CS4104 Applied Machine Learning

Evaluation Measures

Evaluating a Machine Learning Algorithm

- Relevance is assessed relative to the **information need**
- E.g., Information need: *I'm looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine.*
- Query: ***wine red white heart attack effective***
- Evaluate whether the doc addresses the information need, not whether it has these words

zeshan.khan@nu.edu.pk

Dataset

Supervised

- Train Test Data
- Evaluation/Ground Truth

Un-Supervised

- Train Test Data

Standard Datasets

Textual

- GOV2
  - Another TREC/NIST collection
  - 25 million web pages
  - Largest collection that is easily available
  - But still 3 orders of magnitude smaller than what Google/Yahoo/MSN index
- Cross Language Evaluation Forum (CLEF)
  - This evaluation series has concentrated on European languages and cross-language information retrieval.
- TREC (Text Retrieval Conference)
  - 450 Queries/Information Needs
  - 1.89 M Documents
- Reuters-RCV2
- 20 Newsgroups
  - 18941 articles

Image

- Image Net
  - Millions of Images
  - 1000 classes
- Object Net
  - Millions of Images
  - 1000 classes
- MNIST
  - 10 classes

Evaluation Measures

Un-Ranked Results

- Precision
- Recall
- Accuracy
- F-Measure
- MCC
- Jaccard Index

Ranked Results

- Top 5 Accuracy
- Mean Average Precision
- Normalized Discounted Cumulative Gain

Evaluation Measures

- **TP**: True Positive
  - Number of relevant documents retrieved.
- **FP**: False Positive
  - Number of documents retrieved but irrelevant.
- **TN**: True Negative
  - Number of irrelevant documents not retrieved.
- **FN**: False Negative
  - Number of relevant documents not retrieved.

	Relevant	Irrelevant
Retrieved	TP	FP
Not Retrieved	FN	TN

zeshan.khan@nu.edu.pk

Evaluation Measures

- **Precision**: fraction of retrieved docs that are relevant
  - $P = \frac{TP}{TP+FP} = P(relevant|retrieved)$
- **Recall**: fraction of relevant docs that are retrieved
  - $R = \frac{TP}{TP+FN} = P(\text{retrieved}|\text{relevant})$
- **Accuracy**: the fraction of correct retrieval.
  - $Acc = \frac{TP+TN}{TP+TN+FP+FN}$
- **Fall-out**: The proportion of non-relevant documents retrieved.
  - $Fall-out = \frac{FP}{FP+TN}$

	Relevant	Irrelevant
Retrieved	TP	FP
Not Retrieved	FN	TN

Confusion Matrix

Corpus=120 Relevant=100	Retrieved	Relevant Retrieved
Model 1	80	80
Model 2	90	70
Model 3	120	100
Model 4	0	0
Model 5	50	50

Model 1	Relevant	Irrelevant
Retrieved	80	0
Not-Retrieved	20	20

Model 2	Relevant	Irrelevant
Retrieved	70	20
Not-Retrieved	30	0

Model 3	Relevant	Irrelevant
Retrieved	100	20
Not-Retrieved	0	0

Model 4	Relevant	Irrelevant
Retrieved	0	0
Not-Retrieved	100	20

Model 5	Relevant	Irrelevant
Retrieved	50	0
Not-Retrieved	50	20

zeshan.khan@nu.edu.pk

	Relevant	Irrelevant
Retrieved	TP	FP
Not Retrieved	FN	TN

Precision and Recall

	Precision	Recall
Model 1		
Model 2		
Model 3		
Model 4		
Model 5		

$$P = \frac{TP}{TP+FP}$$

$$R = \frac{TP}{TP+FN}$$

Model 1	Relevant	Irrelevant
Retrieved	80	0
Not-Retrieved	20	20

Model 2	Relevant	Irrelevant
Retrieved	70	20
Not-Retrieved	30	0

Model 3	Relevant	Irrelevant
Retrieved	100	20
Not-Retrieved	0	0

Model 4	Relevant	Irrelevant
Retrieved	0	0
Not-Retrieved	100	20

Model 5	Relevant	Irrelevant
Retrieved	50	0
Not-Retrieved	50	20

zeshan.khan@nu.edu.pk

	Relevant	Irrelevant
Retrieved	TP	FP
Not Retrieved	FN	TN

Precision and Reca

	Precision	Recall
Model 1	80/80=1	80/100=0.8
Model 2	70/90=0.78	70/100=0.7
Model 3	100/120=0.83	100/100=1
Model 4	0/0= NA	0/100=0
Model 5	50/50=1	50/100=0.5

