10-610 The KDD Lab

Intro: Outcome Analysis

Sebastian Thrun
Carnegie Mellon University

www.cs.cmu.edu/~10610

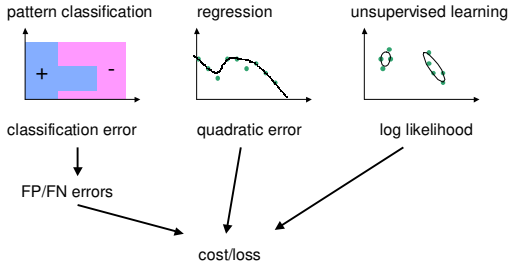
Problem 1

- You find out on testing data, your speech recognizer can recognize sentences with 68% word accuracy, whereas previous recognizers achieve 60%. Would you advice a company to adopt your speech recognizer?

Problem 2

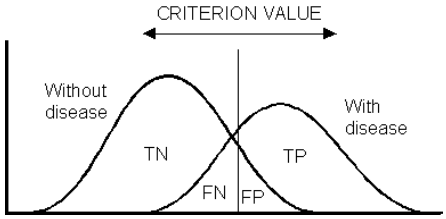
- On testing data, your data mining algorithm can predict emergency C-sections with 68% accuracy, whereas a previous $1,000 test achieves 60% accuracy. Do you recommend to replace the previous test by your new method?

Characterize: What Should We Worry about?



$$\varepsilon_\theta = \int_D L(x, f(x,\theta))\ p_D(x)\ dx$$

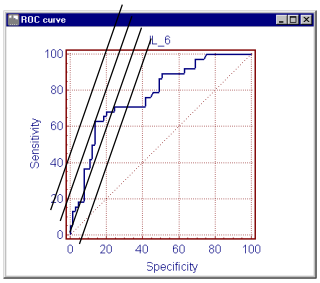
ROC Curves (ROC=Receiver Operating Characteristic)



Error Types

- Type I error, alpha error, false positive: Probability of accepting hypothesis if not true
- Type II error, beta error, false negative: Probability of rejecting hypothesis when it is true

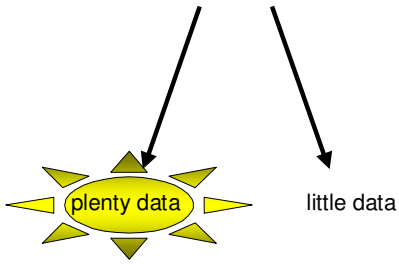
ROC Curves (ROC=Receiver Operating Characteristic)



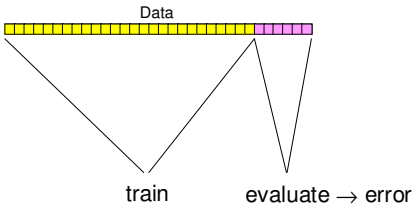
ROC Curves (ROC=Receiver Operating Characteristic)

- **Sensitivity**: probability that a test result will be positive when the disease is present
$$\frac{\text{true positive}}{\text{true positive} + \text{false negative}}$$
- **Specificity**: probability that a test result will be negative when the disease is not present
$$\frac{\text{true negative}}{\text{true negative} + \text{false positive}}$$
- **Positive likelihood ratio**: ratio between the probability of a positive test result given the presence of the disease and the probability of a positive test result given the absence of the disease
$$\frac{\text{true positive}}{\text{true positive}}$$
- **Negative likelihood ratio**: ratio between the probability of a negative test result given the presence of the disease and the probability of a negative test result given the absence of the disease
$$\frac{\text{true negative}}{\text{true negative}}$$
- **Positive predictive value (PPV)**: probability that the disease is present when the test is positive
$$\frac{\text{true positive}}{\text{true positive} + \text{false positive}}$$
- **Negative predictive value (NPV)**: probability that the disease is not present when the test is negative
$$\frac{\text{true negative}}{\text{true negative} + \text{false negative}}$$

Evaluating Machine Learning Algorithms



Holdout Set



- Often also used for parameter optimization

Example:

- Hypothesis misclassifies 12 out of 40 examples in cross validation set S.
- Q: What will the "true" error on future examples?
- A:

Finite Cross-Validation Set

- True error: (true risk) $e_D = \int_D |y - f(x,\theta)|\ p(x,y)\ dx, y$
  D = all data
- Test error: (empirical risk) $\hat{e}_S = \frac{1}{m} \sum_{\langle x,y \rangle \in S} |y - f(x,\theta)|$
  m = #test samples; S = test data

