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

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

Evaluation Measures

	Relevant	Irrelevant
Retrieved	TP	FP
Not Retrieved	FN	TN

- **Precision:** fraction of retrieved docs that are relevant

- **Recall:** fraction of relevant docs that are retrieved

- **Accuracy**: the fraction of correct retrieval.

- **Fall-out**: The proportion of non-relevant documents retrieved.

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

	Relevant	Irrelevant
Retrieved	TP	FP
Not Retrieved	FN	TN

Precision and Recal

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

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

	Relevant	Irrelevant
Retrieved	TP	FP
Not Retrieved	FN	TN

Precision and Recal

	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

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

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

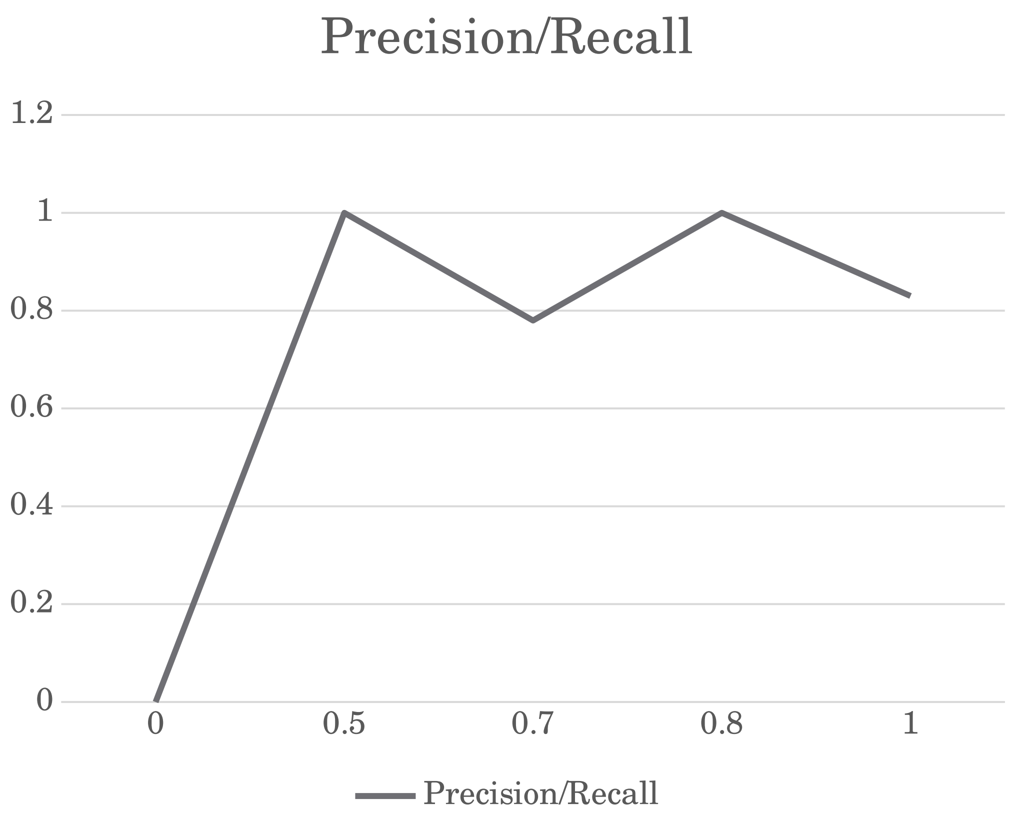
Model 3	Relevant	Irrelevant
Retrieved	100	20
Not-Retrieved	0	0

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

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

Precision 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



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

Accuracy

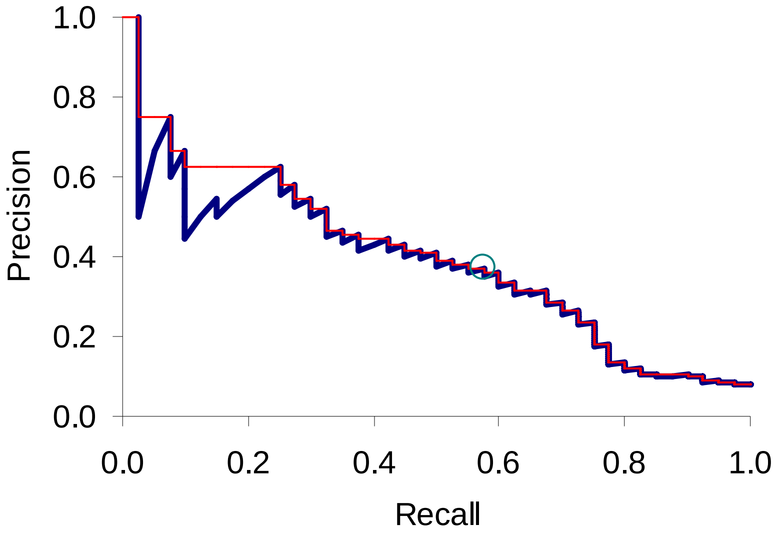
	Accuracy
Model 1	100/120=0.83
Model 2	70/120=0.58
Model 3	100/120=0.83
Model 4	20/120=0.16
Model 5	70/120=0.58

