Admin

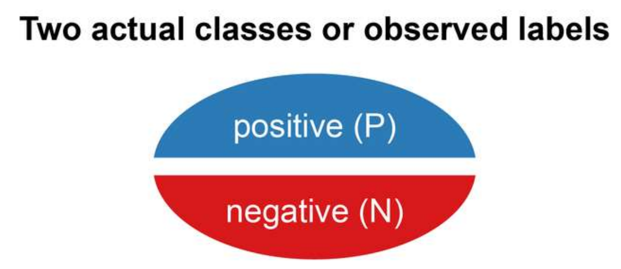
Thursday Presentation

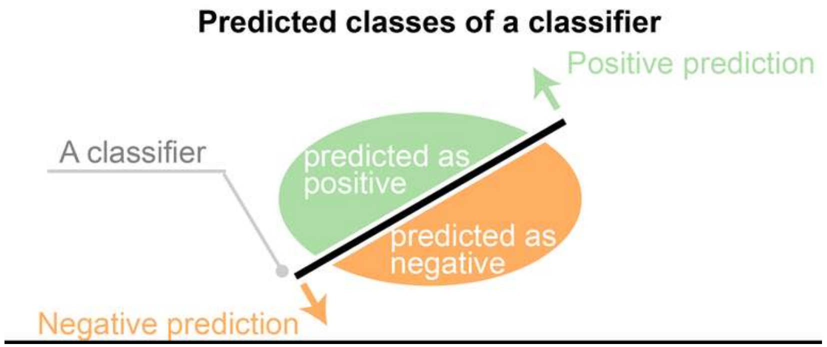
- Yuxuan Zhang,
- Damiana Fitria K,
- Renswick Delvar,
- William Shondelmyer,
- Thomas Legge,
- Vijai Simmon

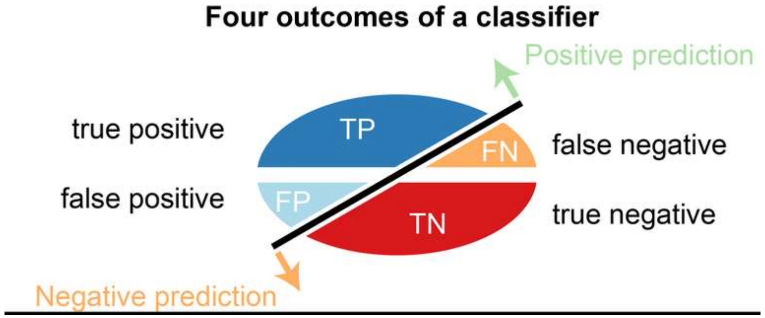
- Paper review due this Friday

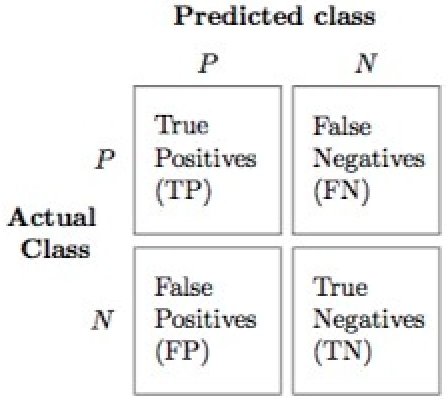
EVALUATION

Classification Systems Evaluation


Two actual classes or observed labels


Predicted classes of a classifier


Four outcomes of a classifier



		Predicted class	
		P	N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}}$$

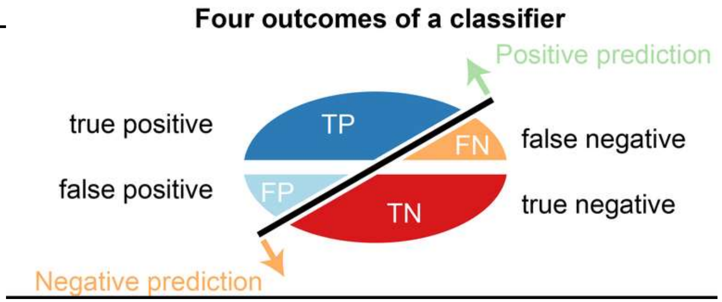
$$\mathrm{ERR} = \frac{\mathrm{FP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}} = \frac{\mathrm{FP} + \mathrm{FN}}{\mathrm{P} + \mathrm{N}}$$