$$P = \frac{TP}{TP+FP}$$

$$R = \frac{TP}{TP+FN}$$

Model 1	Relevant	Irrelevant
Retrieved	80	0
Not-Retrieved	20	20

Model 2	Relevant	Irrelevant
Retrieved	70	20
Not-Retrieved	30	0

Model 3	Relevant	Irrelevant
Retrieved	100	20
Not-Retrieved	0	0

Model 4	Relevant	Irrelevant
Retrieved	0	0
Not-Retrieved	100	20

Model 5	Relevant	Irrelevant
Retrieved	50	0
Not-Retrieved	50	20

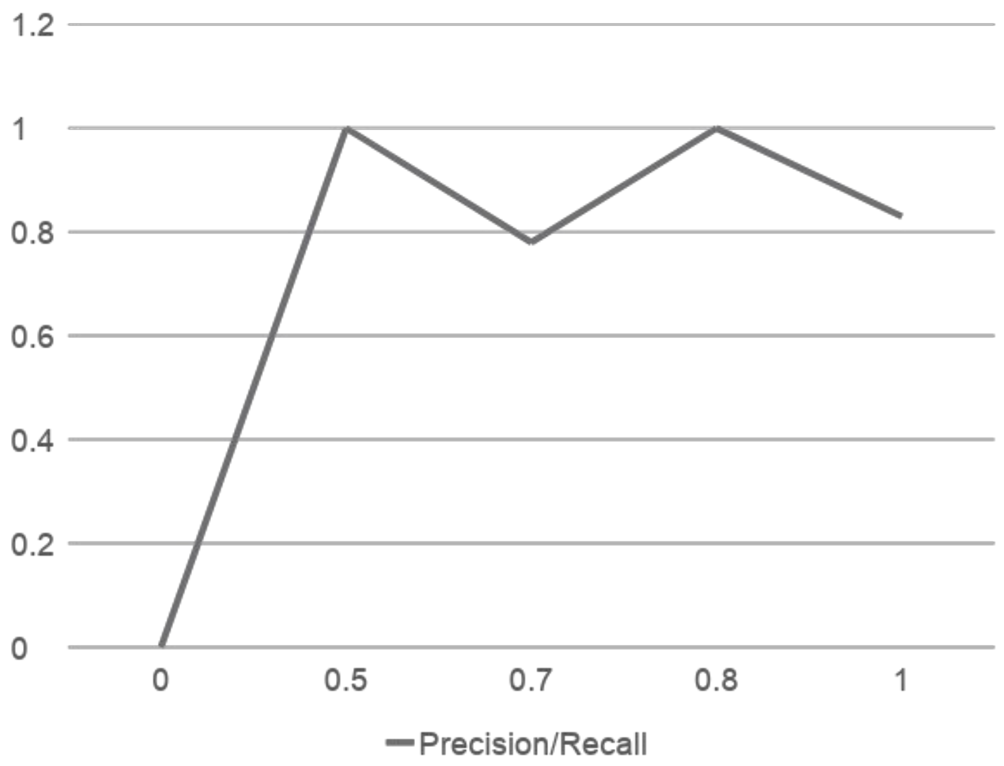
zeshan.khan@nu.edu.pk

Precision and Recall

	Precision	Recall
Model 1	80/80=1	80/100=0.8
Model 2	70/90=0.78	70/100=0.7
Model 3	100/120=0.83	100/100=1
Model 4	0/0= NA	0/100=0
Model 5	50/50=1	50/100=0.5

