Today’s Lecture

- Problem interpreting results: statistical significance
- Problem with datasets: social bias
Current State of NLP

- Emphasis on empirical results
- Statistical significance rarely discussed
Current State of NLP

- Emphasis on empirical results
- Statistical significance rarely discussed
- Large number of architectures, hyperparameters
Current State of NLP

- Emphasis on empirical results
- Statistical significance rarely discussed
- Large number of architectures, hyperparameters
- Datasets re-used many times
Dror et al. (2018) survey

<table>
<thead>
<tr>
<th></th>
<th>ACL 2017</th>
<th>TACL 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total papers</td>
<td>196</td>
<td>37</td>
</tr>
<tr>
<td>Experimental papers</td>
<td>180</td>
<td>33</td>
</tr>
</tbody>
</table>
Dror et al. (2018) survey

<table>
<thead>
<tr>
<th></th>
<th>ACL 2017</th>
<th>TACL 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total papers</td>
<td>196</td>
<td>37</td>
</tr>
<tr>
<td>Experimental papers</td>
<td>180</td>
<td>33</td>
</tr>
<tr>
<td>– reporting significance</td>
<td>63 (35%)</td>
<td>18 (55%)</td>
</tr>
</tbody>
</table>
### Dror et al. (2018) survey

<table>
<thead>
<tr>
<th></th>
<th>ACL 2017</th>
<th>TACL 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total papers</td>
<td>196</td>
<td>37</td>
</tr>
<tr>
<td>Experimental papers</td>
<td>180</td>
<td>33</td>
</tr>
<tr>
<td>– reporting significance</td>
<td>63 (35%)</td>
<td>18 (55%)</td>
</tr>
<tr>
<td>– correctly</td>
<td>36 (20%)</td>
<td>15 (45%)</td>
</tr>
</tbody>
</table>
p-Values

- Probability the result would be at least this extreme, under the null hypothesis
p-Values

- Probability the result would be at least this extreme, under the null hypothesis

NOT:

- Probability the null hypothesis is true
Statistical Significance Testing

- Decide on a **null hypothesis**
- Decide on a **test statistic**
- Decide on a **threshold**

Significance level: probability of incorrectly rejecting null hypothesis (assuming null hypothesis)

Power: probability of correctly rejecting null hypothesis (assuming alternative hypothesis)
Statistical Significance Testing

- Decide on a **null hypothesis**
- Decide on a **test statistic**
- Decide on a **threshold**

**Significance level**: probability of incorrectly rejecting null hypothesis (assuming null hypothesis)
Statistical Significance Testing

- Decide on a **null hypothesis**
- Decide on a **test statistic**
- Decide on a **threshold**

**Significance level**: probability of incorrectly rejecting null hypothesis (assuming null hypothesis)

**Power**: probability of correctly rejecting null hypothesis (assuming alternative hypothesis)
Parametric Tests

- Test statistic follows known distribution (with known parameters)

Paired Student's t-test:

- Paired samples (test datapoints)
- Scores normally distributed
- Null hypothesis: same mean

\[ t = \frac{\bar{D}}{\sigma_D} \sqrt{n} \]

"Student's t-distribution with \( n - 1 \) degrees of freedom"
Parametric Tests

- Test statistic follows known distribution (with known parameters)

- Paired Student’s t-test:
  - Paired samples (test datapoints)
  - Scores normally distributed
  - Null hypothesis: same mean
Parametric Tests

- Test statistic follows known distribution (with known parameters)

- Paired Student’s t-test:
  - Paired samples (test datapoints)
  - Scores normally distributed
  - Null hypothesis: same mean
  - Test statistic: \( t = \frac{\sqrt{n}}{s_D} \bar{X}_D \)
Parametric Tests

- Test statistic follows known distribution (with known parameters)

- Paired Student’s t-test:
  - Paired samples (test datapoints)
  - Scores normally distributed
  - Null hypothesis: same mean
  - Test statistic: \( t = \frac{\sqrt{n}}{s_D} \bar{X}_D \)
  - “Student’s t-distribution with \( n-1 \) degrees of freedom”
Nonparametric Tests