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

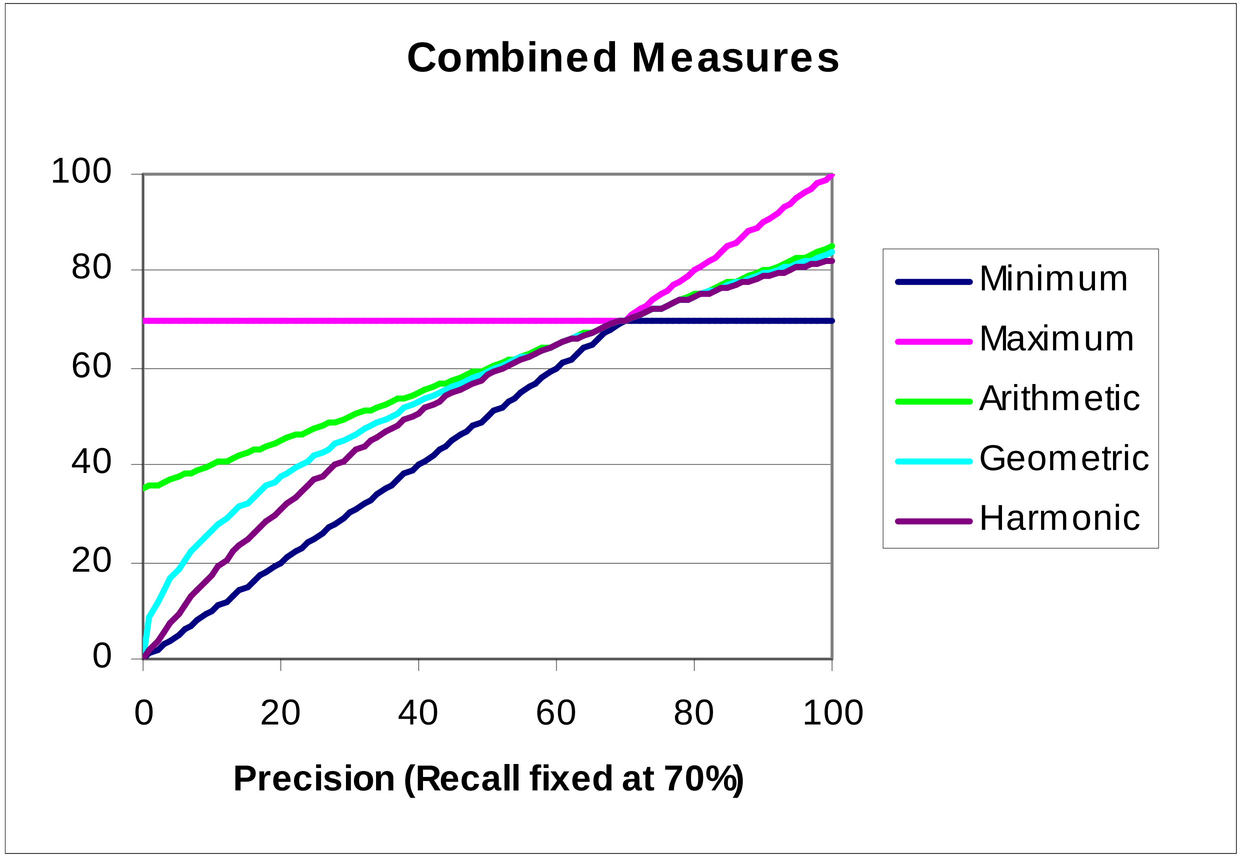
Model 1	Relevant	Irrelevant
Retrieved	80	0
Not-Retrieved	20	20
Model 2	**Relevant**	**Irrelevant**
Retrieved	70	20
Not-Retrieved	30	0
Model 3	**Relevant**	**Irrelevant**
Retrieved	100	20
Not-Retrieved	0	0
Model 4	**Relevant**	**Irrelevant**
Retrieved	0	0
Not-Retrieved	100	20
Model 5	**Relevant**	**Irrelevant**
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



