Why Model Validation?

• We have seen so far
  – Various metrics (e.g., accuracy, F-measure, RMSE, …)
  – Evaluation protocol setups
    • Split Validation
    • Cross Validation
    • Special protocols for time series
    • …

• Today
  – A closer look at evaluation protocols
  – Asking for significance
Some Observations

• Data Mining Competitions often have a hidden test set
  – e.g., Data Mining Cup
  – e.g., many tasks on Kaggle

• Ranking on public test set and ranking on hidden test set may differ

• Example on one Kaggle competition:

https://www.kaggle.com/c/restaurant-revenue-prediction/discussion/14026
Some Observations: DMC 2018

- We had eight teams in Mannheim
- We submitted the results of the best and the third best(!) team
- The third best team(!!!) got among the top 10
  - and eventually scored 2\textsuperscript{nd} worldwide
- Meanwhile, the best local team did not get among the top 10
What is Happening Here?

• We have come across this problem quite a few times
• It’s called *overfitting*
  – Problem: we don’t know the error on the (hidden) test set

but according to the test set, we should have used that one

according to the training dataset, this model is the best one

Overfitting Revisited

• Typical DMC Setup:

  ![](Training Data) ![Test Data]

  we often simulate test data by split or cross validation

• Possible overfitting scenarios:
  – our test partition may have certain characteristics
  – the “official” test data has different characteristics than the training data
Overfitting Revisited

• Typical Kaggle Setup:

- Training Data
- Test Data

  undisclosed part of the test data used for private leaderboard

• Possible overfitting scenarios:
  - solutions yielding good rankings on public leaderboard are preferred
  - models overfit to the public part of the test data
Overfitting Revisited

• Some flavors of overfitting are more subtle than others
• Obvious overfitting:
  – use test partition for training
• Less obvious overfitting:
  – tune parameters against test partition
  – select “best” approach based on test partition
• Even less obvious overfitting
  – use test partition in feature construction, for features such as
    • avg. sales of product per day
    • avg. orders by customer
    • computing trends
Overfitting Revisited

• Typical real world scenario:

Data from the past | The future (no data)

we often simulate test data
by split or cross validation

• Possible overfitting scenarios:
  – Similar to the DMC case, but worse
  – We do not even know the data on which we want to predict
What Unlabeled Test Data can Tell Us

• If we have test data without labels, we can still look at predictions
  – do they look somehow reasonable?

• Task of DMC 2018: predict date of the month in which a product is sold out
  – Solutions for three best (local) solutions:
The Overtuning Problem

- In academia
  - many fields have their established benchmarks
  - achieving outstanding scores on those is required for publication
  - interesting novel ideas may score suboptimally
    - hence, they are not published
  - intensive tuning is required for publication
    - hence, available compute often beats good ideas
The Overtuning Problem

- In real world projects
  - models overfit to past data
  - performance on unseen data is often overestimated
    - i.e., customers are disappointed
  - changing characteristics in data may be problematic
    - drift: e.g., predicting battery lifecycles
    - events not in training data: e.g., predicting sales for next month
  - cold start problem
    - some instances in the test set may be unknown before
    - e.g., predicting product sales for new products
Validating and Comparing Models

• When is a model good?
  – i.e., is it better than random?

• When is a model really better than another one?
  – i.e., is the performance difference by chance or by design?

Some of the following contents are taken from William W. Cohen’s Machine Learning Classes

http://www.cs.cmu.edu/~wcohen/
Confidence Intervals for Models

• Scenario:
  – you have learned a model M1 with an error rate of 0.30
  – the old model M0 had an error rate of 0.35
    (both evaluated on the same test set T)

• Do you think the new model is better?

• What might be suitable indicators?
  – size of the test set
  – model complexity
  – model variance
Size of the Test Set

• Scenario:
  – you have learned a model M1 with an error rate of 0.30
  – the old model M0 had an error rate of 0.35
  (both evaluated on the same test set S)

• Variant A: $|S| = 40$
  – a single error contributes 0.025 to the error rate
  – i.e., M1 got two more example right than M0

• Variant B: $|S| = 2,000$
  – a single error contributes 0.0005 to the error rate
  – i.e., M1 got 100 more examples right than M0
Size of the Test Set

• Scenario:
  – you have learned a model M1 with an error rate of 0.30
  – the old model M0 had an error rate of 0.35
    (both evaluated on the same test set T)

• Intuitively:
  – M1 is better if the error is observed on a larger test set T
  – The smaller the difference in the error, the larger |T| should be

• Can we formalize our intuitions?
What is an Error?

