How do we compare the relative performance among competing models?

Comparing Data Mining Methods

- Frequent problem: we want to know which of the two learning techniques is better
  - How to reliably say Model A is better or worse than Model B?
- We can:
  - Compare on different test sets
  - Compare 10-fold CV estimates
- Both require significance testing.

Significance Tests

- Significance tests tell us how (statistically) confident we can be that there is truly a difference.
- For example:
  - 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

Methods for Comparing Classifiers

- Two models:
  - Model M1: accuracy = 85%, tested on 30 instances
  - Model M2: accuracy = 75%, tested on 5,000 instances
- Can we say M1 is better than M2?
- How much confidence can have in the accuracy of both models?
- Can the difference in performance measure be explained as a result of random fluctuations in the test set?

Confidence Intervals

- We can say: error lies within a certain specified interval within a certain specified confidence
- Example: $S = 750$ successes in $n = 1000$ test examples
- Estimated error rate: 25%
- How close is this to the true error rate?
- With 95% confidence [22.32,27.68]

Confidence Interval for Accuracy

- Prediction can be regarded as a Bernoulli trial with two possible outcomes, correct or incorrect.
- A collection of Bernoulli trials has a Binomial distribution.
- Given the number of correct test predictions x and the number of test instances N, accuracy $acc = x/N$.
- Can we predict the true accuracy of the model from acc?

Confidence Interval for Accuracy

- For large test sets ($N > 30$), the accuracy acc has a normal distribution with mean p and variance $p(1-p)/N$.
- Confidence interval for p is:

$$P\left(Z_{\alpha/2} < \frac{acc - p}{\sqrt{p(1-p)/N}} < Z_{1-\alpha/2}\right)$$
$$= 1 - \alpha$$

Confidence Interval for Accuracy

- Consider a testing set containing 1,000 examples ($N = 1000$).
- 750 examples have been correctly classified ($x = 750$, $acc = 75\%$).
- If we want an 80% confidence level, then the true performance p is between 73.2% and 76.7%.
- If we only have 100 training examples and 75 correctly classified examples, the true performance p is between 69.1% and 80.1%.

Confidence Interval for Accuracy

- Consider a model that produces an accuracy of 80% when evaluated on 100 test instances:
  - $N = 100, acc = 0.8$
  - Let $1 - \alpha = 0.95$ (95% confidence)
  - From probability table, $Z_{\alpha/2} = 1.96$

$1-\alpha$	Z
0.99	2.58
0.98	2.33
0.95	1.96
0.90	1.65

N	50	100	500	1000	5000
p (lower)	0.670	0.711	0.763	0.774	0.789
p (upper)	0.888	0.866	0.833	0.824	0.811

Establishing Confidence Intervals

- If S contains n examples drawn independently and $n \geq 30$, then

- With approximately 95% probability (or confidence), $\text{error}_D(h)$ lies in the interval

$$\text{error}_S(h) \pm 1.96 \sqrt{\frac{\text{error}_D(h)\left(1 - \text{error}_D(h)\right)}{n}}$$

Comparing Performance of Two Models

- Two models, say M1 and M2, which is better?
- M1 is tested on D1 (size $= n_1$), found error rate $= e_1$.
- M2 is tested on D2 (size $= n_2$), found error rate $= e_2$.
- Assume D1 and D2 are independent.
- If n_1 and n_2 are sufficiently large, then:

$$e_1 \sim N(\mu_1, \sigma_1) \qquad e_2 \sim N(\mu_2, \sigma_2)$$

- Approximate:

$$\hat{\sigma}_i = \frac{e_i(1-e_i)}{n_i}$$

Comparing Performance of Two Models

- To test if the difference between the performance of M1 and M2 is statistically significant, we consider $d = e1 - e2$.
- $d \sim N(d_t, \sigma_t)$, where d_t is the true difference.
- Since D1 and D2 are independent:

$$\sigma_t^2 = \sigma_1^2 + \sigma_2^2 \cong \hat{\sigma}_1^2 + \hat{\sigma}_1^2$$
$$= \frac{e_1(1 - e_1)}{n_1} + \frac{e_2(1 - e_2)}{n_2}$$

- At $(1 - \alpha)$ confidence level: $d_t = d \pm Z_{\alpha/2} \hat{\sigma}_t$.