$$P = \frac{TP}{TP+FP}$$

$$R = \frac{TP}{TP+FN}$$



Accuracy

	Accuracy
Model 1	
Model 2	
Model 3	
Model 4	
Model 5	

$$Acc = \frac{TP+TN}{TP+TN+FP+FN}$$

Model 1	Relevant	Irrelevant
Retrieved	80	0
Not-Retrieved	20	20

Model 2	Relevant	Irrelevant
Retrieved	70	20
Not-Retrieved	30	0

Model 3	Relevant	Irrelevant
Retrieved	100	20
Not-Retrieved	0	0

Model 4	Relevant	Irrelevant
Retrieved	0	0
Not-Retrieved	100	20

Model 5	Relevant	Irrelevant
Retrieved	50	0
Not-Retrieved	50	20

zeshan.khan@nu.edu.pk

Accuracy

	Accuracy
Model 1	100/120=0.83
Model 2	70/120=0.58
Model 3	100/120=0.83
Model 4	20/120=0.16
Model 5	70/120=0.58

$$Acc = \frac{TP+TN}{TP+TN+FP+FN}$$

Model 1	Relevant	Irrelevant
Retrieved	80	0
Not-Retrieved	20	20

Model 2	Relevant	Irrelevant
Retrieved	70	20
Not-Retrieved	30	0

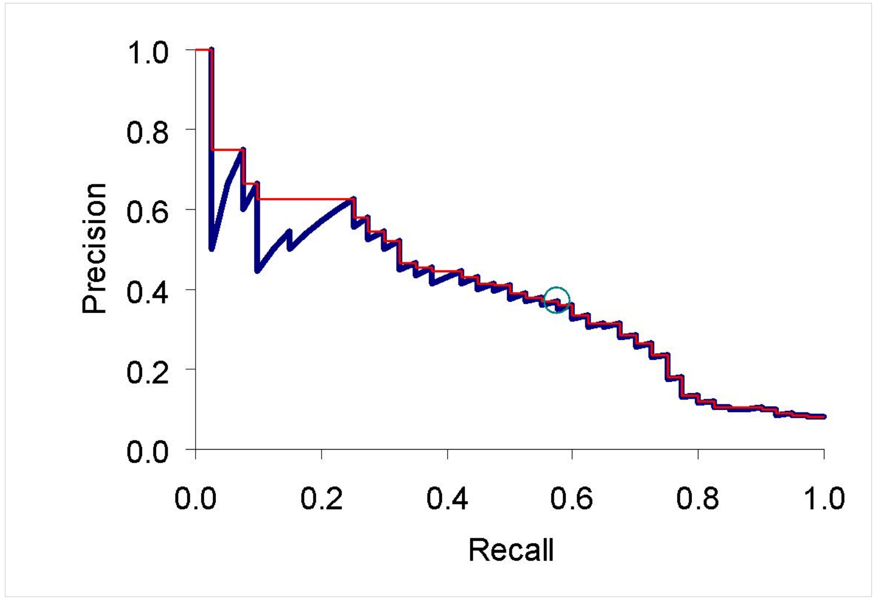
Model 3	Relevant	Irrelevant
Retrieved	100	20
Not-Retrieved	0	0