Confidence Intervals (See Mitchell 97)

If

- S contains m examples, drawn independently
- $m \geq 30$

Then

- With approximately 95% probability, the true error e_D lies in the interval

$$\hat{e}_S \pm 1.96\sqrt{\frac{\hat{e}_S\,(1-\hat{e}_S)}{m}}$$

Example:

- Hypothesis misclassifies 12 out of 40 examples in cross validation set S.
- Q: What will the "true" error on future examples?
- A: With 95% confidence, the true error will be in the interval:

$$[0.16;0.44] \quad \approx \hat{e}_S \pm 1.96\sqrt{\frac{\hat{e}_S\,(1-\hat{e}_S)}{m}}$$

$$m = 40 \qquad \hat{e}_S = \frac{12}{40} = 0.3 \qquad 1.96\sqrt{\frac{\hat{e}_S\,(1-\hat{e}_S)}{m}} \approx 0.14$$

Confidence Intervals (See Mitchell 97)

If

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

Then

- With approximately N% probability, the true error e_D lies in the interval

$$\hat{e}_S \pm z_N\sqrt{\frac{\hat{e}_S\,(1-\hat{e}_S)}{m}}$$

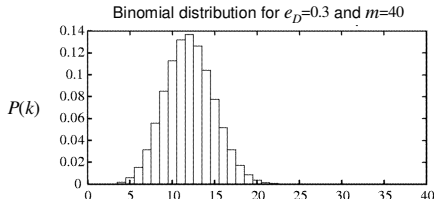
N%	50%	68%	80%	90%	95%	98%	99%
z_N	0.67	1.0	1.28	1.64	1.96	2.33	2.58

Finite Cross-Validation Set

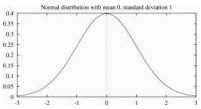
- True error: $e_D = \int_D |y - f(x,\theta)|\; p(x,y)\; dx, y$
- Test error: $\hat{e}_S = \frac{1}{m}\sum_{\langle x,y\rangle \in S} |y - f(x,\theta)|$
- Number of test errors: Is Binomially distributed:

$$p\left(\sum_{\langle x,y\rangle \in S} |y - f(x,\theta)| = k\right) = \frac{m!}{k!(m-k)!}\;(e_D)^k\,(1-e_D)^{n-k}$$

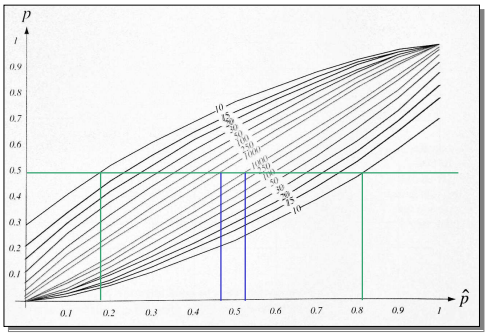
Binomial Distribution



Approximates Normal distribution (Central Limit Theorem)



95% Confidence Intervals



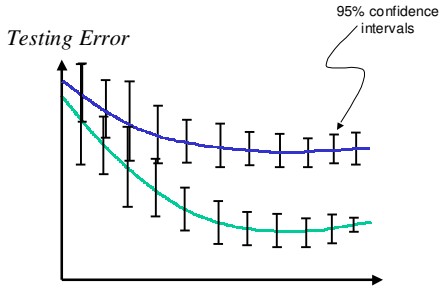
Question

$$\hat{e}_S \pm z_N \sqrt{\frac{\hat{e}_S\,(1-\hat{e}_S)}{m}}$$

- What's the difference between variance and confidence intervals?

Basically a factor $\sqrt{\frac{1}{m}}$

Common Performance Plot

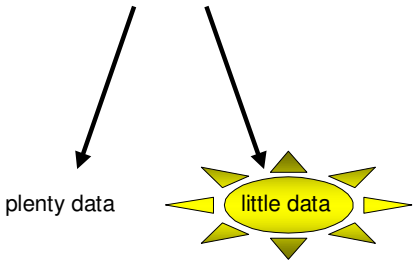


Comparing Different Hypotheses

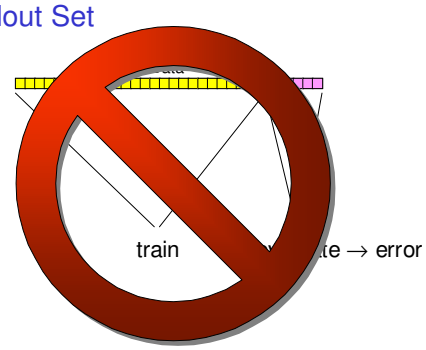
- True difference: $d = e_D(\theta_1) - e_D(\theta_2)$
- Test set difference: $\hat{d} = \hat{e}_S(\theta_1) - \hat{e}_S(\theta_2)$
- 95% Confidence interval:

$$\hat{d} \pm 1.96 \sqrt{\frac{\hat{e}_S(\theta_1)\,(1-\hat{e}_S(\theta_1))}{m_1} + \frac{\hat{e}_S(\theta_2)\,(1-\hat{e}_S(\theta_2))}{m_2}}$$

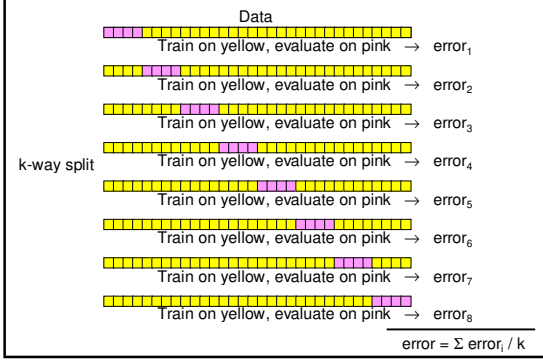
Evaluating Machine Learning Algorithms



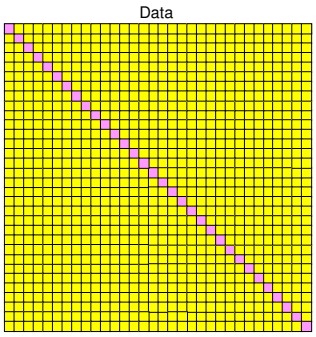
Holdout Set



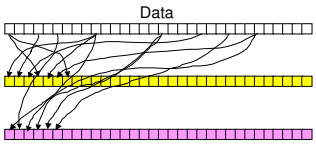
k-fold Cross Validation



The Jackknife



The Bootstrap



- Train on yellow, evaluate on pink → error
- Repeat and average

What's the Problem?

- Confidence intervals assume independence. But our individual estimates are dependent. ☹

Comparing Different Hypotheses: Paired t test

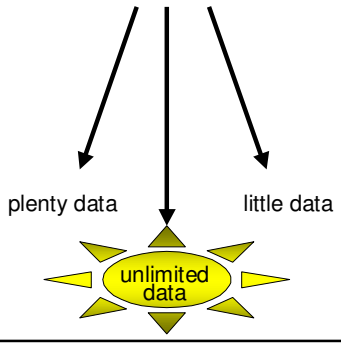
- True difference: $d = e_D(\theta_1) - e_D(\theta_2)$
- For each partition k: $\hat{d}_k = \hat{e}_{S,k}(\theta_1) - \hat{e}_{S,k}(\theta_2)$ (test error for partition k)
- Average: $\hat{d} = \frac{1}{k}\sum_{i=1}^{k} \hat{d}_i$
- N% Confidence interval:

$$\hat{d} \pm t_{N,k-1} \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^{k} (\hat{\delta}_i - \hat{\delta})^2}$$

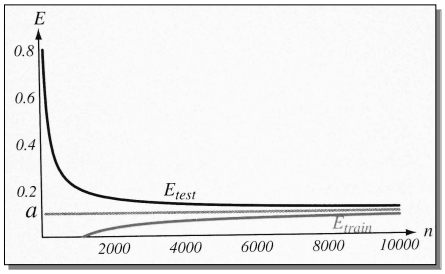
(N is confidence level; k-1 is degrees of freedom)

	90%	95%	98%	99%
$\nu=2$	2.92	4.30	6.96	9.92
$\nu=5$	2.02	2.57	3.36	4.03
$\nu=10$	1.81	2.23	2.76	3.17
$\nu=20$	1.72	2.09	2.53	2.84
$\nu=30$	1.70	2.04	2.46	2.75
$\nu=120$	1.66	1.98	2.36	2.62
$\nu=\infty$	1.64	1.96	2.33	2.58

Evaluating Machine Learning Algorithms



Asymptotic Prediction



- Useful for very large data sets

Summary

- Know your loss function!
- Finite testing data: report confidence intervals
- Scarce data: Repartition training/testing set
- Asymptotic prediction: exponential