Information Retrieval Evaluation

- Data collection
  - TREC
  - Queries; documents labelled as relevant and not-relevant
- Evaluation criteria
  - Precision: Percentage of retrieved documents that are relevant

$$P = \frac{\#of\ \mathrm{Re}levantItems\,\mathrm{Re}trieved}{\#ofItem\,\mathrm{Re}trieved}$$

P = TP/ (TP + FP)



  - Recall: Percentage of all relevant documents that are found by a search

$$R = \frac{\#of\ \mathrm{Re}levantItems\,\mathrm{Re}trieved}{\#of\ \mathrm{Re}levantItemsInCollection}$$

  - R = TP / (TP + FN) = TP/ P

IR evaluation discussion

- Exercise: calculate precision and recall
  - For a query, If a system finds 200 results, among them 50 are relevant.
  - The human labels (model solutions) have 120 relevant documents.
- Why not use Accuracy or Error rate in IR?

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}}$$

- Which is more important in Web search: precision or recall?
- How to compare two IR systems

Evaluation: F measure, AUC, MAP

- F-score is a harmonic mean of precision and recall.

$$\mathrm{F}_1 = \frac{2 \cdot \mathrm{PREC} \cdot \mathrm{REC}}{\mathrm{PREC} + \mathrm{REC}}$$

- AUC: Area under the precision and recall curve
- Top N precision
- MAP: consider ranking, precision, recall
  - Mean of the Average Precision for all queries
  - Average Precision: the mean of the precision when each relevant document is retrieved. (M is the No of relevant documents)

$$\mathrm{MAP}(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} \mathrm{Precision}(R_{jk})$$

  - Average precision is roughly the area under the precision and recall curve
- ARR: the average rank of the documents rated as “relevant”

Evaluation in general

- Information retrieval evaluation methods can be used for evaluation in many other areas

- Recommender can be binary: change rates to positive or negative
  - Precision
  - Top N precision
  - Recall
  - F-measure

Recommender Systems Evaluation

- Consider ranking score

- MAE: mean absolute error

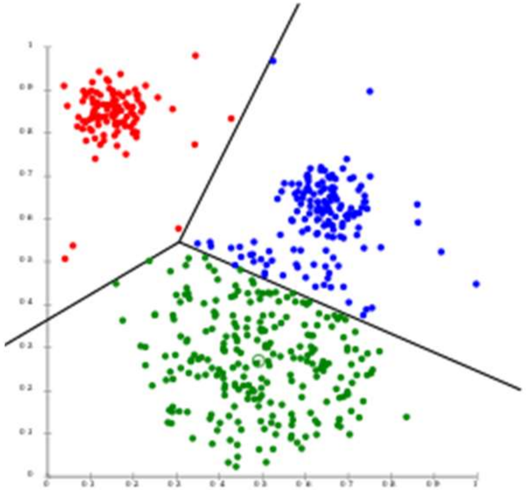
$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|\,.$$

Clustering systems Evaluation

- No labels (for training)
- Labels are not used in training
- Use labels only for evaluation

Rand Index = (TP + TN)/ (TP+ TN + FN + FP)

- Typically consider document pairs rather than individual document
- Pair of documents: same class label in the same cluster TP