Model 4	Relevant	Irrelevant
Retrieved	0	0
Not-Retrieved	100	20

Model 5	Relevant	Irrelevant
Retrieved	50	0
Not-Retrieved	50	20

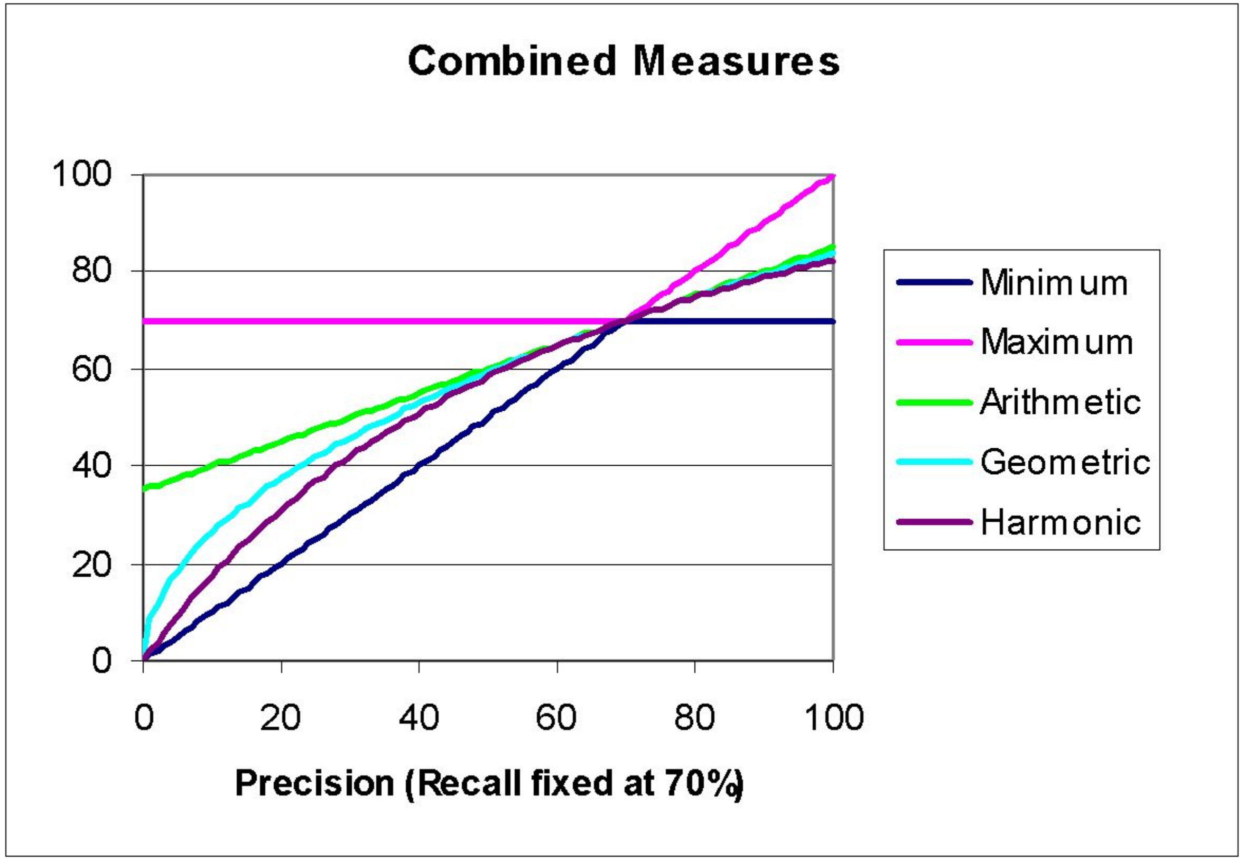
Precision/Recall

- You can get high recall (but low precision) by retrieving all docs for all queries!
- Recall is a non-decreasing function of the number of docs retrieved
- In a good system, precision decreases as either the number of docs retrieved or recall increases
  - This is not a theorem, but a result with strong empirical confirmation



zeshan.khan@nu.edu.pk

Comminated Measures



Weighted Harmonic Mean (F-Measure)

- Combined measure that assesses precision/recall tradeoff is **F measure** (weighted harmonic mean):
- $F = \frac{1}{\alpha * \frac{1}{P} + (1-\alpha) * \frac{1}{R}}$
- $\beta^2 = \frac{1-\alpha}{\alpha}$
- $F = \frac{(\beta^2+1)PR}{\beta^2 P + R}$
- People usually use balanced F_1 measure
- $\beta = 1$ or $\alpha = \frac{1}{2}$
- $F1 = \frac{2PR}{P+R}$

	Relevant	Irrelevant
Retrieved	TP	FP
Not Retrieved	FN	TN

F1-Score, F1-Meas

	Precision	Recall	F1-Measure
Model 1	1	0.8	(2*1*0.8)/1.8=0.89
Model 2	0.78	0.7	(2*0.78*0.7)/1.48=0.74
Model 3	0.83	1	(2*0.83*1)/1.83=0.91
Model 4	NA	0	NA
Model 5	1	0.5	(2*1*0.5)/1.5=0.67