• Ultimately, we want to minimize the error on unseen data (D)
  - but we cannot measure it directly
• As a proxy, we use a sample S
  - in the best case: \( \text{error}_S = \text{error}_D \leftrightarrow |\text{error}_S - \text{error}_D| = 0 \)
  - or, more precisely: \( \mathbb{E}[|\text{error}_S - \text{error}_D|] = 0 \) for each S
• In many cases, our models are overly optimistic
  - i.e., \( \text{error}_D - \text{error}_S > 0 \)
What is an Error?

- In many cases, our models are overly optimistic
  - i.e., \( \text{error}_D - \text{error}_S > 0 \)

- Most often, the model has overfit to \( S \)

- Possible reasons:
  - \( S \) is a subset of training data (drastic)
  - \( S \) has been used in feature engineering and/or parameter tuning
  - we have trained and tuned three models only on \( T \), and pick the one which is best on \( S \)
What is an Error?

- Ultimately, we want to minimize the error on unseen data (D) but we cannot measure it directly.
- As a proxy, we use a sample S
  - unbiased model: \( E[|\text{error}_D - \text{error}_S|] = 0 \) for each S
- Even for an unbiased model, there is usually some variance given S
  - i.e. \( E[(\text{error}_S - E[\text{error}_S])^2] > 0 \)
  - intuitively: we measure (slightly) different errors on different S
Back to our Example

• Scenario:
  – you have learned a model M1 with an error rate of 0.30
  – the old model M0 had an error rate of 0.35
    (both evaluated on the same test set T)

• Old question:
  – is M1 better than M0?

• New question:
  – how likely is it the error of M1 is lower just by chance?
    • either: due to bias in M1, or due to variance
Back to our Example

• New question:
  – how likely is it the error of M1 is lower *just by chance*?
    • either: due to bias in M1, or due to variance

• Consider this a random process:
  – M1 makes an error on example x
  – Let us assume it actually has an error rate of 0.3
    • i.e., M1 follows a binomial with its maximum at 0.3

• Test:
  – what is the probability of actually observing 0.3 or 0.35 as error rates?
Binomial Distribution for M1

• We can easily construct those binomial distributions given n and p

\[ P(r) = \frac{n!}{r!(n-r)!} error_D(h)^r (1 - error_D(h))^{n-r} \]

Probability of observing an error of 0.3 (12/40): 0.137

Probability of observing an error of 0.35 (14/40): 0.104
From the Binomial to Confidence Intervals

- New question:
  - what values are we likely to observe? (e.g., with a probability of 95%)
  - i.e., we look at the symmetric interval around the mean that covers 95%

```
P(r) = \frac{n!}{r!(n-r)!} error_D(h)^r (1 - error_D(h))^{n-r}
```
• With a probability of 95%, we observe 7 to 17 errors
  – corresponds to $[0.175 ; 0.425]$ as a confidence interval

• All observations in that interval are considered likely
  – i.e., an observed error rate of 0.35 might also correspond to an actual error rate of 0.3

• Back to our example
  – on a test sample of $|S|=40$, we cannot say whether M1 or M0 is better
Simplified Calculation (z Test)

• The central limit theorem states that
  – a binomial distribution can be approximated by a Gaussian normal distribution
    • with \( \mu = np \), \( \sigma = \sqrt{\frac{p(1-p)}{n}} \)
  – for sufficiently large \( n \)
    • rule of thumb: sufficiently large equals \( n > 30 \)

\[ p \text{ in our case: error} \]
Simplified Calculation (z Test)

- The central limit theorem states that
  - a binomial distribution can be approximated by a Gaussian normal distribution
  - Gaussian distributions are simple to compute