- No assumptions about distribution

- Sign test:
  - Paired samples (test datapoints)
  - System A better or system B better
  - Null hypothesis: equal chance
  - Test statistic: $n$
Nonparametric Tests

- No assumptions about distribution

- Sign test:
  - Paired samples (test datapoints)
  - System A better or system B better
  - Null hypothesis: equal chance

Test statistic: $n$
Nonparametric Tests

- No assumptions about distribution

- Sign test:
  - Paired samples (test datapoints)
  - System A better or system B better
  - Null hypothesis: equal chance
  - Test statistic: $n$
Nonparametric Tests

- No assumptions about distribution

- Sign test:
  - Paired samples (test datapoints)
  - System A better or system B better
  - Null hypothesis: equal chance
  - Test statistic: $n$
  - Binominal distribution
Multiple Tests

- If we test many systems, we expect some will pass
If we test many systems, we expect some will pass

Bonferroni correction:
- Replace nominal significance level $\alpha$ with $\frac{\alpha}{m}$
Base Rate Fallacy

- Evaluate 1000 systems
  - 900 similar to baseline
  - 100 better than baseline
Base Rate Fallacy

- Evaluate 1000 systems
  - 900 similar to baseline
  - 100 better than baseline

- Perform statistical test
  - Significance level: 5%
  - Power: 80%
Base Rate Fallacy

- Evaluate 1000 systems
  - 900 similar to baseline
  - 100 better than baseline

- Perform statistical test
  - Significance level: 5% → 45 pass
  - Power: 80% → 80 pass
Base Rate Fallacy

- Evaluate 1000 systems
  - 900 similar to baseline
  - 100 better than baseline

- Perform statistical test
  - Significance level: 5% → 45 pass
  - Power: 80% → 80 pass

- Probability system is better, given it passed the test: 64%
Base Rate Fallacy

- Evaluate 1000 systems
  - 960 similar to baseline
  - 40 better than baseline

- Perform statistical test
  - Significance level: 5% → 48 pass
  - Power: 80% → 32 pass

- Probability system is better, given it passed the test: 40%
Base Rate Fallacy

- Evaluate 1000 systems
  - 1000 similar to baseline
  - 0 better than baseline

- Perform statistical test
  - Significance level: 5% → 50 pass
  - Power: 80% → 0 pass

- Probability system is better, given it passed the test: 0%
A significant difference may not be a large difference
Effect Size

- A significant difference may not be a large difference
- e.g. a coin toss
  - Coins not perfectly symmetric
  - Probability of heads not exactly 50%
  - Difference so small we don’t care
Publication Bias

- Hard to publish negative results...
Publication Bias

- Hard to publish negative results...
- Authors may hide failed experiments
Publication Bias

- Hard to publish negative results...
- Authors may hide failed experiments
- MPhil project and L101 mini-project: Don’t hide! Negative results are okay!
Summary of Significance Testing

- Significance testing is important but underused in NLP!

- Choice of test:
  - Parametric (e.g. paired Student’s t-test)
  - Nonparametric (e.g. sign test)
  - Multiple tests (e.g. Bonferroni correction)

- Be careful:
  - Base rate fallacy
  - Effect size
  - Publication bias
Back to the Beginning...

- Task
- Data
- Model
- Training

Real-world application? 13
Back to the Beginning...

- Task
- Data
- Model
- Training

Most NLP papers
Back to the Beginning...

- Task
- Data
- Model
- Training

What if this goes wrong?

Most NLP papers
Back to the Beginning...

- Task
- Data
- Model
- Training

What if this goes wrong?

Most NLP papers

Real-world application?
Task: Predict death from pneumonia
Task: Predict death from pneumonia

Pattern in data: asthma reduces risk
Caruana et al. (2015)

- Task: Predict death from pneumonia
- Pattern in data: asthma reduces risk
- Real reason: asthma patients sent to Intensive Care Unit, reducing risk
Task: Predict death from pneumonia

Pattern in data: asthma reduces risk

Real reason: asthma patients sent to Intensive Care Unit, reducing risk

Shallow models (e.g. logistic regression) → can identify and fix such problems
Bias

- Bias (statistics): expected value differs from true value
- Bias (law): unfair or undesirable prejudice
Bias

“Bias is a social issue first, and a technical issue second.”