Model 1	Relevant	Irrelevant
Retrieved	80	0
Not-Retrieved	20	20

Model 2	Relevant	Irrelevant
Retrieved	70	20
Not-Retrieved	30	0

Model 3	Relevant	Irrelevant
Retrieved	100	20
Not-Retrieved	0	0

Model 4	Relevant	Irrelevant
Retrieved	0	0
Not-Retrieved	100	20

Model 5	Relevant	Irrelevant
Retrieved	50	0
Not-Retrieved	50	20

Matthews Correlation Coefficient (MCC)

- $MCC = \frac{(TP*TN)-(FP*FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$
- Perfect return 1
- Worst results -1
- 0 for the random resuls

Jaccard Index (JI)

- $JI = \frac{Intersection}{Union} * 100$
- Intersection: The number of common elements in the prediction and ground truths
- Union: Total number of distinct values in predicted and ground truths
- Range of value: 0 to 100
- 0 for Worst
- 100 for Best

Evaluation Measures

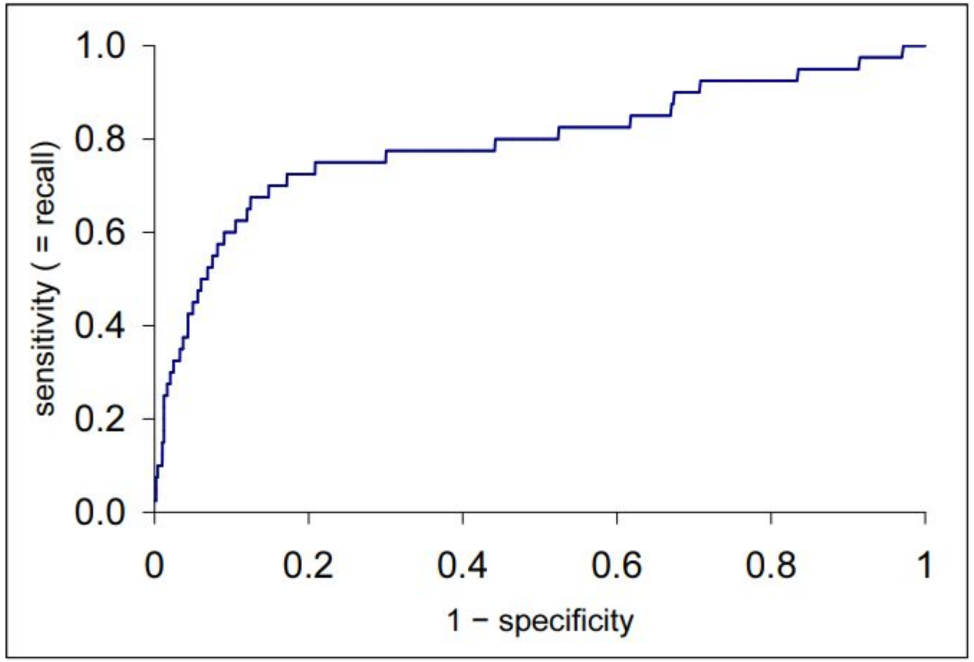
TPR (Sensitivity): True Positive Rate

$$TPR = \frac{TP}{TP+FN}$$

FPR: False Positive Rate $FPR = \frac{FP}{FP+TN}$

$$Specificity = \frac{TN}{TN+FP}$$

ROC Curve: A curve of TPR on FPR



zeshan.khan@nu.edu.pk

Evaluation of ranked results

- Evaluation of ranked results:
  - The system can return any number of results
  - By taking various numbers of the top returned documents (levels of recall), the evaluator can produce a *precision-recall curve*

Ranking Evaluation Measures

- Top 5 Accuracy
- Interpolated precision (P_{interp}): The highest precision at recall r.
- Average Interpolated precision (P_{interp}): The highest precision at recall r.
- R-precision: Precision at cut-off R (top R relevant documents).
- Precision at K: Precision from top k documents retrieved.
- BREAK-EVEN POINT:
- Average Precision AveP: The precision average of thee ranked documents.
- Mean average precision (MAP): Average precision for top k documents.
- Cumulative Gain (CG):
- Discounted Cumulative Gain (DCG):
- Normalized Discounted Cumulative Gain (NDCG):

zeshan.khan@nu.edu.pk

Accuracy

- Top 1 Accuracy
  - The accuracy considering top 1 element as true
- Top 5 Accuracy
  - The accuracy considering top 5 element as true