80% of area (probability) lies in $\mu \pm 1.28\sigma$

N% of area (probability) lies in $\mu \pm z_N\sigma$

<table>
<thead>
<tr>
<th>N%</th>
<th>50%</th>
<th>68%</th>
<th>80%</th>
<th>90%</th>
<th>95%</th>
<th>98%</th>
<th>99%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$z_N$</td>
<td>0.67</td>
<td>1.00</td>
<td>1.28</td>
<td>1.64</td>
<td>1.96</td>
<td>2.33</td>
<td>2.58</td>
</tr>
</tbody>
</table>
Simplified Confidence Intervals

• Given that we have |S|=n, and an observed error\(_S\)  
  – With p\% probability, error\(_D\) is in [error\(_S\) – y, error\(_S\) + y]  
  – With y= \( z_N \sqrt{\frac{error_S(1-error_S)}{n}} \)

<table>
<thead>
<tr>
<th>N%:</th>
<th>50%</th>
<th>68%</th>
<th>80%</th>
<th>90%</th>
<th>95%</th>
<th>98%</th>
<th>99%</th>
</tr>
</thead>
<tbody>
<tr>
<td>( z_N ):</td>
<td>0.67</td>
<td>1.00</td>
<td>1.28</td>
<td>1.64</td>
<td>1.96</td>
<td>2.33</td>
<td>2.58</td>
</tr>
</tbody>
</table>

• Given our example
  – error\(_S\) = 0.30, n=40  
  \( \rightarrow \) with 95\% probability, error\(_D\) is in [0.158, 0.442]
Working with Confidence Intervals

• Given that we have $|S|=n$, and an observed error $\bar{s}$
  – With p% probability, error $D$ is in $[error_s - y, error_s + y]$
  – With $y = \frac{z_{\frac{p}{2}} \sqrt{error_s (1-error_s)}}{n}$

Observation: the interval shrinks with growing n

• Recap: we had two scenarios, $|S| = 40$ and $|S| = 2000$
  – Interval for n=40: error $D$ is in [0.158, 0.442]
  – Interval for n=2000: error $D$ is in [0.280, 0.320]

• So, for $|S|=2000$, the probability that error $D$ is lower than 0.35 is >95%
Working with Confidence Intervals

• Comparing M0 and M1

|S|=40

|S|=2000

• For |S|=2000, the confidence intervals do not overlap
  – i.e., with 95% probability, M1 is better than M0
  – but we cannot make such a statement for |S|=40
Occam's Razor Revisited

- Named after William of Ockham (1287-1347)
- A fundamental principle of science
  - if you have two theories
  - that explain a phenomenon equally well
  - choose the simpler one

- Example:
  - phenomenon: the street is wet
  - theory 1: it has rained
  - theory 2: a beer truck has had an accident, and beer has spilled. The truck has been towed, and magpies picked the glass pieces, so only the beer remains
Occam's Razor Revisited

• Let’s rephrase:
  – if you have two models
  – where none is *significantly* better than the other
  – choose the simpler one

• Indicators for simplicity:
  – less features used
  – less variables used
    • hidden neurons in an ANN
    • no. of trees in a Random Forest
    • ...

Heiko Paulheim
Model Variance

• What happens if you repeat an experiment...
  – ...on a different test set?
  – ...on a different training set?
  – ...with a different random seed?

• Some methods may have higher \textit{variance} than others
  – if your result was good, was just luck?
  – what is your actual estimate for the future?

• Typically, we need more than one experiment!
Model Variance

• Scenario:
  – you have learned a model M1 with an error rate of 0.30
  – the old model M0 had an error rate of 0.35
    (this time: in 10-fold cross validation)

• Variant A:
  – M0:
    | F1  | F2  | F3  | F4  | F5  | F6  | F7  | F8  | F9  | F10 | Ø  |
    | 0.37 | 0.28 | 0.38 | 0.40 | 0.27 | 0.42 | 0.26 | 0.39 | 0.41 | 0.29 | 0.35 |

  – M1_A:
    | F1  | F2  | F3  | F4  | F5  | F6  | F7  | F8  | F9  | F10 | Ø  |
    | 0.28 | 0.30 | 0.31 | 0.32 | 0.25 | 0.32 | 0.27 | 0.32 | 0.33 | 0.30 | 0.30 |
Model Variance

- Scenario:
  - you have learned a model M1 with an error rate of 0.30
  - the old model M0 had an error rate of 0.35
    (this time: in 10-fold cross validation)

