Slides for “Data Mining”
by
I. H. Witten and E. Frank
Predicting performance

- Assume the estimated error rate is 25%. How close is this to the true error rate?
  - Depends on the amount of test data
- Prediction is just like tossing a (biased!) coin
  - “Head” is a “success”, “tail” is an “error”
- In statistics, a succession of independent events like this is called a *Bernoulli process*
  - Statistical theory provides us with confidence intervals for the true underlying proportion
Confidence intervals

- We can say: \( p \) lies within a certain specified interval with a certain specified confidence

- Example: \( S=750 \) successes in \( N=1000 \) trials
  - Estimated success rate: 75%
  - How close is this to true success rate \( p \)?
    - Answer: with 80% confidence \( p \in [73.2, 76.7] \)

- Another example: \( S=75 \) and \( N=100 \)
  - Estimated success rate: 75%
  - With 80% confidence \( p \in [69.1, 80.1] \)
Mean and variance

- Mean and variance for a Bernoulli trial: $p, p(1-p)$
- Expected success rate $f=S/N$
- Mean and variance for $f$: $p, p(1-p)/N$
- For large enough $N$, $f$ follows a Normal distribution
- $c\%$ confidence interval $[-z \leq X \leq z]$ for random variable with 0 mean is given by:
  \[
  \Pr[-z \leq X \leq z] = c
  \]
- With a symmetric distribution:
  \[
  \Pr[-z \leq X \leq z] = 1 - 2 \times \Pr[X \geq z]
  \]
Confidence limits

- Confidence limits for the normal distribution with 0 mean and a variance of 1:

<table>
<thead>
<tr>
<th>Pr[X ≥ z]</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1%</td>
<td>3.09</td>
</tr>
<tr>
<td>0.5%</td>
<td>2.58</td>
</tr>
<tr>
<td>1%</td>
<td>2.33</td>
</tr>
<tr>
<td>5%</td>
<td>1.65</td>
</tr>
<tr>
<td>10%</td>
<td>1.28</td>
</tr>
<tr>
<td>20%</td>
<td>0.84</td>
</tr>
<tr>
<td>40%</td>
<td>0.25</td>
</tr>
</tbody>
</table>

- Thus:

  \[ \Pr[-1.65 \leq X \leq 1.65] = 90\% \]

- To use this we have to reduce our random variable \( f \) to have 0 mean and unit variance
Transforming $f$

- Transformed value for $f$:
  \[
  \frac{f - p}{\sqrt{p(1 - p)/N}}
  \]
  (i.e. subtract the mean and divide by the standard deviation)

- Resulting equation:
  \[
  \Pr\left[-z \leq \frac{f - p}{\sqrt{p(1 - p)/N}} \leq z\right] = c
  \]

- Solving for $p$:
  \[
  p = \left(f + \frac{z^2}{2N} \pm z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}}\right) \div \left(1 + \frac{z^2}{N}\right)
  \]
Examples

- $f = 75\%, N = 1000, c = 80\%$ (so that $z = 1.28$):
  \[ p \in [0.732, 0.767] \]

- $f = 75\%, N = 100, c = 80\%$ (so that $z = 1.28$):
  \[ p \in [0.691, 0.801] \]

- Note that normal distribution assumption is only valid for large $N$ (i.e. $N > 100$)

- $f = 75\%, N = 10, c = 80\%$ (so that $z = 1.28$):
  \[ p \in [0.549, 0.881] \]

  (should be taken with a grain of salt)
Comparing data mining schemes

- Frequent question: which of two learning schemes performs better?
- Note: this is domain dependent!
- Obvious way: compare 10-fold CV estimates
- Problem: variance in estimate
- Variance can be reduced using repeated CV
- However, we still don’t know whether the results are reliable
Significance tests

