Machine Learning: Experimental Evaluation
Motivation

• Evaluating the performance of learning systems is important because:
  – Learning systems are usually designed to predict the class of “future” unlabeled data points.
  – In some cases, evaluating hypotheses is an integral part of the learning process (example, when pruning a decision tree)
Issues

How well can a classifier be expected to perform on novel data?

Choice of performance measure

How close is the estimated performance to the true performance?

Comparing Classifiers
Performances of a given hypothesis

Confusion matrix (also called a contingency table)

- **TP**: Number of True positives
- **FP**: Number of False positives
- **TN**: Number of True Negatives
- **FN**: Number of False Negative

\[
\text{Accuracy} = \frac{(TP + TN)}{n}
\]

\[
n = TP + TN + FP + FN
\]
Performances of a given hypothesis

- Precision, Recall, F-measure

- $P = \frac{TP}{TP + FP}$
- $R = \frac{TP}{TP + FN}$
- $F = \frac{2(P \times R)}{P + R}$
Receiver Operating Characteristic curve (or ROC curve.) It is a plot of the true positive rate against the false positive rate different possible cutpoints of a diagnostic test. (e.g. increasing learning set, using different algorithm parameters..)
Evaluating an hypothesis

- ROC and accuracy not enough
- How well will the learned classifier perform on novel data?
- Performance on the training data is not a good indicator of performance on future data
Testing on the training data is not appropriate. If we have a limited set of training examples, possibly the system will misclassify new unseen instances.
Difficulties in Evaluating Hypotheses when only limited data are available

- **Bias in the estimate**: The observed accuracy of the learned hypothesis over the training examples is a poor estimator of its accuracy over future examples ==> we must test the hypothesis on a test set chosen independently of the training set and the hypothesis.

- **Variance in the estimate**: Even with a separate test set, the measured accuracy can vary from the true accuracy, depending on the makeup of the particular set of test examples. The smaller the test set, the greater the expected variance.
Variance
Questions to be Considered

- Given the observed accuracy of a hypothesis over a limited sample of data, **how well does this estimate** its accuracy over additional examples?
- Given that one hypothesis outperforms another over some sample data, **how probable is it** that this hypothesis is more accurate, in general?
- When data is limited what is the **best way to use** this data to both learn a hypothesis and estimate its accuracy?
Estimating Hypothesis Accuracy

Two Questions of Interest:

- Given a hypothesis $h$ and a data sample containing $n$ examples drawn at random according to distribution $\mathcal{D}$, what is the best estimate of the accuracy of $h$ over future instances drawn from the same distribution? $\Rightarrow$ sample error vs. true error

- What is the probable error in this accuracy estimate? $\Rightarrow$ confidence intervals = error in estimating error
Sample Error and True Error

- **Definition 1:** The *sample error* (denoted $error_s(h)$ or $e_S(h)$) of hypothesis $h$ with respect to target function $f$ and data sample $S$ is:
  
  $$ error_s(h) = \frac{1}{n} \times \sum_{x \in S} \delta(f(x), h(x)) = \frac{r}{n} $$

  where $n$ is the number of examples in $S$, and the quantity $\delta(f(x), h(x))$ is 1 if $f(x) \neq h(x)$, and 0, otherwise.

- **Definition 2:** The *true error* (denoted $error_D(h)$, or $p$) of hypothesis $h$ with respect to target function $f$ and distribution $D$, is the probability that $h$ will misclassify an instance drawn at random according to $D$.
  
  $$ error_D(h) = p = Pr_{x \in D}[f(x) \neq h(x)] $$
Example

Hypothesis $h$ misclassifies 12 of the 40 examples in $S$

$$error_S(h) = \frac{12}{40} = .30$$

What is $error_D(h)$?
error_D is a random variable

- **Compare:**
  - Toss a coin \( n \) times and it turns up heads \( r \) times with probability of each toss turning heads \( p = P(\text{heads}) \). (Bernoulli trial)
  - Randomly draw \( n \) samples, and \( h \) (the learned hypothesis) misclassifies \( r \) over \( n \) samples with probability of misclassification \( p = \text{error}_D(h) \). What is the (apriori) probability of \( r \) misclassifications over \( n \) trials?

