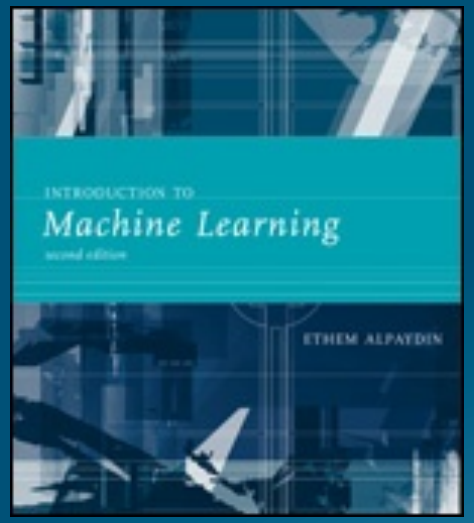


Lecture Slides for

INTRODUCTION TO

Machine Learning

2nd Edition

ETHEM ALPAYDIN

alpaydin@boun.edu.tr
http://www.cmpe.boun.edu.tr/~ethem/i2ml2e

Edited by Zehra Çataltepe, ITU BLG527E Machine Learning
Sep 25, 2013

CHAPTER 19:

Design and Analysis of Machine Learning Experiments

Introduction

- Questions:
  - Assessment of the expected error of a learning algorithm:
    - Is the error rate of 1-NN less than 2%?
  - Comparing the expected errors of two algorithms:
    - Is *k*-NN more accurate than MLP ?
    - Should k be 1 or 3 in kNN?
- Training/validation/test sets
- Resampling methods: *K*-fold cross-validation

Important Notes

- Using our experiments we only show that a particular algorithm is better than others for this specific dataset. No algorithm can be the best on all possible datasets *(see NFL (No Free Lunch) Theorems, Wolpert 1995)*
- Once you decide on learning algorithm, parameter setting using the training-validation partitioned data, use ALL (training+validation) data to train your final model.
- Use a separate test set (not used for validation) to report the expected test error, not the validation error. *(In papers, people do report validation error though.)*

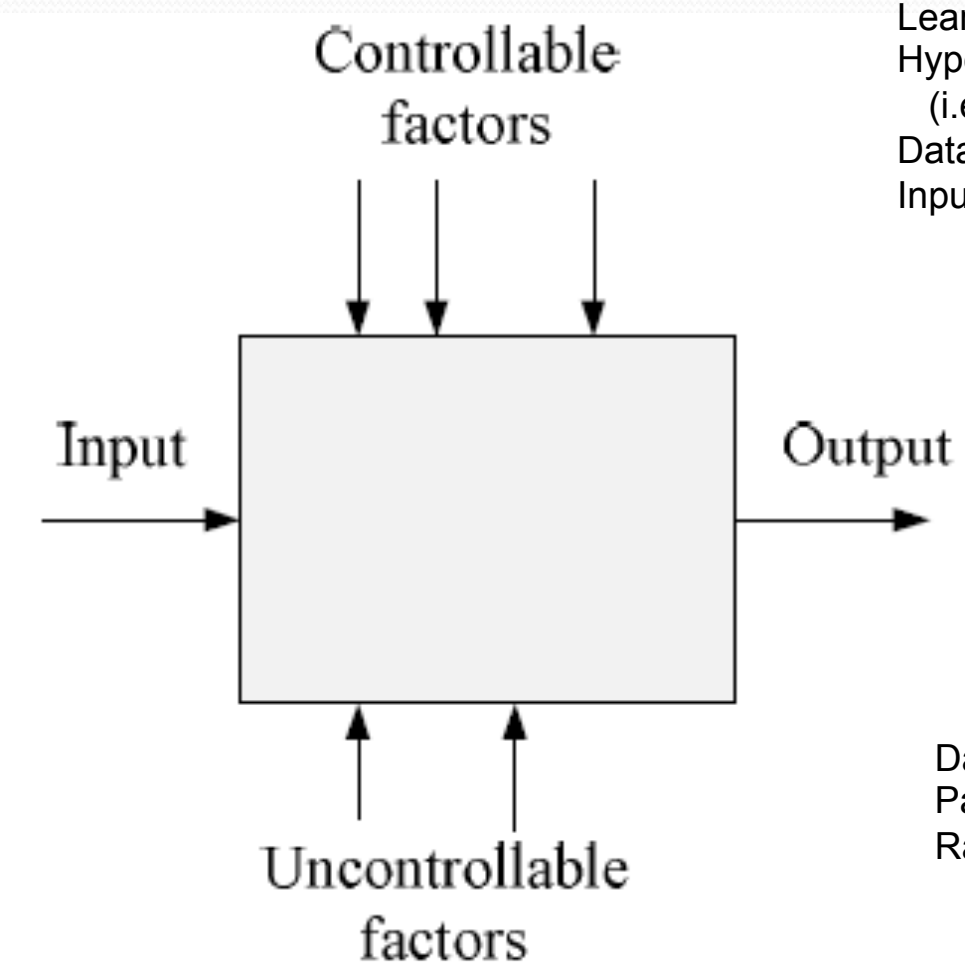
Algorithm Preference

- Criteria (Application-dependent):
  - Misclassification error, or risk (loss functions)
  - Training time/space complexity
  - Testing time/space complexity
  - Interpretability
  - Easy programmability
- Cost-sensitive learning (Elkan 2001)