- Significance tests tell us how confident we can be that there really is a difference
- Null hypothesis: there is no “real” difference
- Alternative hypothesis: there is a difference
- A significance test measures how much evidence there is in favor of rejecting the null hypothesis
- Let’s say we are using 10-fold CV
- Question: do the two means of the 10 CV estimates differ significantly?
Paired t-test

- Student’s t-test tells whether the means of two samples are significantly different
- Take individual samples using cross-validation
- Use a paired t-test because the individual samples are paired
  - The same CV is applied twice

William Gosset
Born: 1876 in Canterbury; Died: 1937 in Beaconsfield, England
Obtained a post as a chemist in the Guinness brewery in Dublin in 1899. Invented the t-test to handle small samples for quality control in brewing. Wrote under the name "Student".
Distribution of the means

- $x_1, x_2, \ldots, x_k$ and $y_1, y_2, \ldots, y_k$ are the $2k$ samples for a $k$-fold CV
- $m_x$ and $m_y$ are the means
- With enough samples, the mean of a set of independent samples is normally distributed

- Estimated variances of the means are $\sigma_x^2/k$ and $\sigma_y^2/k$

- If $\mu_x$ and $\mu_y$ are the true means then
  \[
  \frac{m_x - \mu_x}{\sqrt{\sigma_x^2 / k}} \quad \frac{m_y - \mu_y}{\sqrt{\sigma_y^2 / k}}
  \]
  are approximately normally distributed with mean 0, variance 1
Student’s distribution

- With small samples \((k < 100)\) the mean follows \textit{Student’s distribution with \(k-1\) degrees of freedom}

- Confidence limits:

<table>
<thead>
<tr>
<th>Pr([X \geq z])</th>
<th>(z)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1%</td>
<td>4.30</td>
</tr>
<tr>
<td>0.5%</td>
<td>3.25</td>
</tr>
<tr>
<td>1%</td>
<td>2.82</td>
</tr>
<tr>
<td>5%</td>
<td>1.83</td>
</tr>
<tr>
<td>10%</td>
<td>1.38</td>
</tr>
<tr>
<td>20%</td>
<td>0.88</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pr([X \geq z])</th>
<th>(z)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1%</td>
<td>3.09</td>
</tr>
<tr>
<td>0.5%</td>
<td>2.58</td>
</tr>
<tr>
<td>1%</td>
<td>2.33</td>
</tr>
<tr>
<td>5%</td>
<td>1.65</td>
</tr>
<tr>
<td>10%</td>
<td>1.28</td>
</tr>
<tr>
<td>20%</td>
<td>0.84</td>
</tr>
</tbody>
</table>
Distribution of the differences

- Let $m_d = m_x - m_y$
- The difference of the means ($m_d$) also has a Student’s distribution with $k-1$ degrees of freedom
- Let $\sigma_d^2$ be the variance of the difference
- The standardized version of $m_d$ is called the $t$-statistic:

$$t = \frac{m_d}{\sqrt{\sigma_d^2 / k}}$$

- We use $t$ to perform the $t$-test
Performing the test

• Fix a significance level $\alpha$
  • If a difference is significant at the $\alpha\%$ level, there is a $(100-\alpha)\%$ chance that there really is a difference

• Divide the significance level by two because the test is two-tailed
  • I.e. the true difference can be $+$ve or $-$ve

• Look up the value for $z$ that corresponds to $\alpha/2$

• If $t \leq -z$ or $t \geq z$ then the difference is significant
  • I.e. the null hypothesis can be rejected
Unpaired observations

- If the CV estimates are from different randomizations, they are no longer paired
- (or maybe we used $k$-fold CV for one scheme, and $j$-fold CV for the other one)
- Then we have to use an *un* paired t-test with $\min(k, j) - 1$ degrees of freedom
- The $t$-statistic becomes:

$$t = \frac{m_d}{\sqrt{\frac{\sigma_d^2}{k}}} \quad \Rightarrow \quad t = \frac{m_x - m_y}{\sqrt{\frac{\sigma_x^2}{k} + \frac{\sigma_y^2}{j}}}$$