\[
\text{error}_S(h) = \frac{r}{n}
\]

**Probability of observing \( r \) misclassified examples:**

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

\( P(r) \) is the probability of \( r \) misclassifications and \( (n-r) \) good classifications. However it depends on the value of \( \text{error}_D = p \) which is **unknown**!
Binomial distribution

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

Expected, or mean value of \( X \), \( E[X] \), is

\[ E[X] = \sum_i iP(i) = np \]

Variance of \( X \) is

\[ Var(X) = E[(X - E[X])^2] = np(1 - p) \]

Standard deviation of \( X \), \( \sigma_X \), is

\[ \sigma_X = \sqrt{E[(X - E[X])^2]} = \sqrt{np(1 - p)} \]

But again, \( p = \text{error}_\mathcal{D}(h) \), the probability that a single item is misclassified is UNKNOWN!
Bias and variance

Notice that also the estimator $\text{error}_S(h)$ of $p$ is a random variable! *If we perform many experiments we could get different values!*

$\text{error}_S(h)$ follows a *Binomial* distribution, with

- **mean** $\mu_{\text{error}_S(h)} = \text{error}_D(h) = p$
- **standard deviation** $\sigma_{\text{error}_S(h)}$

\[
\sigma_{\text{error}_S(h)} = \sqrt{\frac{\text{error}_D(h)(1-\text{error}_D(h))}{n}}
\]

\[
\sigma_{\text{error}_S(h)} \approx \sqrt{\frac{\text{error}_S(h)(1-\text{error}_S(h))}{n}}
\]

$\text{error}_S(h)$ is an *unbiased estimator* for $\text{error}_D(h)$

the **bias** of an *estimator* is the difference between this estimator's *expected value* and the true value of the parameter being estimated
More details

Why is this true??

\[ \sigma_s = \frac{\sigma_D}{n} = \frac{1}{n} \sqrt{np(1-p)} \approx \sqrt{\frac{r(1-r)}{n(n-1)}} = \sqrt{\frac{\text{error}_S(h)(1-\text{error}_S(h))}{n}} \]
In other words, if you perform several experiments, $\operatorname{E}($errorS(h)) $\to p$
Variance in Test Accuracy

- When the number of trials is at least 30, the central limit theorem ensures that the (binomial) distribution of $\text{error}_S(h)$ for different random samples of size $n$ will be closely approximated by a normal (Guassian) distribution, with expected value $\text{error}_D(h)$.
Central Limit Theorem (μ=expected value of the true distribution)

In this experiment, each figure shows the distribution of values with a growing number of dices (from 1 to 50), and a large number of tosses. X axis is the sum of dices’ values when tossing n dices, and Y is the frequency of occurrence of that value.
Summary

• Number of dices ➔ number of examples to test
• Outcome of tossing n dices ➔ errors of the classifier over n test examples
• The sample error \( \text{error}_S = r / n \) is a random variable following a Binomial distribution: if we repeat the experiment on different samples of the same dimension n we would obtain different values for error\(_S\) with probability \( P_{\text{binomial}}(r_i / n) \)
• The expected value of error\(_S\) (for different experiments) equals \( np \), where \( p \) is the real (unknown) error rate
• If the number of examples n is >30, then the underlying binomial distribution approximates a Gaussian distribution with same expected value
Confidence interval for an error estimate

- **Confidence interval**
  \[ LB \leq |error_D(h(x)) - error_S(h(x))| \leq UB \]

- Provides a measure of the minimum and maximum expected **discrepancy** between the **measured and real** error rate

- **Def**: an **N% confidence interval** for an estimate \( e \) is an interval \([LB, UB]\) that includes \( e \) with probability N% (with probability N% we have \( LB \leq e \leq UB \))

- In our case, \( e = |error_D(h(x)) - error_S(h(x))| \) is a random variable representing the **difference** between true and estimated error (\( e \) is the error in estimating the error!!)