(Crawford, 2017)
Demographic Bias

- Region
- Social Class
- Gender
- Age
- Ethnicity
Hovy and Søgaard (2015)

- POS-tagging
Hovy and Søgaard (2015)

- POS-tagging

- Training data:
  - Wall Street Journal (English)
  - Frankfurter Rundschau (German)
Hovy and Søgaard (2015)

- POS-tagging

- Training data:
  - Wall Street Journal (English)
  - Frankfurter Rundschau (German)

- Test data:
  - Trustpilot reviews
  - Age, gender, location
### H&S (2015) – German Results

<table>
<thead>
<tr>
<th>Group</th>
<th>TreeT</th>
<th>CRF++</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under 35</td>
<td>.874</td>
<td>.859</td>
</tr>
<tr>
<td>Over 45</td>
<td><strong>.894</strong></td>
<td>.870</td>
</tr>
<tr>
<td>Men</td>
<td>.885</td>
<td>.861</td>
</tr>
<tr>
<td>Women</td>
<td>.882</td>
<td><strong>.868</strong></td>
</tr>
<tr>
<td>Highest-prob region</td>
<td>.885</td>
<td>.865</td>
</tr>
<tr>
<td>Lowest-prob region</td>
<td>.889</td>
<td>.874</td>
</tr>
</tbody>
</table>
## H&S (2015) – English Results

<table>
<thead>
<tr>
<th>Group</th>
<th>TreeT</th>
<th>CRF++</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under 35</td>
<td>.879</td>
<td>.882</td>
</tr>
<tr>
<td>Over 45</td>
<td><strong>.883</strong></td>
<td><strong>.884</strong></td>
</tr>
<tr>
<td>Men</td>
<td>.882</td>
<td>.886</td>
</tr>
<tr>
<td>Women</td>
<td>.880</td>
<td>.881</td>
</tr>
<tr>
<td>Highest-prob region</td>
<td>.883</td>
<td>.886</td>
</tr>
<tr>
<td>Lowest-prob region</td>
<td>.882</td>
<td>.885</td>
</tr>
</tbody>
</table>
POS-tagging on Twitter data

<table>
<thead>
<tr>
<th>Group</th>
<th>Stanf.</th>
<th>Gate</th>
<th>Ark</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAVE</td>
<td>.614</td>
<td>.791</td>
<td>.775</td>
</tr>
<tr>
<td>non-AAVE</td>
<td>.745</td>
<td>.833</td>
<td>.779</td>
</tr>
</tbody>
</table>
Corpora reflect social biases:

- Uncontroversial (e.g. pleasant/unpleasant association with flowers, insects, etc.)
- Prejudiced (e.g. pleasant/unpleasant association with gender, ethnicity, etc.)
- Status quo (e.g. association between gender and career)
Corpora reflect social biases:

- Uncontroversial (e.g. pleasant/unpleasant association with flowers, insects, etc.)
- Prejudiced (e.g. pleasant/unpleasant association with gender, ethnicity, etc.)
- Status quo (e.g. association between gender and career)

Distributional semantic vectors reflect social biases
The Guardian (2017):
“Computer says no: Irish vet fails oral English test needed to stay in Australia”
Decision Making


- Bias in training data vs. bias in decisions
Summary of Bias and Ethics

- Social bias (not statistical bias)
  - Training data
  - Model predictions
- POS-tagging & demographic groups
- Distributional semantics & associations
Course Summary

- Naive Bayes, Topic Classification
- HMM, POS-Tagging
- Logistic Regression, MEMM, NER
- Decision Boundaries, SVM, Kernels
- K-Means, LDA, WSI, Topic Discovery
- Distributional Semantics
- CNN, RNN, Hyperparameter Tuning
- Statistical Significance, Social Bias
Still To Come

- Last 3 sessions – reading seminars
- Mini-project