Comminated Measures



Weighted Harmonic Mean (F-Measure)

- Combined measure that assesses precision/recall tradeoff is **F measure** (weighted harmonic mean):

- People usually use balanced F_1 measure

- or

	Relevant	Irrelevant
Retrieved	TP	FP
Not Retrieved	FN	TN

F1-Score, F1-Measu

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

$$F1=\frac{2PR}{P+R}$$

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)

- Perfect return 1
- Worst results -1
- 0 for the random resuls

Jaccard Index (JI)

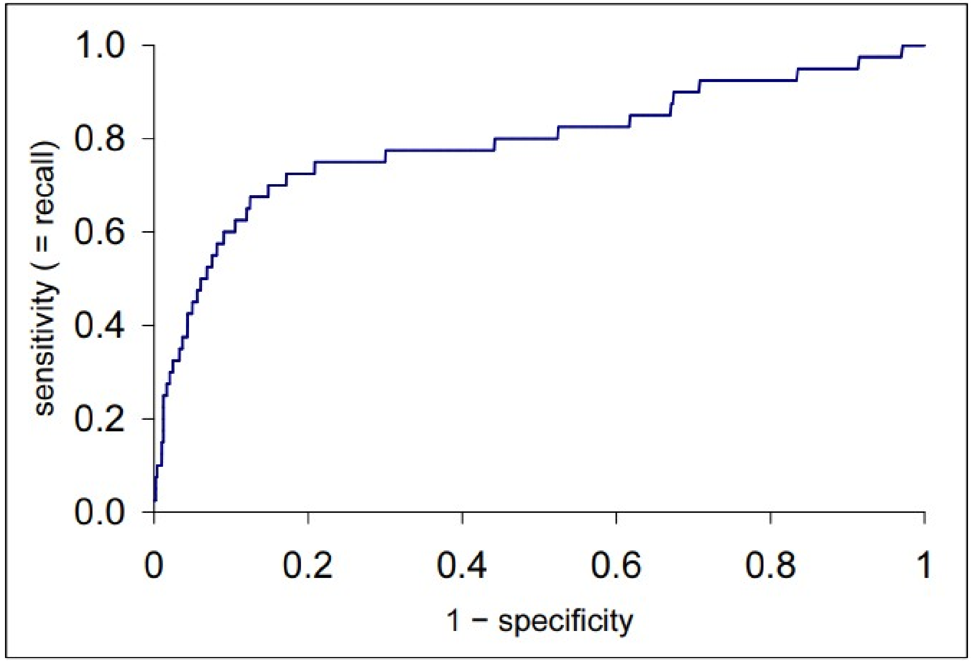
- Intersection: The number of common elements in the prediction and ground truths
- Union: Total number of distinct values in predicted and ground truths
- Range of value: 0 to 100
- 0 for Worst
- 100 for Best

Evaluation Measures

- TPR (Sensitivity): True Positive Rate

- FPR: False Positive Rate

- ROC Curve: A curve of TPR on FPR



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 (): The highest precision at recall r.
- Average Interpolated precision (): 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):

Accuracy

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

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

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

Example ()

Cumulative Gain

- Cumulative Gain (CG):
  - is the relevancy of the documents in ranked retrieved.
- Example

Discounted Cumulative Gain

- Discounted Cumulative Gain (DCG):

- Example

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

Normalized Discounted Cumulative Gain

- Normalized Discounted Cumulative Gain (NDCG):

- Example

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

Cluster Evaluation

- Silhouette Index
- Davies Bouldin
- Calinski Harabasz

Silhouette Index

- Measurement of consistency of clusters
- Mean Distance Inner/Intra Cluster
  - Evaluation of the assignment of p
- Mean Distance Outer
  - Evaluation of the assignment of p with near most cluster
- Silhouette Value of p

- Value of Silhouette (-1,+1)

Davies Bouldin

- is the centroid of

Calinski Harabasz

- number of points in cluster k
- centroid of cluster k
- centroid of all clusters
- total points