

## Intro.

- Experiments: to verify hypotheses
- Hypothesis [wilson '52]:
  - Trail idea, a tentative concerning the nature of things
  - Until it has been *tested*, it should not be confused with a *law*
- Plausibility is not a substitute for evidence, however great may be the emotional wish to believe.

## Experiments 101

Based on Zobel's "writing for CS" Chpt. 8 &  
Alpaydin's "Intro. to Machine Learning" Chpt. 14

## Defending hypotheses

- Testing hypothesis & supporting evidence
- Relate hypothesis to evidence; e.g.,
  - *H*: "Rev-Tree is more accurate than previous Tree-classifiers on textual data"
  - Supporting E: "Rev-Tree achieves 98% classification accuracy on 20-NG" (*is it ok?*)
  - Missing: "state-of-the-art D-tree achieves ??% on 20-NG"

## Stating hypotheses

- Clearly
- Precisely
- Unambiguous
- Testable
- Vulnerable to falsification → more convincing

## Designing fair experiments

- When hypo' least likely hold?
- Other interpretations of results?
- Check negative, failed experiemnts
- Results sensible (e.g., boundary conditions)
- Don't draw undue conclusions

## Evidence

Four types:

1. Analysis or Proof
2. Modeling
3. Simulation
4. Experiment

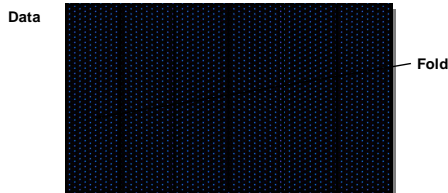
### Describing experiments

- Experiments are only valuable if carefully described
- Experiment – verifiable and reproducible:
  - If based on code: consider publish it
  - Must decide which results to report
  - Unethical to conceal failures
  - Consider sketching “route to hypothesis”

### Designing robust experiments

- Minimize effect of extraneous factors
- Example:
  - choice of test data
  - Baseline algs'

### K-fold cross-validation



### Accessing Classification Alg.'s

- Standard benchmarks:
  - UCI repository
  - NIST digits
  - 20NG, Reuters
  - Calgary Corpus

### Which hyper-parameter?

#### Validation folds

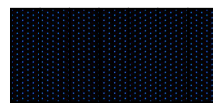


Training Set

Do CV – however, only using training set

Example: pre-pruning of decision tree

### CV: the i'th fold



Test Set

Training Set

Value of "k":

- 10-fold; 5-fold
- Leave one out
- How about 2 fold cv?

### CV – pseudo code (cont.)

```

createFolds(S, N) {
1: Randomly partition S into N subset of equal size and tags proportions,
   S1, S2, ..., SN;
2: Construct N training/test partitions (Sitrain, Sitest)i=1N such that Sitrain =
   S \ Si and Sitest = Si;
3: return (Sitrain, Sitest)i=1N.
}
    
```

### CV – pseudo code

```

Require: 1. A labeled dataset S = {(xi, yi)}i=1m
         2. A learning algorithm A(θ)
         3. A set Θ of feasible hyper-parameter vectors
         4. Nf = number of folds
         5. Nv = number of validation folds
Ensure: An estimate  $\hat{L}_A$  for the best achievable error of the learning algorithm
      A trained on a set of a size |S|(1 - 1/Nf).

experiment(S, A, Θ, Nf, Nv) {
1: (Sitrain, Sitest)i=1Nf = createFolds(S, Nf)
2: for 1 ≤ i ≤ Nf do
3:   Let θ* = findBestHyper(Sitrain, Nv, Θ)
4:   Compute LA(i) = calcError(A, θ*, Sitrain, Sitest)
5: end for
6: return  $\frac{1}{Nf} \sum_i \hat{L}_A(i)$ 
}
    
```