Example of Comparing Two Models

- Given M1 with $n_1 = 30$ and $e_1 = 0.15$ and M2 with $n_2 = 5000$ and $e_2 = 0.25$, $d = 0.1$ (2-sided test). Thus,

$$\hat{\sigma}_d = \frac{0.15(1-0.15)}{30} + \frac{0.25(1-0.25)}{5000} = 0.0043$$

- At 95% confidence level, $Z_{\alpha/2} = 1.96$:

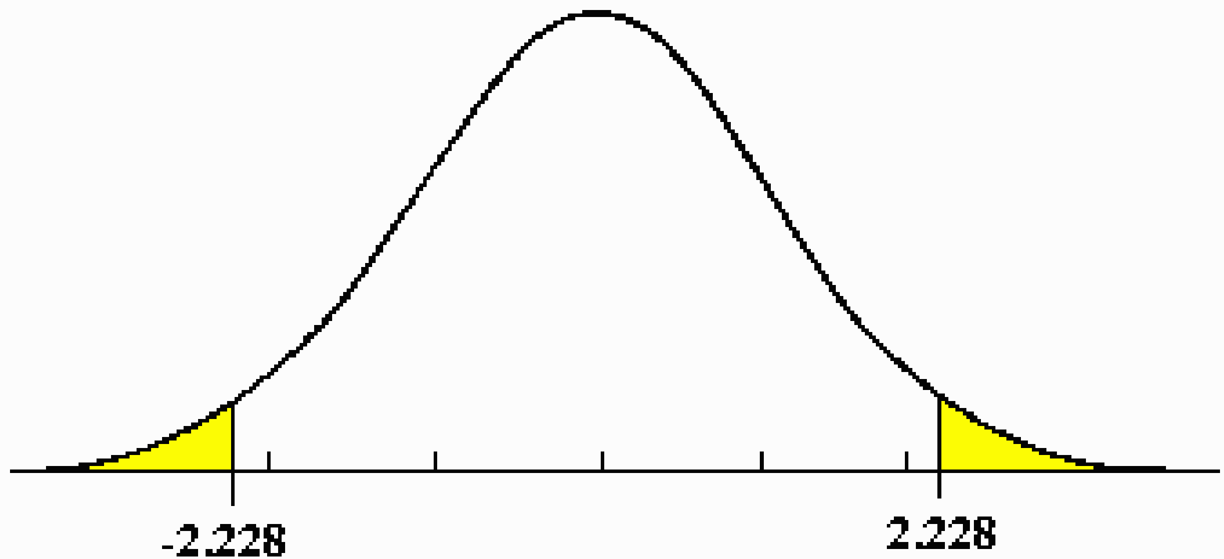
$$d_t = 0.100 \pm 1.96 \cdot \sqrt{0.0043} = 0.100 \pm 0.128.$$

- The interval contains 0, therefore the difference may not be statistically significant.

Student's t-Test

- Student's t-test tells us whether the means of two samples are significantly different
- Take individual samples from the sets of all possible cross-validation estimates
- Use a paired t-test because the individual samples are paired
  - The same CV is applied twice

Two-tailed t-Test



Comparing Two Classifiers

- Suppose we want to compare the performance of two classifiers using the k-fold cross-validation approach.
  - Assume we did 10-fold CV for two classifiers
- We want to know if there is a statistically significant difference between the two means.