- Variant B:
  - M0:
    - F1: 0.37
    - F2: 0.28
    - F3: 0.38
    - F4: 0.40
    - F5: 0.27
    - F6: 0.42
    - F7: 0.26
    - F8: 0.39
    - F9: 0.41
    - F10: 0.29
    - Ø: 0.35
  - M1\(_B\):
    - F1: 0.17
    - F2: 0.29
    - F3: 0.18
    - F4: 0.53
    - F5: 0.28
    - F6: 0.49
    - F7: 0.27
    - F8: 0.29
    - F9: 0.19
    - F10: 0.31
    - Ø: 0.30

lucky shots

total fails
Model Variance

- **M0:**
  
<table>
<thead>
<tr>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
<th>F5</th>
<th>F6</th>
<th>F7</th>
<th>F8</th>
<th>F9</th>
<th>F10</th>
<th>Ø</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.37</td>
<td>0.28</td>
<td>0.38</td>
<td>0.40</td>
<td>0.27</td>
<td>0.42</td>
<td>0.26</td>
<td>0.39</td>
<td>0.41</td>
<td>0.29</td>
<td>0.35</td>
</tr>
</tbody>
</table>

- **M1_A:**
  
<table>
<thead>
<tr>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
<th>F5</th>
<th>F6</th>
<th>F7</th>
<th>F8</th>
<th>F9</th>
<th>F10</th>
<th>Ø</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.28</td>
<td>0.30</td>
<td>0.31</td>
<td>0.32</td>
<td>0.25</td>
<td>0.32</td>
<td>0.27</td>
<td>0.32</td>
<td>0.33</td>
<td>0.30</td>
<td>0.30</td>
</tr>
</tbody>
</table>

- **M1_B:**
  
<table>
<thead>
<tr>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
<th>F5</th>
<th>F6</th>
<th>F7</th>
<th>F8</th>
<th>F9</th>
<th>F10</th>
<th>Ø</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.17</td>
<td>0.29</td>
<td>0.18</td>
<td>0.53</td>
<td>0.28</td>
<td>0.49</td>
<td>0.27</td>
<td>0.29</td>
<td>0.19</td>
<td>0.31</td>
<td>0.30</td>
</tr>
</tbody>
</table>

- **Some observations:**
  - Standard deviations (M0: 0.06, M1_A: 0.03, M1_B: 0.12)
  - Pairwise competition:
    - M1_A outperforms M0 in 7/10 cases
    - but: M0 also outperforms M1_B in 6/10 cases!
  - Worst case of M1_A is below that of M0, but worst case of M1_B is above
Model Variance

• Why is model variance important?
  – recap: confidence intervals
  – risk vs. gain (use case!)
  – often, training data differs
    • even if you use cross or split validation during development
    • you might still train a model on the entire training data later
General Comparison of Methods

• Practice: finding a good method for a given problem
• Research: finding a good method for a class of problems

https://xkcd.com/664/
General Comparison of Methods

• Practice: finding a good method for a given problem
• Research: finding a good method for a \textit{class of problems}

• Typical research paper:
  – Method M is better than state of the art S on a problem class P
  – Evaluation: show results of M on a subset of P
  – Claim that M is significantly better than S

let's look closer
General Comparison of Methods

• De facto gold standard paper: Demšar, 2006
  – >8,000 citations on Google scholar
  – one of the most cited papers in JMLR in general

Statistical Comparisons of Classifiers over Multiple Data Sets

Abstract

While methods for comparing two learning algorithms on a single data set have been scrutinized for quite some time already, the issue of statistical tests for comparisons of more algorithms on multiple data sets, which is even more essential to typical machine learning studies, has been all but ignored. This article reviews the current practice and then theoretically and empirically examines several suitable tests. Based on that, we recommend a set of simple, yet safe and robust non-parametric tests for statistical comparisons of classifiers: the Wilcoxon signed ranks test for comparison of two classifiers and the Friedman test with the corresponding post-hoc tests for comparison of more classifiers over multiple data sets. Results of the latter can also be neatly presented with the newly introduced CD (critical difference) diagrams.

