Data Science Bootcamp, 11th January 2017

Evaluation Approaches

Kriste Krstovski

Columbia University
Data Science Institute

How do we evaluate and compare models?

- Performance Measures
  - Extrinsically
    - Precision and recall
  - Intrinsically
    - Log-likelihood and perplexity
- Statistical Measures
  - Significance
    - Randomization test
  - Correlation
    - Linear and rank correlation

Relevance Measures

- Precision

$$P = \frac{|Relevant\ \&\ Retrieved|}{|Retrieved|}$$

- Recall

$$R = \frac{|Relevant\ \&\ Retrieved|}{|Relevant|}$$

Model 1

1.	Doc K	0.134
2.	Doc A	0.187
3.	Doc M	0.203
4.	Doc Z	0.329
5.	Doc L	0.348
6.	Doc T	0.452
7.	Doc E	0.484
8.	Doc F	0.522
9.	Doc S	0.593
10.	Doc J	0.643
. ...		...
20.	Doc P	1.322
. ...		...
26.	...	...

Model 2

1.	Doc M	12.132
2.	Doc Q	9.881
3.	Doc P	9.343
4.	Doc K	9.108
5.	Doc U	8.884
6.	Doc J	8.756
7.	Doc F	7.453
8.	Doc Z	7.332
9.	Doc S	7.128
10.	Doc H	6.845
. ...		...
20.	Doc O	4.087
. ...		...
26.	...	...

Relevance Set

Doc F
Doc J
Doc K
Doc M
Doc P

Relevance Measures

- Precision

$$P = \frac{|Relevant \ \& \ Retrieved|}{|Retrieved|}$$

- Recall

$$R = \frac{|Relevant \ \& \ Retrieved|}{|Relevant|}$$

Model 1

1.	Doc K	0.134
2.	Doc A	0.187
3.	Doc M	0.203
4.	Doc Z	0.329
5.	Doc L	0.348
6.	Doc T	0.452
7.	Doc E	0.484
8.	Doc F	0.522
9.	Doc S	0.593
10.	Doc J	0.643
….		…
20.	Doc P	1.322
….		…
26.	…	…

Model 2

1.	Doc M	12.132
2.	Doc Q	9.881
3.	Doc P	9.343
4.	Doc K	9.108
5.	Doc U	8.884
6.	Doc J	8.756
7.	Doc F	7.453
8.	Doc Z	7.332
9.	Doc S	7.128
10.	Doc H	6.845
….		…
20.	Doc O	4.087
….		…
26.	…	…

Relevance Set

Doc F
Doc J
Doc K
Doc M
Doc P

$$P@10 = \frac{4}{10} = 0.4$$

$$R@10 = \frac{4}{5} = 0.8$$

Relevance Measures

- Precision

$$P = \frac{|Relevant\ \&\ Retrieved|}{|Retrieved|}$$

- Recall

$$R = \frac{|Relevant\ \&\ Retrieved|}{|Relevant|}$$

Model 1

1.	Doc K	0.134
2.	Doc A	0.187
3.	Doc M	0.203
4.	Doc Z	0.329
5.	Doc L	0.348
6.	Doc T	0.452
7.	Doc E	0.484
8.	Doc F	0.522
9.	Doc S	0.593
10.	Doc J	0.643
….		…
20.	Doc P	1.322
….		…
26.	…	…

Model 2

1.	Doc M	12.132
2.	Doc Q	9.881
3.	Doc P	9.343
4.	Doc K	9.108
5.	Doc U	8.884
6.	Doc J	8.756
7.	Doc F	7.453
8.	Doc Z	7.332
9.	Doc S	7.128
10.	Doc H	6.845
….		…
20.	Doc O	4.087
….		…
26.	…	…

