CS 391L: Machine Learning: Experimental Evaluation

Raymond J. Mooney

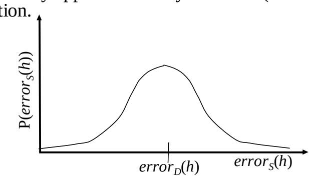
University of Texas at Austin

Evaluating Inductive Hypotheses

- Accuracy of hypotheses on training data is obviously biased since the hypothesis was constructed to fit this data.
- Accuracy must be evaluated on an independent (usually disjoint) test set.
- The larger the test set is, the more accurate the measured accuracy and the lower the variance observed across different test sets.

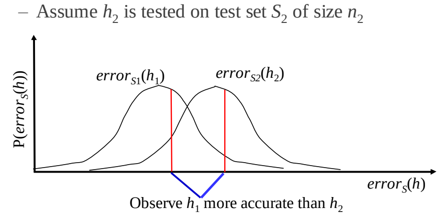
Variance in Test Accuracy

- Let $error_S(h)$ denote the percentage of examples in an independently sampled test set S of size n that are incorrectly classified by hypothesis h.
- Let $error_D(h)$ denote the true error rate for the overall data distribution D.
- When n is at least 30, the *central limit theorem* ensures that the distribution of $error_S(h)$ for different random samples will be closely approximated by a normal (Guassian) distribution.



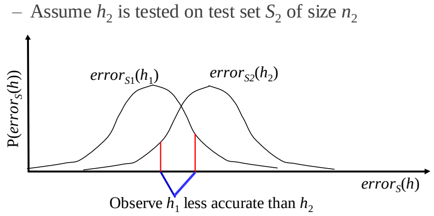
Comparing Two Learned Hypotheses

- When evaluating two hypotheses, their observed ordering with respect to accuracy may or may not reflect the ordering of their true accuracies.
  - Assume h_1 is tested on test set S_1 of size n_1
  - Assume h_2 is tested on test set S_2 of size n_2



Comparing Two Learned Hypotheses

- When evaluating two hypotheses, their observed ordering with respect to accuracy may or may not reflect the ordering of their true accuracies.
  - Assume h_1 is tested on test set S_1 of size n_1
  - Assume h_2 is tested on test set S_2 of size n_2



Statistical Hypothesis Testing

- Determine the probability that an empirically observed difference in a statistic could be due purely to random chance assuming there is no true underlying difference.
- Specific tests for determining the significance of the difference between two means computed from two samples gathered under different conditions.
- Determines the probability of the ***null hypothesis***, that the two samples were actually drawn from the same underlying distribution.
- By scientific convention, we ***reject the null hypothesis*** and say the difference is ***statistically significant*** if the probability of the null hypothesis is less than 5% ($p < 0.05$) or alternatively we accept that the difference is due to an underlying cause with a ***confidence*** of $(1 - p)$.

One-sided vs Two-sided Tests

- One-sided test assumes you expected a difference in one direction (A is better than B) and the observed difference is consistent with that assumption.
- Two-sided test does not assume an expected difference in either direction.
- Two-sided test is more conservative, since it requires a larger difference to conclude that the difference is significant.

7

Z-Score Test for Comparing Learned Hypotheses

- Assumes h_1 is tested on test set S_1 of size n_1 and h_2 is tested on test set S_2 of size n_2.
- Compute the difference between the accuracy of h_1 and h_2

$$d = \left| error_{S_1}(h_1) - error_{S_2}(h_2) \right|$$

- Compute the standard deviation of the sample estimate of the difference.

$$\sigma_d = \sqrt{\frac{error_{S_1}(h_1)\cdot(1-error_{S_1}(h_1))}{n_1} + \frac{error_{S_2}(h_2)\cdot(1-error_{S_2}(h_2))}{n_2}}$$

- Compute the z-score for the difference

$$z = \frac{d}{\sigma_d}$$

8

Z-Score Test for Comparing Learned Hypotheses (continued)

- Determine the confidence in the difference by looking up the highest confidence, *C*, for the given z-score in a table.

confidence level	50%	68%	80%	90%	95%	98%	99%
z-score	0.67	1.00	1.28	1.64	1.96	2.33	2.58

- This gives the confidence for a two-tailed test, for a one tailed test, increase the confidence half way towards 100%

$$C' = (100 - \frac{(100-C)}{2})$$

9

Sample Z-Score Test 1

Assume we test two hypotheses on different test sets of size 100 and observe:

$$error_{S_1}(h_1) = 0.20 \qquad error_{S_2}(h_2) = 0.30$$

$$d = \left| error_{S_1}(h_1) - error_{S_2}(h_2) \right| = \left| 0.2 - 0.3 \right| = 0.1$$

$$\sigma_d = \sqrt{\frac{error_{S_1}(h_1)\cdot(1-error_{S_1}(h_1))}{n_1} + \frac{error_{S_2}(h_2)\cdot(1-error_{S_2}(h_2))}{n_2}}$$

$$= \sqrt{\frac{0.2\cdot(1-0.2)}{100} + \frac{0.3\cdot(1-0.3)}{100}} = 0.0608$$

$$z = \frac{d}{\sigma_d} = \frac{0.1}{0.0608} = 1.644$$

Confidence for two-tailed test: 90%
Confidence for one-tailed test: (100 – (100 – 90)/2) = 95%