Keywords: comparative studies, statistical methods, Wilcoxon signed-ranks test, Friedman test, multiple comparisons tests
Example

• New Method M vs. State of the Art Method S
  – Tested on 12 different problems
  – Depicted: error rate

• Observations:
  – error rate alone might not be telling
  – problems are not directly comparable

<table>
<thead>
<tr>
<th>Problem</th>
<th>M</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.09</td>
<td>0.11</td>
</tr>
<tr>
<td>2</td>
<td>0.71</td>
<td>0.72</td>
</tr>
<tr>
<td>3</td>
<td>0.77</td>
<td>0.69</td>
</tr>
<tr>
<td>4</td>
<td>0.21</td>
<td>0.44</td>
</tr>
<tr>
<td>5</td>
<td>0.37</td>
<td>0.37</td>
</tr>
<tr>
<td>6</td>
<td>0.85</td>
<td>0.92</td>
</tr>
<tr>
<td>7</td>
<td>0.62</td>
<td>0.65</td>
</tr>
<tr>
<td>8</td>
<td>0.58</td>
<td>0.55</td>
</tr>
<tr>
<td>9</td>
<td>0.79</td>
<td>0.89</td>
</tr>
<tr>
<td>10</td>
<td>0.12</td>
<td>0.16</td>
</tr>
<tr>
<td>11</td>
<td>0.09</td>
<td>0.15</td>
</tr>
<tr>
<td>12</td>
<td>0.19</td>
<td>0.24</td>
</tr>
<tr>
<td>Avg.</td>
<td>0.45</td>
<td>0.49</td>
</tr>
</tbody>
</table>

simpler problem

harder problem
Example

• Observation:
  – 9 times: M outperforms S
  – 2 times: S outperforms M
  – 1 tie

• Just looking at those outcomes
  – Null hypothesis: M and S are equally good
    • i.e., probability of M outperforming S is 0.5
  – What is the likelihood of M outperforming S in 9 or more out of 11 cases?
    • analogy: what is the likelihood of 9 or more heads in 11 coin tosses?
      → known as sign test
Example

• We’ve already seen something similar
  – what is the likelihood of that outcome (9/11 wins for M) by chance?
  – let’s look at confidence intervals

<table>
<thead>
<tr>
<th>Problem</th>
<th>M</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.09</td>
<td>0.11</td>
</tr>
<tr>
<td>2</td>
<td>0.71</td>
<td>0.72</td>
</tr>
<tr>
<td>3</td>
<td>0.77</td>
<td>0.69</td>
</tr>
<tr>
<td>4</td>
<td>0.21</td>
<td>0.44</td>
</tr>
<tr>
<td>5</td>
<td>0.37</td>
<td>0.37</td>
</tr>
<tr>
<td>6</td>
<td>0.85</td>
<td>0.92</td>
</tr>
<tr>
<td>7</td>
<td>0.62</td>
<td>0.65</td>
</tr>
<tr>
<td>8</td>
<td>0.58</td>
<td>0.55</td>
</tr>
<tr>
<td>9</td>
<td>0.79</td>
<td>0.89</td>
</tr>
<tr>
<td>10</td>
<td>0.12</td>
<td>0.16</td>
</tr>
<tr>
<td>11</td>
<td>0.09</td>
<td>0.15</td>
</tr>
<tr>
<td>12</td>
<td>0.19</td>
<td>0.24</td>
</tr>
<tr>
<td>Avg.</td>
<td>0.45</td>
<td>0.49</td>
</tr>
</tbody>
</table>

- M wins: \( \frac{9}{11} \pm 1.96 \sqrt{\frac{\frac{9}{11} \cdot \left(1 - \frac{9}{11}\right)}{11}} \rightarrow [0.70, 0.93] \)

- S wins: \( \frac{2}{11} \pm 1.96 \sqrt{\frac{\frac{2}{11} \cdot \left(1 - \frac{2}{11}\right)}{11}} \rightarrow [0.07, 0.30] \)

• Looks safe, but... \( n<30! \)
Example

- Observation:
  - 9 times: M outperforms S
  - 2 times: S outperforms M
  - 1 tie