### CV – pseudo code (cont.)

```

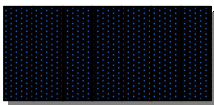
calcError(A, θ, Strain, Stest) {
1: Let C be the classifier generated by A(θ) trained with Strain
2: return  $\sum_{(x_i, y_i) \in S_{test}} I(y_i = C(x_i))$  (I is the indicator function)
}
    
```

### CV – pseudo code (cont.)

```

findBestHyper(S, N, Θ) {
1: (Sitrain, Sitest)i=1N = createFolds(S, N)
2: for θ ∈ Θ do
3:   for 1 ≤ i ≤ N do
4:     LA(i, θ) = calcError(A, θ, Sitrain, Sitest)
5:   end for
6:   Compute the median LAmed(θ) of {LA(i, θ)}
7: end for
8: return θ* = arg minθ LAmed(θ)
}
    
```

### CV: Reminder (on board)



Training Set

Test Set

Value of "k":  
•10-fold; 5-fold  
•Leave one out  
•How about 2 fold cv?

## Measuring the error

- Loss function - here consider 0/1-loss
- Confusion Matrix:

Classification \ True Class	Yes	No
Yes	TP: true positive	FN: false negative
No	FP: false positive	TN: true negative

## Other methods

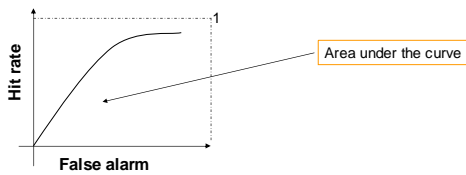
- 5X2 Cross-Validation
  - Divide dataset randomly to: training & test
  - Repeat 5 times
- Bootstrapping
  - Draw sample with replacement
  - Approximately 2/3 training & 1/3 test
- Hold-out
  - Similar to 5X2 CV

## Presenting results: Receiver Operating Characteristic curves

- “Hit rate” -vs- “False alarm”

$$\frac{|TP|}{|TP| + |FN|}$$

$$\frac{|FP|}{|FP| + |TN|}$$



## Statistical Significance

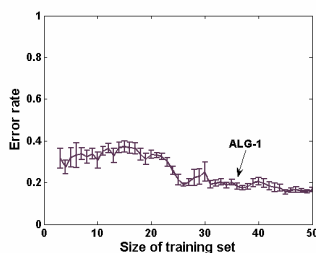
(Description on board)

- Bernoulli experiment
- Mean and standard-error-of-the-mean
- Confidence intervals
- Comparing two classifiers (McNemar's, and k-fold CV paired t Tests)
- Non-parametric tests (sign tests)

Truth \ Decision	Accept	Reject
True	Correct	Type I Error
False	Type II Error	Correct (power)

## Presenting Results: Learning curve

- Learning Curve



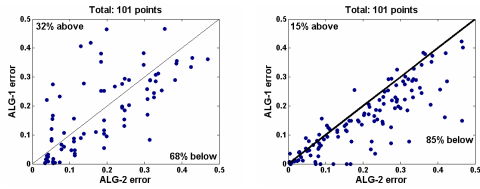
## Results: Cost Curves [Drummond & Holt]

- P(+)- fraction of positive instances
- c(+|-)- cost of false positive
- c(-|+)- cost of false negative

$$PCF = \frac{P(+)\ c(-|+)}{P(+)\ c(-|+) + P(-)\ c(+|-)}$$

## Presenting results (cont.)

- Scatter plots



## Presenting Results

- Table:

data	Alg.1	Alg.2	Alg.3
text	0.14 ± 0.01	0.20 ± 0.10	0.15 ± 0.10
usps	0.90 ± 0.06	<b>0.30 ± 0.00</b>	0.85 ± 0.00
monk	0.11 ± 0.01	<b>0.02 ± 0.01</b>	<b>0.02 ± 0.02</b>

**Table 1.** The error (%) of ALG-1,2,3.  
The lowest error in each row appears in bold font.