Interpolated precision (P_{interp})

P_{interp}

The highest precision at recall r.

$$P_{interp}(r) = \max_{r1 \le r} P(r1)$$

Example

$P = \{0.1, 0.5, 0.6, 0.6, 0.5, 0.5, 0.7, 0.9, 1\}$

$R = \{1, 0.9, 0.7, 0.5, 0.4, 0.4, 0.3, 0.1, 0\}$

$P_{interp}(0) = 1$

$P_{interp}(0.5) = 0.6$

$P_{interp}(0.9) = 0.5$

$P_{interp}(1) = 0.1$

zeshan.khan@nu.edu.pk

Interpolated precision (P_{interp})

$AvgP_{interp}$

Commonly used 11 point average interpolation precision

$$AvgP_{interp} = \frac{1}{11}\sum_{i\in(0,0.1,0.2,\ldots,1.0)} \max_{r\geq i} P(r)$$

Example

$$P = \{0.1,0.5,0.6,0.6,0.5,0.5,0.7,0.9,1\}$$

$$R = \{1,0.9,0.7,0.5,0.4,0.4,0.3,0.1,0\}$$

$$AvgP_{interp} = \frac{1}{11}\sum_{i\in\{0,0.1,0.2,\ldots,1.0\}}\{1,0.9,0.7,0.7,0.6,0.6,0.6,0.6,0.5,0.5,0.1\}$$

$$AvgP_{interp} = 0.62$$

zeshan.khan@nu.edu.pk

R-precision

- If we have a known (though perhaps incomplete) set of relevant documents of size *Rel*, then calculate precision of the top *Rel* docs returned
- Perfect system could score 1.0.

Average Precision AveP

AveP

- The precision average of thee ranked documents.
- $AveP = \sum_{k=1}^{n} \frac{(P(k) \times rel(k))}{docs-count}$
- $P(k)$: The precision at cut-off k
- $rel(k) = 1\ if\ doc_k\ is\ relavant\ zero\ otherwise$

Example

- $D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8$
- $q_1 = \{\mathbf{1}, 0.5, \mathbf{0.66}, \mathbf{0.75}, \mathbf{0.8}, 0.66, 0.57, \mathbf{0.63}\}$
- $rel_{q1} = \{1,0,1,1,1,0,0,1\}$
- $AveP = \frac{1+0.66+0.75+0.80+0.63}{5} = 0.77$
- $q_2 = \{\mathbf{1}, 0.5, \mathbf{0.66}, 0.75, \mathbf{0.8}, \mathbf{0.66}, 0.57, \mathbf{0.63}\}$
- $rel_{q2} = \{1,0,1,0,1,1,0,1\}$
- $AveP = \frac{1+0.66+0.80+0.66+0.63}{5} = 0.75$

zeshan.khan@nu.edu.pk

Mean average precision (MAP)

MAP

- Average of the precision value obtained for the top k documents, each time a relevant doc is retrieved
- Avoids interpolation, use of fixed recall levels
- MAP for query collection is arithmetic average.
  - Macro-averaging: each query counts equally
- $MAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q}$

EXAMPLE $(D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8)$

- $q_1 = \{\mathbf{1}, 0.5, \mathbf{0.66}, \mathbf{0.75}, \mathbf{0.8}, 0.66, 0.57, \mathbf{0.63}\}$
- $rel_{q1} = \{1,0,1,1,1,0,0,1\}$
- $AveP = \frac{1+0.66+0.75+0.80+0.63}{5} = 0.77$
- $q_2 = \{\mathbf{1}, 0.5, \mathbf{0.66}, 0.75, \mathbf{0.8}, \mathbf{0.66}, 0.57, \mathbf{0.63}\}$
- $rel_{q2} = \{1,0,1,0,1,1,0,0\}$
- $AveP = \frac{1+0.66+0.80+0.63}{4} = 0.75$
- $MAP = \frac{1}{2}(0.77 + 0.75) = 0.76$

zeshan.khan@nu.edu.pk

Cumulative Gain

Cumulative Gain (CG):

- $CG = \sum_{i=1}^{docs} rel_i$
- rel_i is the relevancy of the i^{th} documents in ranked retrieved.