- Just looking at those outcomes
  - Null hypothesis: M and S are equally good
    - i.e., probability of M outperforming S is 0.5
  - What is the likelihood of M outperforming S in 9 or more out of 11 cases?
    - analogy: what is the likelihood of 9 or more heads in 11 coin tosses?
    - Here: 0.03
      → i.e., with a probability >0.95, this is not an outcome by chance
Sign Test

• Observation:
  – 9 times: M outperforms S
  – 2 times: S outperforms M
  – 1 tie

• Sign test looks at those outcomes as binary experiments
  – null hypothesis: M is not better than S, i.e., M outperforming S is as likely as M not outperforming S

<table>
<thead>
<tr>
<th>Problem</th>
<th>M</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.09</td>
<td>0.11</td>
</tr>
<tr>
<td>2</td>
<td>0.71</td>
<td>0.72</td>
</tr>
<tr>
<td>3</td>
<td>0.77</td>
<td>0.69</td>
</tr>
<tr>
<td>4</td>
<td>0.21</td>
<td>0.44</td>
</tr>
<tr>
<td>5</td>
<td>0.37</td>
<td>0.37</td>
</tr>
<tr>
<td>6</td>
<td>0.85</td>
<td>0.92</td>
</tr>
<tr>
<td>7</td>
<td>0.62</td>
<td>0.65</td>
</tr>
<tr>
<td>8</td>
<td>0.58</td>
<td>0.55</td>
</tr>
<tr>
<td>9</td>
<td>0.79</td>
<td>0.89</td>
</tr>
<tr>
<td>10</td>
<td>0.12</td>
<td>0.16</td>
</tr>
<tr>
<td>11</td>
<td>0.09</td>
<td>0.15</td>
</tr>
<tr>
<td>12</td>
<td>0.19</td>
<td>0.24</td>
</tr>
<tr>
<td><strong>Avg.</strong></td>
<td><strong>0.45</strong></td>
<td><strong>0.49</strong></td>
</tr>
</tbody>
</table>

Table 3: Critical values for the two-tailed sign test at $\alpha = 0.05$ and $\alpha = 0.10$. A classifier is significantly better than another if it performs better on at least $w_\alpha$ data sets.
Sign Test – Variants

• Some variations:
  – We used $N = \text{wins} + \text{losses}$ (standard sign test).
    Some use: $N = \text{wins} + \text{losses} + \text{ties}$

• With that variant, we would not conclude significance at $p < 0.05$

<table>
<thead>
<tr>
<th>Problem</th>
<th>$M$</th>
<th>$S$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.09</td>
<td>0.11</td>
</tr>
<tr>
<td>2</td>
<td>0.71</td>
<td>0.72</td>
</tr>
<tr>
<td>3</td>
<td>0.77</td>
<td>0.69</td>
</tr>
<tr>
<td>4</td>
<td>0.21</td>
<td>0.44</td>
</tr>
<tr>
<td>5</td>
<td>0.37</td>
<td>0.37</td>
</tr>
<tr>
<td>6</td>
<td>0.85</td>
<td>0.92</td>
</tr>
<tr>
<td>7</td>
<td>0.62</td>
<td>0.65</td>
</tr>
<tr>
<td>8</td>
<td>0.58</td>
<td>0.55</td>
</tr>
<tr>
<td>9</td>
<td>0.79</td>
<td>0.89</td>
</tr>
<tr>
<td>10</td>
<td>0.12</td>
<td>0.16</td>
</tr>
<tr>
<td>11</td>
<td>0.09</td>
<td>0.15</td>
</tr>
<tr>
<td>12</td>
<td>0.19</td>
<td>0.24</td>
</tr>
<tr>
<td>Avg.</td>
<td>0.45</td>
<td>0.49</td>
</tr>
</tbody>
</table>

Table 3: Critical values for the two-tailed sign test at $\alpha = 0.05$ and $\alpha = 0.10$. A classifier is significantly better than another if it performs better on at least $w_\alpha$ data sets.
Sign Test – Variants

• Observation: some wins/losses are rather marginal

• Stricter variant:
  – perform significance test for each dataset (as shown earlier today)
  – regard only significant wins/losses

• In our example:
  – Let’s assume the results on problem 1, 3, 4, 6, 7, 9, 10, 11, 12 are significant