Comparing Algorithms A and B

- Partition data D into k stratified disjoint subsets $T_1, T_2, \ldots, T_k$ of equal size.
- For $i = 1$ to k do

Use T_i as the testing set, and the remaining data for training set S_i

$$S_i \leftarrow \{D - T_i\}$$
$$h_A \leftarrow L_A(S_i)$$
$$h_B \leftarrow L_B(S_i)$$
$$\delta_i \leftarrow \mathrm{error}_i(h_A) - \mathrm{error}_i(h_B)$$

Return $\bar{\delta}$, where

$$\bar{\delta} \equiv \frac{1}{k}\sum_{i=1}^{k} \delta_i$$

Comparing Classifiers A and B

- The difference of the means also has a Student's distribution with $k-1$ degrees of freedom
- N% confidence interval for δ: $\bar{\delta} \pm t_{N,k-1}\ s_{\bar{\delta}}$

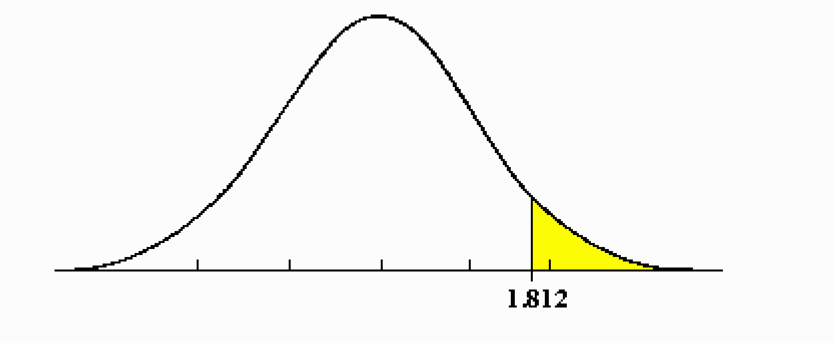
$$s_{\bar{\delta}} \equiv \sqrt{\frac{1}{k(k-1)}\sum_{i=1}^{k}\left(\delta_i - \bar{\delta}\right)^2}$$

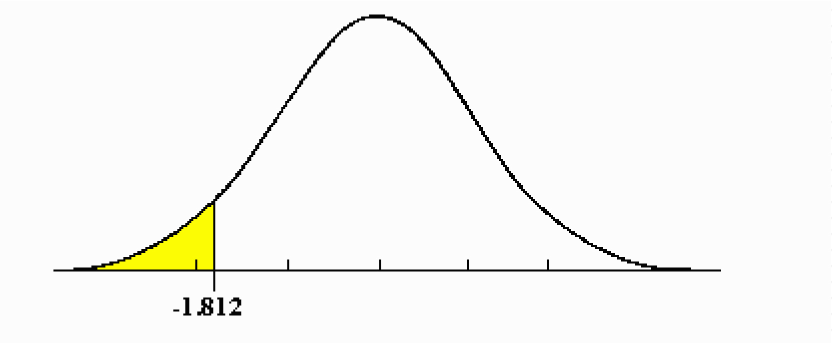
$$t_{N,K-1} = \frac{\bar{\delta}}{s_{\bar{\delta}}}$$

Performing the t-Test

1. Fix a significance level α
   - If a difference is significant at the $\alpha\%$ level, there is a $(100-\alpha)\%$ chance that there really is a difference
2. Divide the significance level by two, because the test is two-tailed
   - i.e., the true difference can be positive or negative
3. If $t_{N,k-1} < -t$ or $t \geq t_{N,k-1}$ then the difference is significant
   - i.e., the null hypothesis can be rejected

One-tailed t-Test





Unpaired Observations

- If the CV estimates are from different randomizations, they are no longer paired
- Then we have to use an unpaired t-test with $\min(k, j) - 1$ degrees of freedom
- The t-statistic becomes:

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

Evaluation Measures Summary

- What you should know?
  - Confidence intervals
  - Evaluation schemes—hold-out, 10-fold CV, bootstrap, etc.
  - Significance tests
  - Different evaluation measures for classification
    - Error/accuracy, ROC, f-measure, lift curves, cost-sensitive classification