Example

$D_1, D_2, D_3, D_4, D_5, D_6$

$relevance = \{3,2,3,0,1,2\}$

$CG_6 = rel_1, rel_2, rel_3, rel_4, rel_5, rel_6$

$CG_6 = 3 + 2 + 3 + 0 + 1 + 2 = 11$

Discounted Cumulative Gain

Discounted Cumulative Gain (DCG):

- $DCG = \sum_{i=1}^{docs} \frac{rel_i - 1}{\log_2 i+1}$

Example

$D_1, D_2, D_3, D_4, D_5, D_6$

$relevance = \{3,2,3,0,1,2\}$

$\sum(3,1.3,1.5,0,0.4,0.7) = 6.9$

$DCG = 6.9$

i			
1	3	1	3
2	2	1.6	1.3
3	3	2	1.5
4	0	2.3	0
5	1	2.6	0.4
6	2	2.8	0.7

zeshan.khan@nu.edu.pk

Normalized Discounted Cumulative Gain

Normalized Discounted Cumulative Gain (NDCG):

- $nDCG = \frac{\sum_{i=1}^{docs} \frac{rel_i - 1}{\log_2 i+1}}{\sum_{i=1}^{docs_{sorted}} \frac{rel_i - 1}{\log_2 i+1}}$

Example

$D_1, D_2, D_3, D_4, D_5, D_6$

$relevance = \{3,2,3,0,1,2\}$

$\sum(3,1.8,1,0.9,0.4,0) = 7.2$

$nDCG = \frac{6.9}{7.2} = 0.96$

i			
1	3	1	3
2	3	1.6	1.8
3	2	2	1
4	2	2.3	0.9
5	1	2.6	0.4
6	0	2.8	0

zeshan.khan@nu.edu.pk

Cluster Evaluation

- Silhouette Index
- Davies Bouldin
- Calinski Harabasz

Silhouette Index

- Measurement of consistency of clusters
- Mean Distance Inner/Intra Cluster
  - $\mathrm{sim}(p) = \frac{1}{|C_L|} \sum_{i \in C_L} d(i,p) : p \in C_L$
  - Evaluation of the assignment of p
- Mean Distance Outer
  - $diff(p) = \min_{L \neq M} \frac{1}{|C_M|} \sum_{i \in C_M} d(i,p) : p \in C_L$
  - Evaluation of the assignment of p with near most cluster
- Silhouette Value of p
  - $s(p) = \frac{diff(p) - sim(p)}{\max(diff(p), sim(p))}\ if\ |C_L| > 1$
  - $s(p) = 0\ if |C_L| = 1$

- $s(p) = \begin{cases} 1 - \frac{sim(p)}{diff(p)} & if \quad sim(p) < diff(p) \\ 0 & if \quad sim(p) = diff(p) \\ \frac{diff(p)}{sim(p)} - 1 & if \quad sim(p) > diff(p) \end{cases}$
- $s(P) = \mathrm{mean}_{p \in P}\, s(p)$
- Value of Silhouette (-1,+1)

Davies Bouldin

- $DB = \frac{1}{|c|} \sum_{i=1}^{|c|} R_i$
- $R_i = \max_{j=1,..,nc, i \neq j} R_{ij}, i = [1, \dots, |c|]$
- $R_{ij} = \frac{s_i + s_j}{d_{ij}}$
- $d_{ij} = d(v_i, v_j)$
- $s_i = \frac{1}{|c_i|} \sum_{x \in c_i} d(x, v_i)$
- v_i is the centroid of c_i

Calinski Harabasz

-
- $CH = \frac{\left(\frac{\left(\sum_{k=1}^{K} n_k |c_k - c|^2\right)}{K-1}\right)}{\frac{\sum_{k=1}^{K} \sum_{i=1}^{n_k} |d_i - c_k|^2}{N-K}}$

- n_k number of points in cluster k
- c_k centroid of cluster k
- c centroid of all clusters
- N total points