

COMP348

Evaluation of Statistical Corpus-Based Approaches

Training vs Test Data

- For pretty much all evaluation, you want to divide your data into at least two sets: training and test
- Training data is the one (typically marked-up) that you calculate your statistics over
 - e.g. $\Pr(\text{FS} \mid w_1^2)$ from earlier example
- Test data is separate (not marked-up)
- You may also have a third set (DevTest) to help develop your system

Precision and Recall

- Errors by systems making a binary choice can often be broken into two types.
 - Selecting something when it's not supposed to be.
 - Not selecting something when it is supposed to be.
- Examples:
 - If the task is to identify documents as relevant or not, the system can mistakenly classify irrelevant documents as relevant, or relevant documents as irrelevant.
 - If the task is to identify whether a full stop is the end of a sentence or not, the system can mistakenly classify abbreviations as end of sentence markers, or vice versa.

Precision and Recall

- Can group results of system into four categories: tp (true positive), fp (false positive), fn (false negative), tn (true negative)

	actual	
system	target	not target
selected	tp	fp
not selected	fn	tn

Precision and Recall

- **Formulas:**
 - $\text{precision} = \text{tp} / (\text{tp} + \text{fp})$
 - $\text{recall} = \text{tp} / (\text{tp} + \text{fn})$
- **Example:**
 - You have a system to decide on sentence segmentation based on identifying full stops as either sentence delimiters or abbreviation markers. It tells you that there are 30 sentence delimiters, but 5 are wrong. It also missed 10 sentence delimiters, which it indicated were abbreviation markers. There were also 10 full stops which were correctly identified as abbreviation markers.
 - $\text{precision} = 25 / 30$
 - $\text{recall} = 25 / 35$

Accuracy

- **Accuracy is the number correctly classified out of the whole set**
 - $\text{acc} = (\text{tp} + \text{tn}) / (\text{tp} + \text{fp} + \text{tn} + \text{fn})$
 - From previous example, accuracy is 35/50
- **Sometimes used (inaccurately) for precision**

Spam Filtering

- **Another classification task (the topic of next week's lectures) is separating spam from non-spam**
- **Exercise:**
 - Assume your system processes 1000 emails. It classifies 640 as spam, of which 480 are correct. It misses 120 spam emails. What are the precision and recall (of the spam detection) and accuracy of the system?

Baselines

- **Typically, you want to compare the performance of your system to something else**
 - This could be something simple: for example, what if you just guessed the category of the full-stop?
 - Alternatively, you might compare it against a standard algorithm

Hypothesis Testing

- Often, want to check whether some experimental result, say the output of a system, is likely to have happened by chance.
 - Is it “significant”?
- Examples:
 - Is the mean outcome of this experiment significantly different from the expected?
 - Are two success rates significantly different from each other?
- Postulate H_0 , the null hypothesis, to test

Hypothesis Testing

- Basic idea
 - Want to calculate some normalised statistic in testing whether two things are different
 - The normalising is done by dividing the difference by a measure of dispersion (typically, standard deviation)
 - This value is then compared against a standard distribution

Basic z-Test: Example

- H_0 : mean score of Wangimatta High School Maths Ext 1 students is the same as the mean for all students
- Data:
 - \bar{x} = mean score of Wangimatta students = 42.1
 - μ = mean score of all students (2003) = 39.8
 - σ = standard deviation (sqrt of variance) for all students = 7.6
 - n = number of students in sample = 52
- Statistic:
 - $z = (42.1 - 39.8) / (7.6 / \sqrt{52}) = 2.18$

Basic z-Test: Tables

- How do you know if this is significant?
 - Consult a z-statistic table

level of significance α		z-value
two-sided	one-sided	
0.001 (i.e. 0.1%)	0.0005	3.29
0.002	0.001	3.09
0.0026	0.0013	3.00
0.01	0.005	2.58
0.02	0.01	2.33
0.0456	0.0228	2.00
0.05	0.025	1.96
0.10	0.05	1.64
0.20	0.10	1.28
0.318	0.159	1.00

Basic z-Test: Example

- Two-sided is for testing whether your value of interest is just different; one-sided is for testing whether it's larger (or smaller)
- For our example, we're just testing for difference
 - Our value of 2.18 is >1.96 , so there's less than a 5% likelihood it's just random chance that the Wangimatta students' mean score is different

z-Test of Proportion

- Can also test proportions
- You have some proportion p , and you want to compare it against some (hypothesised) population proportion p_0
- z-statistic is
 - $z = (p - p_0) / \sqrt{p_0 * (1 - p_0) / n}$
- Why?
 - It's based on a binomial distribution, and the denominator is the standard deviation of that distribution
 - This is approximate, but OK for sample > 30

z-Test of Proportion

- Assume you have a classification system with overall accuracy rate 0.7258. Is this different from a (hypothesised) population accuracy rate of 0.6?
- Data:
 - $p = 0.7258$
 - $p_0 = 0.6$
 - $n = 620$
- Statistic:
 - $z = 6.39$ (way off the scale)

z-Test of Two Proportions

- In the previous example, we were comparing a sample proportion against a known population proportion
- What about if comparing two samples, with proportion p_1 from the first sample (of size n_1) and p_2 from the second sample (size n_2)?
- Also a z-statistic:
 - $z = (p_1 - p_2) / \sqrt{P * (1 - P) * (1/n_1 + 1/n_2)}$
 - where $P = (p_1 * n_1 + p_2 * n_2) / (n_1 + n_2)$

z-Test of Two Proportions

- Say we want to compare the WebKR overall accuracy against another system AS, whose sample of 90 files has an accuracy of 0.75. Is this difference significant?
- Data
 - $p1 = 0.7258$
 - $p2 = 0.75$
 - $n1 = 620$
 - $n2 = 90$
- Statistic:
 - $z = -0.4826$
 - Not significant at any typical level

Test of Two Sample Means

- What if we want to compare two sample of heights of people and see if they are different?
- Requires a different statistic (t-test)