- If \( e \) obeys a **gaussian distribution**, the confidence interval can be easily estimated
Confidence interval

- Given an hypothesis $h(x)$, if $error_S(h(x))$ obeys a **gaussian** with mean $\mu = error_D(h(x)) = p$ and variance $\sigma$ (which we said is approximately true for $n > 30$) then the estimated value $error_S(h(x))$ on a sample $S$ of $n$ examples, $r/n$, with **probability N\%** will lie in the following interval: $\mu \pm z_N \sigma$

- $z_N$ is half of the length of the interval around $\mu$ that includes $N\%$ of the total probability mass.

Again, remember that the Gaussian shows the probability distribution of our outcomes (outcomes are the measured error values $r/n$)

Of course, we don’t know where our estimator $r/n$ is placed in this distribution, because we don’t actually know $p$ nor $\sigma$, we only know that $r/n$ lies somewhere in a gaussian-shaped distribution.
Probability mass in a gaussian distribution

Probability mass as a function of $\sigma$

$\mu \pm z N \sigma$

So, what we actually KNOW is that WHEREVER our estimated error is in this Gaussian, it will be with 68% probability at a distance $\pm 1 \sigma$ from the true error rate $p$, with 80% probability at a distance $\pm 1.28 \sigma$, with 98% probability at a distance $\pm 2.33 \sigma$, etc. And even though we don’t know $\sigma$, we have seen that it can be approximated with:

$$\sqrt{\frac{r}{n} \left(1 - \frac{r}{n}\right)}$$

0 and standard deviation 1

<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>
Finding the confidence interval

- N% of the mass lies between $\mu \pm z_N \sigma$
- 80% lies in $\mu \pm 1.28 \sigma$

<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>

- For a Gaussian with average 0 and standard deviation 1:

$$\sigma_s = \frac{\sigma_D}{n} = \frac{1}{n} \sqrt{np(1-p)} \approx \sqrt{\frac{r(1-r)}{n} \frac{1}{n}} = \sqrt{\frac{\text{error}_s(h)(1-\text{error}_s(h))}{n}}$$
Finding the confidence interval

\[
[LB, UB] = \left[ (\mu - Z_n \sigma), (\mu + Z_n \sigma) \right] \approx \\
\left[ e(h) - Z^N_n \sqrt{\frac{e(h)(1-e(h))}{n}}, e(h) + Z^N_n \sqrt{\frac{e(h)(1-e(h))}{n}} \right]
\]

\[ e(h) = error_S(h) \]
Calculating the N% Confidence Interval: Example

Consider the following example:

- A classifier has a 13% chance of making an error
- A sample $S$ containing 100 instances is drawn
- We can now compute, that with 90% confidence we can say that the true error lies in the interval,

$$\left[0.13 - 1.64 \sqrt{\frac{0.13(1-0.13)}{100}}, 0.13 + 1.64 \sqrt{\frac{0.13(1-0.13)}{100}}\right] = [0.075, 0.19]$$

<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>
Example 2

Given the following extract from a scientific paper on multimodal emotion recognition:

We trained the classifiers with 156 samples and tested with 50 samples from three subjects.

Table 3. Emotion recognition results for 3 subjects using 156 training and 50 testing samples.

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Number of Classes</th>
<th>Classifier</th>
<th>Correctly classified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face*</td>
<td>67</td>
<td>C4.5</td>
<td>78 %</td>
</tr>
<tr>
<td>Body*</td>
<td>140</td>
<td>BayesNet</td>
<td>90 %</td>
</tr>
</tbody>
</table>

For the Face modality, what is $n$? What is $\text{error}_s(h)$?