Personalized Search Evaluation

- In lab setting

    10-500 users

- Quantitative & Qualitative
- System performance
- User evaluation, system usability
- Data sets

    open web corpora, in-lab generated logs,

    TREC collection, search engine query logs

    subset of annotated documents from specific sites

Evaluation in reality, in practise

- A/B testing
- A/B testing (sometimes called split testing) is experimenting and comparing two types or variations of an online or offline campaign such as a landing page, ad text, a headline, call-to-action or just about any other element of a marketing campaign or ad.
- By displaying two variations of your campaign, you can see which one attracts more interaction and conversions from your customers.
  - e.g. CTR (clickthrough rate): the number of clicks that your ad receives divided by the number of times your ad is shown

Query expansion

- Query expansion (QE) is the process of reformulating a seed query to improve retrieval performance in information retrieval operations.

Query expansion

Existing research and their limitations

- Two approaches
    - Relevance feedback
        - Pseudo relevance feedback
            - Challenge: what if the top N are not relevant
    - Thesaurus-based
        - May cause concept drifting
        - May improve recall, but often hurt precision

Exploiting Underrepresented Query Aspects for Automatic Query Expansion

Daniel Crabtree, Peter Andreae, Xiaoying Gao

Victoria University of Wellington

New Zealand

Main idea

- Identify query aspects
  - Air New Zealand flight to Canada
  - Black bear attacks
- Build a vocabulary for each aspect
  - Search results for each aspect
  - Frequent words and co-related words with aspects
- Identify Under-represented aspects
  - Vocabulary coverage in original results
- Expand the under-represented aspects

Example: Find Query Aspects

- Black Bear Attacks People

Black Bear

Attacks

People

Black	
Black Bear	
“Black Bear”	“Bear Black”
Black Bear Attacks	
“Black Bear Attacks”	“Attacks Black Bear”
	“Attacks Bear Black”
	“Black Attacks Bear”
	“Bear Attacks Black”
	“Bear Black Attacks”
Attacks People	
“Attacks People”	“People Attacks”
People	

Find Query Aspects

Global Document Analysis

- Web Frequency of Query Subsequences
- Relative Frequency Based Heuristic Determines Aspects

Example: Black Bear Attacks People

- Number of occurrences anywhere on page
  - Black Bear Attacks
- Number of occurrences as a phrase
  - “Black Bear Attacks”
- Number of occurrences of alternative phrases: permutations
  - “Attacks Black Bear”
  - “Attacks Bear Black”
  - “Black Attacks Bear”
  - “Bear Attacks Black”
  - “Bear Black Attacks”

Identify Underrepresented Aspects

- Different aspects -> different vocabulary
  - Car
    - Automobile
    - Petrol
    - Motor
    - Safety
    - Transport
  - Black Bear
    - Animal
    - Mammal
    - Diet
    - Habitat
    - Cub

Identify Underrepresented Aspects

1) Build Vocabulary Model for each Aspect

- Sub-queries: aspect and each aspect pair (pairs address polysemy)
- Most frequent words with strongest aspect relationship

Analyze Original Document Set

- Score: Coverage of each aspects vocabulary model
- Low score -> underrepresented aspect

Black Bear Attacks People

Black Bear Aspect

Sub-queries

Black Bear	
Black Bear	Attacks
Black Bear	People

Black Bear Vocabulary

★ Animal
Mammal
★ Diet
★ Habitat
Cub

Find Refinement

1) Pick terms from underrepresented aspect(s)

- or if none, stop and use original query

2) Run queries for each term

- New Query = original query + term

3) Determine aspect representation scores

- Use documents from new query
- Use vocabulary models identified earlier

4) Apply Best Scoring Refinement

Advantages and disadvantages

- Work for hard queries when no relevant documents are retrieved

- Many sub-queries, not efficient for real time search

- Better way to
  - Identify aspects
  - Built aspect vocabulary model
  - Better way to match the model with documents

How to improve it using more recent research on word similarities, etc.

Apply it to a different problem

- Large language models

- Prompt based fine tuning

- Automatically refine the prompt