<table>
<thead>
<tr>
<th>Problem</th>
<th>M</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.09</td>
<td>0.11</td>
</tr>
<tr>
<td>2</td>
<td>0.71</td>
<td>0.72</td>
</tr>
<tr>
<td>3</td>
<td>0.77</td>
<td>0.69</td>
</tr>
<tr>
<td>4</td>
<td>0.21</td>
<td>0.44</td>
</tr>
<tr>
<td>5</td>
<td>0.37</td>
<td>0.37</td>
</tr>
<tr>
<td>6</td>
<td>0.85</td>
<td>0.92</td>
</tr>
<tr>
<td>7</td>
<td>0.62</td>
<td>0.65</td>
</tr>
<tr>
<td>8</td>
<td>0.58</td>
<td>0.55</td>
</tr>
<tr>
<td>9</td>
<td>0.79</td>
<td>0.89</td>
</tr>
<tr>
<td>10</td>
<td>0.12</td>
<td>0.16</td>
</tr>
<tr>
<td>11</td>
<td>0.09</td>
<td>0.15</td>
</tr>
<tr>
<td>12</td>
<td>0.19</td>
<td>0.24</td>
</tr>
<tr>
<td>Avg.</td>
<td>0.45</td>
<td>0.49</td>
</tr>
</tbody>
</table>

Table 3: Critical values for the two-tailed sign test at α = 0.05 and α = 0.10. A classifier is significantly better than another if it performs better on at least \( w_\alpha \) data sets.
Wilcoxon Signed-Rank Test

- Observation: some wins/losses are rather marginal
- Wilcoxon Signed-Rank Test
  - takes margins into account

- Approach:
  - rank results by *absolute* difference
  - sum up ranks for positive and negative outcomes
    - best case: all outcomes positive $\rightarrow$ sum of negative ranks = 0
    - still good case: all negative outcomes are marginal $\rightarrow$ sum of negative ranks is low

<table>
<thead>
<tr>
<th>Problem</th>
<th>M</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.09</td>
<td>0.11</td>
</tr>
<tr>
<td>2</td>
<td>0.71</td>
<td>0.72</td>
</tr>
<tr>
<td>3</td>
<td>0.77</td>
<td>0.69</td>
</tr>
<tr>
<td>4</td>
<td>0.21</td>
<td>0.44</td>
</tr>
<tr>
<td>5</td>
<td>0.37</td>
<td>0.37</td>
</tr>
<tr>
<td>6</td>
<td>0.85</td>
<td>0.92</td>
</tr>
<tr>
<td>7</td>
<td>0.62</td>
<td>0.65</td>
</tr>
<tr>
<td>8</td>
<td>0.58</td>
<td>0.55</td>
</tr>
<tr>
<td>9</td>
<td>0.79</td>
<td>0.89</td>
</tr>
<tr>
<td>10</td>
<td>0.12</td>
<td>0.16</td>
</tr>
<tr>
<td>11</td>
<td>0.09</td>
<td>0.15</td>
</tr>
<tr>
<td>12</td>
<td>0.19</td>
<td>0.24</td>
</tr>
<tr>
<td>Avg.</td>
<td>0.45</td>
<td>0.49</td>
</tr>
</tbody>
</table>
Wilcoxon Signed-Rank Test

- Computation: rank results
  - sum up $R^-$ and $R^+$
  - ties are ignored
  - equal ranks are averaged

- $R^- = 11.5$, $R^+ = 54.5$

<table>
<thead>
<tr>
<th>Problem</th>
<th>M</th>
<th>S</th>
<th>Delta</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.09</td>
<td>0.11</td>
<td>-0.02</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>0.71</td>
<td>0.72</td>
<td>-0.01</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>0.77</td>
<td>0.69</td>
<td>0.08</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>0.21</td>
<td>0.44</td>
<td>-0.23</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0.37</td>
<td>0.37</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>6</td>
<td>0.85</td>
<td>0.92</td>
<td>-0.07</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>0.62</td>
<td>0.65</td>
<td>-0.03</td>
<td>8.5</td>
</tr>
<tr>
<td>8</td>
<td>0.58</td>
<td>0.55</td>
<td>0.03</td>
<td>8.5</td>
</tr>
<tr>
<td>9</td>
<td>0.79</td>
<td>0.89</td>
<td>-0.1</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>0.12</td>
<td>0.16</td>
<td>-0.04</td>
<td>7</td>
</tr>
<tr>
<td>11</td>
<td>0.09</td>
<td>0.15</td>
<td>-0.06</td>
<td>5</td>
</tr>
<tr>
<td>12</td>
<td>0.19</td>
<td>0.24</td>
<td>-0.05</td>
<td>6</td>
</tr>
<tr>
<td>Avg.</td>
<td>0.45</td>
<td>0.49</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Wilcoxon Signed-Rank Test