<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>
Solution

Precision is 0.78 hence error rate $r/n$ is 0.22; the test set has 50 instances, hence $n=50$

Choose e.g. to compute the N% confidence interval with $N=0.95$

Given that $error_s(h)=0.22$ and $n=50$, and $z_N=1.96$ for $N=95$, we can now say that with 95% probability $error_D(h)$ will lie in the interval:

$$\left[0.22 - 1.96 \sqrt{\frac{0.22(1-0.22)}{50}}, 0.22 + 1.96 \sqrt{\frac{0.22(1-0.22)}{50}}\right] = \left[0.11, 0.34\right]$$
One sided bound

• We might be interested not in a confidence interval with both an upper and a lower bound, but instead in the upper or lower limit only. For instance, what is the probability that $error_D(h)$ is at most $U$.

• In the confidence interval example, we found that with $(100-a)=95\%$ confidence

$$L = 0.11 \leq error_D(h) \leq 0.34 = U$$

• Using the symmetry property of a normal distribution, we now find that $error_D(h) \leq U=0.34$ with confidence $(100-a/2)=97.5\%$.

$a=(100-95)\%=5\%$ is the percentage of the gaussian area outside lower and upper extremes of the curve; $a/2=2.5\%$. 
One sided two sided bounds
Upper bound
Evaluating error

- Evaluating the error rate of an hypothesis
- Evaluating alternative hypotheses
- Evaluating alternative learning algorithms
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 $n_2$

\[
P(\text{error}_{S_1}(h_1)) > P(\text{error}_{S_2}(h_2))
\]

Observe $h_1$ more accurate than $h_2$
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 $S_3$ of size $n_1$
  - Assume $h_2$ is tested on test set $S_4$ of size $n_2$

Observe $h_1$ less accurate than $h_2$
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 $h_2$
  \[ \hat{d} = \left| \text{error}_{S_1}(h_1) - \text{error}_{S_2}(h_2) \right| \]
- Compute the standard deviation of the sample’s estimate of the difference.
  \[ \sigma_d = \sqrt{\frac{\text{error}_{S_1}(h_1) \cdot (1 - \text{error}_{S_1}(h_1))}{n_1} + \frac{\text{error}_{S_2}(h_2) \cdot (1 - \text{error}_{S_2}(h_2))}{n_2}} \]
Z-Score Test for Comparing Learned Hypotheses

- \( d^\wedge \pm Z_N \sqrt{\text{error}_{S1}(h_1)(1-\text{error}_{S1}(h_1))/n_1 + \text{error}_{S2}(h_2)(1-\text{error}_{S2}(h_2))/n_2} = d^\wedge \pm Z_N \sigma \)

- Contrary to the previous case, where only one value of the difference was known (i.e. we did know the sample error, but not the true error) we now know both values of the difference (both sample errors of the two hypotheses). Therefore, \( d \) is known, and \( \sigma_d \) can be approximated, as before.

- We can then compute \( Z_N \)

\[ Z_N = \frac{d^\wedge}{\sigma_d} \]
Z-Score Test for Comparing Learned Hypotheses (continued)

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

<table>
<thead>
<tr>
<th>confidence level</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-score</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>
Example

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

\[ \text{error}_{S_1}(h_1) = 0.20 \quad \text{error}_{S_2}(h_2) = 0.30 \]

\[ d = \left| \text{error}_{S_1}(h_1) - \text{error}_{S_2}(h_2) \right| = |0.2 - 0.3| = 0.1 \]

\[ \sigma_d = \sqrt{\frac{\text{error}_{S_1}(h_1) \cdot (1 - \text{error}_{S_1}(h_1))}{n_1} + \frac{\text{error}_{S_2}(h_2) \cdot (1 - \text{error}_{S_2}(h_2))}{n_2}} \]