10

Sample Z-Score Test 2

Assume we test two hypotheses on different test sets of size 100 and observe:

$$error_{S_1}(h_1) = 0.20 \qquad error_{S_2}(h_2) = 0.25$$

$$d = \left| error_{S_1}(h_1) - error_{S_2}(h_2) \right| = \left| 0.2 - 0.25 \right| = 0.05$$

$$\sigma_d = \sqrt{\frac{error_{S_1}(h_1)\cdot(1-error_{S_1}(h_1))}{n_1} + \frac{error_{S_2}(h_2)\cdot(1-error_{S_2}(h_2))}{n_2}}$$

$$= \sqrt{\frac{0.2\cdot(1-0.2)}{100} + \frac{0.25\cdot(1-0.25)}{100}} = 0.0589$$

$$z = \frac{d}{\sigma_d} = \frac{0.05}{0.0589} = 0.848$$

Confidence for two-tailed test: 50%
Confidence for one-tailed test: (100 – (100 – 50)/2) = 75%

11

Z-Score Test Assumptions

- Hypotheses can be tested on different test sets; if same test set used, stronger conclusions might be warranted.
- Test sets have at least 30 independently drawn examples.
- Hypotheses were constructed from independent training sets.
- Only compares two specific hypotheses regardless of the methods used to construct them. Does not compare the underlying learning methods in general.

12

Comparing Learning Algorithms

- Comparing the average accuracy of hypotheses produced by two different learning systems is more difficult since we need to average over multiple training sets. Ideally, we want to measure:

$$E_{S\subset D}(error_D(L_A(S)) - error_D(L_B(S)))$$

where $L_X(S)$ represents the hypothesis learned by method L from training data S.

- To accurately estimate this, we need to average over multiple, independent training and test sets.
- However, since labeled data is limited, generally must average over multiple splits of the overall data set into training and test sets.

13

K-Fold Cross Validation

Randomly partition data D into k disjoint equal-sized subsets $P_1 \ldots P_k$

For i from 1 to k do:

Use P_i for the test set and remaining data for training

$S_i = (D - P_i)$

$h_A = L_A(S_i)$

$h_B = L_B(S_i)$

$\delta_i = \text{error}_{P_i}(h_A) - \text{error}_{P_i}(h_B)$

Return the average difference in error:

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

14

K-Fold Cross Validation Comments

- Every example gets used as a test example once and as a training example k–1 times.
- All test sets are independent; however, training sets overlap significantly.
- Measures accuracy of hypothesis generated for $[(k-1)/k]\cdot|D|$ training examples.
- Standard method is 10-fold.
- If k is low, not sufficient number of train/test trials; if k is high, test set is small and test variance is high and run time is increased.
- If k=|D|, method is called ***leave-one-out*** cross validation.

15

Significance Testing

- Typically k<30, so not sufficient trials for a z test.
- Can use (***Student's***) ***t-test***, which is more accurate when number of trials is low.
- Can use a ***paired*** t-test, which can determine smaller differences to be significant when the training/sets sets are the same for both systems.
- However, both z and t test's assume the trials are independent. Not true for k-fold cross validation:
  - Test sets are independent
  - Training sets are **not** independent
- Alternative statistical tests have been proposed, such as McNemar's test.
- Although no test is perfect when data is limited and independent trials are not practical, some statistical test that accounts for variance is desirable.

16

Sample Experimental Results

Which experiment provides better evidence that SystemA is better than SystemB?

Experiment 1

	SystemA	SystemB	Diff
Trial 1	87%	82%	+5%
Trail 2	83%	78%	+5%
Trial 3	88%	83%	+5%
Trial 4	82%	77%	+5%
Trial 5	85%	80%	+5%
Average	85%	80%	+5%

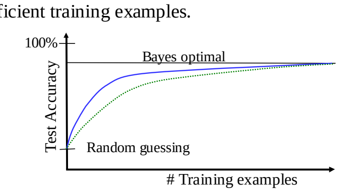
Experiment 2

	SystemA	SystemB	Diff
Trial 1	90%	82%	+8%
Trail 2	93%	76%	+17%
Trial 3	80%	85%	–5%
Trial 4	85%	75%	+10%
Trial 5	77%	82%	– 5%
Average	85%	80%	+5%

17

Learning Curves

- Plots accuracy vs. size of training set.
- Has maximum accuracy (Bayes optimal) nearly been reached or will more examples help?
- Is one system better when training data is limited?
- Most learners eventually converge to Bayes optimal given sufficient training examples.



18

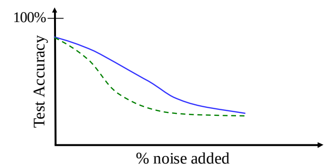
Cross Validation Learning Curves

Split data into k equal partitions
For trial i = 1 to k do:
  Use partition i for testing and the union of all other partitions for training.
  For each desired point p on the learning curve do:
    For each learning system L
      Train L on the first p examples of the training set and record
        training time, training accuracy, and learned concept complexity.
      Test L on the test set, recording testing time and test accuracy.
Compute average for each performance statistic across k trials.
Plot curves for any desired performance statistic versus training set size.
Use a paired t-test to determine significance of any differences between any two systems for a given training set size.

19

Noise Curves

- Plot accuracy versus noise level to determine relative resistance to noisy training data.
- Artificially add category or feature noise by randomly replacing some specified fraction of category or feature values with random values.



20

Experimental Evaluation Conclusions

- Good experimental methodology is important to evaluating learning methods.
- Important to test on a variety of domains to demonstrate a general bias that is useful for a variety of problems. Testing on 20+ data sets is common.
- Variety of freely available data sources
  - UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html
  - KDD Cup (large data sets for data mining) http://www.kdnuggets.com/datasets/kddcup.html
  - CoNLL Shared Task (natural language problems) http://www.ifarm.nl/signll/conll/
- Data for real problems is preferable to artificial problems to demonstrate a useful bias for real-world problems.
- Many available datasets have been subjected to significant feature engineering to make them learnable.

21