Relevance Set

Doc F
Doc J
Doc K
Doc M
Doc P

$$P@10 = \frac{4}{10} = 0.4$$

$$R@10 = \frac{4}{5} = 0.8$$

$$P@10 = \frac{5}{10} = 0.5$$

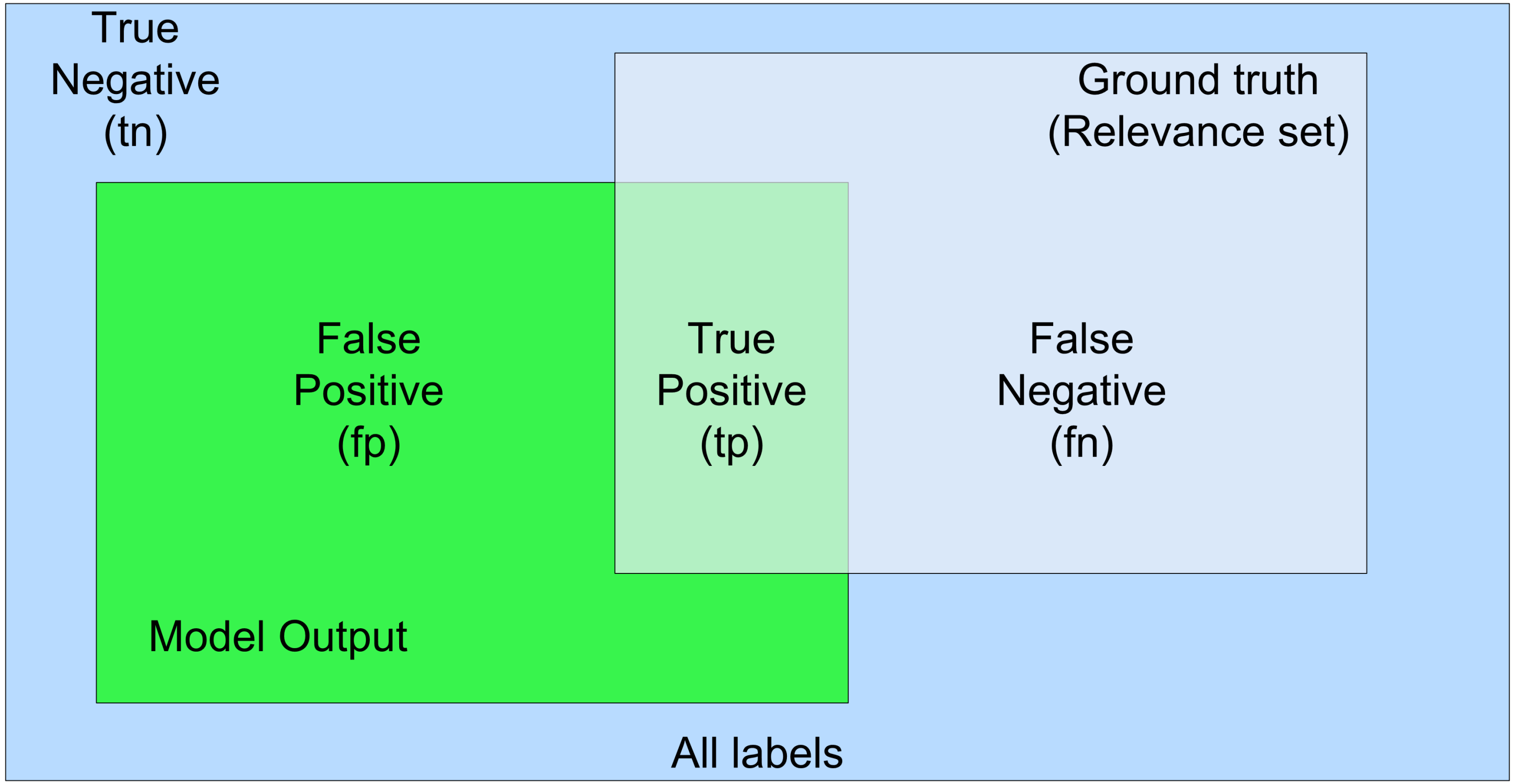
$$R@10 = \frac{5}{5} = 1.0$$

Relevance Measures

- Binary Classification

$$P = \frac{|tp|}{|tp + fp|}$$

$$R = \frac{|tp|}{|tp + fn|}$$



Log-likelihood & Perplexity

- Log-likelihood

$$\mathcal{L}(x) \;=\; \sum_{i=1}^{n} \log p(x_i|\theta)$$

- Perplexity

$$perplexity(x) \;=\; \exp\left\{ -\frac{\mathcal{L}(x)}{|n|} \right\}$$

Log-likelihood & Perplexity

- Perplexity

$$perplexity(x) \;=\; \exp\left\{-\frac{\mathcal{L}(x)}{|n|}\right\}$$

- Perplexity for topic models

$$perplexity(D) \;=\; \exp\left\{-\frac{\sum_{d=1}^{M}\log p(w_d)}{\sum_{d=1}^{M}N_d}\right\}$$

How do we evaluate and compare models?

- Performance Measures
    - Extrinsically
        - Precision and recall
    - Intrinsically
        - Log-likelihood and perplexity
- **Statistical Measures**
    - **Significance**
        - **Randomization test**
    - Correlation
        - Linear and rank correlation

Randomization Test

- Also known as the permutation test
- Determine whether the difference in the test statistic used to judge two models is statistically significant or not
- Null hypothesis is that the two models are identical

Randomization Test Components

- Test statistics by which models/systems are judged
  - e.g. difference in the mean of some metric
- Distribution of the test statistic under the null-hypothesis
- Significance level
  - How likely a difference value as large or larger than our experiment's difference value could have occurred under the null hypothesis

Randomization Test Algorithm

- Create the distribution of the test statistic under the null hypothesis
    1. Repeat n times:
        1. Go over each data point in the results set
        2. Randomly choose an evaluation result from the two model results for that data point
        3. Repeat the process twice (once for each model) and compute the mean for each of the newly generated set of evaluation results
        4. Compute the difference between the two means (i.e. the test statistic)
        5. Store test statistic for the n-th iteration
- Go over the n generated test statistics and count the number of times their values were larger than our original test statistic
- Compute p-value

Measuring Correlation

- Linear correlation
  - Measures linear dependence between two sets of values
  - Pearson's R
- Rank correlation
  - Measures similarity between two rankings
  - Spearman's ρ

In This Lab Session

- Learn how to compute precision and recall
- Evaluate topic models using log-likelihood
- Learn how to implement the randomization test
- Compare model performance using linear and rank correlation