- Computation: rank results
  - sum up \( R^- \) and \( R^+ \)
  - ties are ignored
  - equal ranks are averaged

- \( R^- = 11.5, \ R^+ = 54.5 \)
- We use the one-tailed test
  - because we want to test if \( M \) is *better* than \( S \)
- \( 11.5 < 17 \)
  → the results are significant

\[
\begin{array}{cccccc}
\alpha_{\text{two-tailed}} & \leq 0.10 & \alpha_{\text{one-tailed}} & \leq 0.05 & \alpha_{\text{one-tailed}} & \leq 0.025 \\
5 & 0 & 0 & 0 & 0 & 0 \\
6 & 2 & 2 & 0 & 0 & 0 \\
7 & 3 & 3 & 0 & 0 & 0 \\
8 & 5 & 5 & 1 & 0 & 0 \\
9 & 8 & 8 & 3 & 1 & 1 \\
10 & 10 & 10 & 3 & 3 & 3 \\
11 & 13 & 13 & 5 & 5 & 5 \\
12 & 17 & 17 & 7 & 7 & 7 \\
13 & 21 & 21 & 9 & 9 & 9 \\
14 & 25 & 25 & 11 & 11 & 11 \\
15 & 30 & 30 & 13 & 13 & 13 \\
16 & 35 & 35 & 15 & 15 & 15 \\
17 & 41 & 41 & 17 & 17 & 17 \\
18 & 47 & 47 & 19 & 19 & 19 \\
19 & 53 & 53 & 21 & 21 & 21 \\
20 & 60 & 60 & 23 & 23 & 23 \\
21 & 67 & 67 & 25 & 25 & 25 \\
22 & 75 & 75 & 27 & 27 & 27 \\
23 & 83 & 83 & 29 & 29 & 29 \\
24 & 91 & 91 & 31 & 31 & 31 \\
25 & 100 & 100 & 33 & 33 & 33 \\
26 & 110 & 110 & 35 & 35 & 35 \\
27 & 119 & 119 & 37 & 37 & 37 \\
28 & 130 & 130 & 39 & 39 & 39 \\
29 & 140 & 140 & 41 & 41 & 41 \\
30 & 151 & 151 & 43 & 43 & 43 \\
\end{array}
\]

Tests for Comparing Approaches

• Summary
  – Simple z test only reliable for many datasets (>30)
  – Sign test does not distinguish large and small margins
  – Wilcoxon signed-rank test
    • works also for small samples (e.g., half a dozen datasets)
    • considers large and small margins
Take Aways

• Results in Data Mining are often reduced to a single number
  – e.g., accuracy, error rate, F1, RMSE
  – result differences are often marginal

• Problem of unseen data
  – we can only guess/approximate the true performance on unseen data
  – makes it hard to select between approaches

• Helpful tools
  – confidence intervals
  – significance tests
  – Occam’s Razor
What’s Next?

• The Data Mining Cup is up and running
  – From next week on, we’ll discuss your results together
    • We’ll be using ZOOM for that
  – Please use our custom ILIAS plugin (see yesterday’s e-mail)
    • please upload your best solution so far each week before the lecture slot
    • thanks for Beta Testing ;-)  
    • in case of problems, get in touch with Nico

• Final exam
  – we have no information yet
Further Offerings in the next Semester

• Machine Learning (Gemulla)
• Web Data Integration (Bizer)
• Relational Learning (Meilicke and Stuckenschmidt)
• Network Analysis (Hulpus and Stuckenschmidt)
• Text Analytics (Ponzetto and Colleagues)
• Image Processing (Keuper)
• Process Mining and Analysis (van der Aa & Rehse)
Questions?
Data Mining II
Model Validation