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

\[ z = \frac{d}{\sigma_d} = \frac{0.1}{0.0608} = 1.644 \]

From table, if \( z = 1.64 \) confidence level is 90%
Example 2

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

\[ \text{error}_{S_1}(h_1) = 0.20 \quad \text{error}_{S_2}(h_2) = 0.25 \]

\[ d = \left| \text{error}_{S_1}(h_1) - \text{error}_{S_2}(h_2) \right| = |0.2 - 0.25| = 0.05 \]

\[ \sigma_d = \sqrt{\frac{\text{error}_{S_1}(h_1) \cdot (1 - \text{error}_{S_1}(h_1))}{n_1} + \frac{\text{error}_{S_2}(h_2) \cdot (1 - \text{error}_{S_2}(h_2))}{n_2}} \]

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

\[ z = \frac{d}{\sigma_d} = \frac{0.05}{0.0589} = 0.848 \]

Confidence between 50% and 68%
Z-Score Test Assumptions: summary

- Hypotheses can be tested on different test sets; if the same test set is used, stronger conclusions might be warranted.
- Test set(s) must have at least 30 independently drawn examples (to apply central limit theorem).
- Hypotheses must be 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.
Evaluating error

- Evaluating the error rate of an hypothesis
- Evaluating alternative hypotheses
- Evaluating alternative learning algorithms
Comparing 2 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:

\[ E_{S \subseteq D}(error_D(L_A(S)) - error_D(L_B(S))) \]

where \( L_X(S) \) represents the hypothesis learned by learning algorithm \( L_X \) 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 (**K-fold cross validation**).
K-fold cross validation of an hypothesis

Partition the test set in k equally sized random samples
K-fold cross validation (2)

At each step, learn from Li and test on Ti, then compute the error

\[ e_1, e_2, e_3 \]
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 (at each step, you leave out one example).
Use K-Fold Cross Validation to evaluate different learners

Randomly partition data $D$ into $k$ disjoint equal-sized (N) 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:

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

Error bound is computed as:

$$\delta \pm t N, k - 1 \cdot \sigma_{\overline{\delta}}$$

where $t$ has the same role as $z$
**Sample Experimental Results**

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

### Experiment 1

<table>
<thead>
<tr>
<th>Trial</th>
<th>SystemA (%)</th>
<th>SystemB (%)</th>
<th>Diff (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial 1</td>
<td>83%</td>
<td>78%</td>
<td>+5%</td>
</tr>
<tr>
<td>Trial 2</td>
<td>88%</td>
<td>83%</td>
<td>+5%</td>
</tr>
<tr>
<td>Trial 3</td>
<td>82%</td>
<td>77%</td>
<td>+5%</td>
</tr>
<tr>
<td>Trial 4</td>
<td>85%</td>
<td>80%</td>
<td>+5%</td>
</tr>
<tr>
<td>Average</td>
<td>85%</td>
<td>80%</td>
<td>+5%</td>
</tr>
</tbody>
</table>

Experiment 1 has $\sigma=0$, therefore we have a perfect confidence in the estimate of $\delta$.

### Experiment 2

<table>
<thead>
<tr>
<th>Trial</th>
<th>SystemA (%)</th>
<th>SystemB (%)</th>
<th>Diff (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial 1</td>
<td>93%</td>
<td>82%</td>
<td>+17%</td>
</tr>
<tr>
<td>Trial 2</td>
<td>80%</td>
<td>85%</td>
<td>-5%</td>
</tr>
<tr>
<td>Trial 3</td>
<td>85%</td>
<td>85%</td>
<td>+10%</td>
</tr>
<tr>
<td>Trial 4</td>
<td>77%</td>
<td>82%</td>
<td>-5%</td>
</tr>
<tr>
<td>Average</td>
<td>85%</td>
<td>80%</td>
<td>+5%</td>
</tr>
</tbody>
</table>
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.
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.
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.