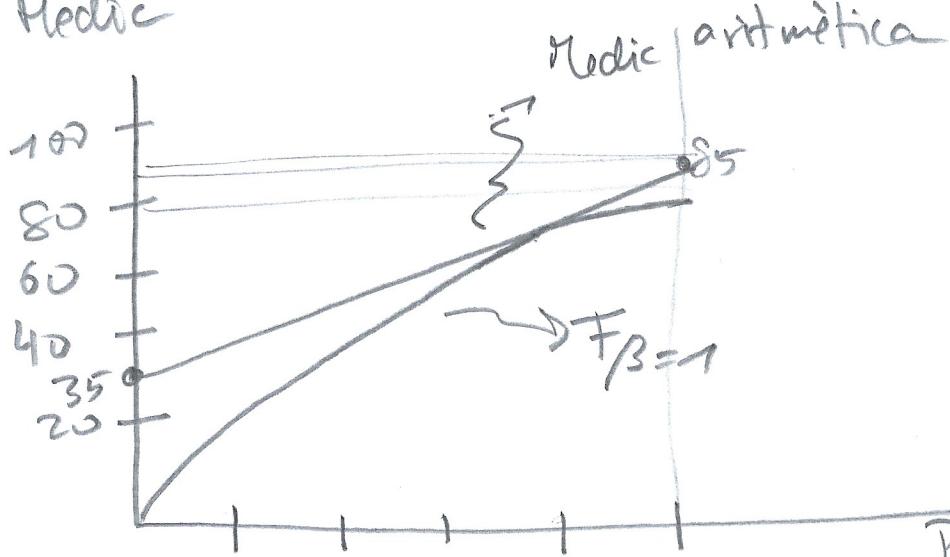
Se usa ampliamente

En IR el número de No Rel \gg Rel \rightarrow

Ocasionalmente Accuracy (Accuracy $\rightarrow 1$) si se entregan todos como No Relevantes, Esto en la práctica no vale. Por esto Precision es una mejor métrica que Accuracy en IR.

$$F_{\beta=1} = \frac{2PR}{P+R} \quad \text{Medie harmònica de } P \text{ y } R$$

Medie



(Recall Fijo al 70%)

Si $P \neq R$ bajan, la medie harmònica penaliza mas que la medie aritmética:

$$\overline{AP(q_j)} = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} \text{Precision}(R_{jk})$$

$q_j \in Q$ tiene $\{d_1, \dots, d_{m_j}\}$ relevantes

R_{jk} es el rank de documentos recuperados desde la posición j hasta la posición k para la query q_j

- Se suele aplicar hasta un umbral de corte, por ejemplo cut 1000 en TREC Eval

$AP@1000$, Es la expresión anterior proporcionando $MAP@1000$ en el cut 1000 del rank.

Pa1K	Pa25	Pa10	Pa20
Pa2K	Pa25	Pa10	Pa20

R-Precision \equiv R-Prec \equiv PaRel, donde
 Rel es el num. de relevantes para la query
 Perfect System \rightarrow R-Prec = 1

DCG Example

- discounted gain $\frac{\text{rel}_i}{\log_2 i} \rightarrow$ apartir de $i=2$,
 \hookrightarrow en $i=1$
 discounted gain = rel_1

• DCG =

$$\text{rel}_1 + \sum_{i=2}^n \frac{\text{rel}_i}{\log_2 i} = \underline{\underline{\text{DCG}_{\text{ap}}}}$$

$$\bullet \text{ NDCG}_{\text{ap}} = \frac{\text{DCG}_{\text{ap}}}{\text{IDCG}_{\text{ap}}}$$

SIGNIFICANCE TESTS

Error Tipos I: Se rechaza la hipótesis nula siendo cierta la hipótesis nula.

Muchos errores Tipos I \rightarrow El test tiene poca exactitud

Error Tipos II: Se acepta la hipótesis nula siendo falsa la hipótesis nula.

Muchos errores Tipos II \rightarrow El test tiene poca potencia,
i.e., no rechaza una hipótesis nula que debería
rechazar

nivel de significancia α ($0'1, 0'05, 0'01$)

p -value $< \alpha$, la probabilidad de que el estadístico del test tome un valor tan extremo si la hipótesis nula fuera cierta es muy baja y $< \alpha$,
por tanto rechaza nula la hipótesis nula con poca probabilidad de equivocarse.

Region of Rejection

El t-test supone una distribución normal de los valores del estadístico bajo la hipótesis nula.

$\alpha = 0.05$. El área sombreada supone el 5% del área bajo la curva.

Los valores t del estadístico que caen en el eje t en la zona sombreada se corresponden con $p\text{-value} < 0.05$.

Por cada valor t del estadístico

$$p\text{-value} = \frac{\text{área bajo curva a la derecha de } t}{\text{área total}}$$

t-test

distribución normal $\text{mean} = 0$
 $\text{desviación estandar} = 1$

$\bar{B}-A$ medida de las diferencias

$\sigma_{\bar{B}-A}$ desviación estandar de las diferencias

$$\sigma = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{N}}$$

Wilcoxon Signed Rank-Test

Hipótesis nula: la suma de los ranks positivos es igual a la suma de los ranks negativos

Para computar el p-value, computar todos los posibles signed rank y sus sumas, y en el histograma de sumas ver donde cae el valor χ del estadístico posible signed ranks en el grupo.

-1, -2, -3 ... 7, 8, 9

+1, +2, +3 ... +7, +8, +9

...

Para valores de $n > 20$, en vez de computar todos el histograma se suele ajustar por una gaussiana.

- El t-test supone que las diferencias se pueden ordenar y que su magnitud importa
- wilcoxon solo supone que las diferencias se puedan ordenar
- Sgn-t-test solo supone que lo que importa es el signo de las diferencias.

Todos los tests suponen que los guerris son independientes