Elkan, C. (2001, August). The foundations of cost-sensitive learning. In International joint conference on artificial intelligence (Vol. 17, No. 1, pp. 973-978). LAWRENCE ERLBAUM ASSOCIATES LTD.

determine the classifier decision when each class may have a different cost. Resampling/reweighting of instances is proposed to achieve the optimal decision boundaries based on a specific cost matrix.

Factors and Response



Learning algorithm
Hyperparameters of the algorithm
(i.e. no of hidden layers/unit in a MLP)
Dataset used
Input representation

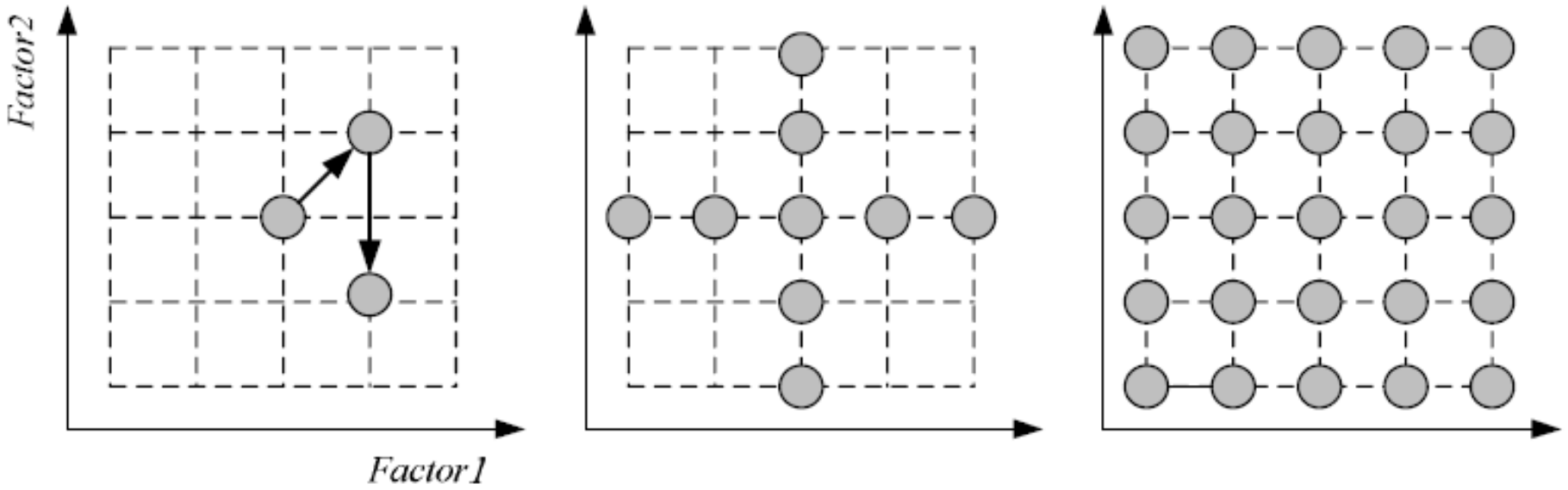
used to generate the response variable
- classification error
- expected risk
- precision, recall
- auc

Data noise
Particular resampled training set
Randomness in optimization process

Factors and Response: Example

- Controllable Factors:
  - PCA to reduce dimension to d
  - Knn classifier with k
- Response:
  - Classification error on validation set
- Find the setting of k and d for the best response

Strategies of Experimentation



(a) Best guess (b) One factor at a time (c) Factorial design

F Factors with L levels each:
Cost: O(LF)

Grid search
Best approach
F Factors with L levels each:
Cost: $O(L^F)$ (☹ !!!)

Response surface design for approximating and maximizing the response function in terms of the controllable factors

Experimental Design

- Randomization
  - independent order of experiments)
- Replication
  - For the same configuration, run the experiment a number of times → *see Cross Validation*
- Blocking
  - Reduce variability due to nuisance factors
  - *Pairing, paired testing*: compare algorithms trained on the same resampled training subsets

Guidelines for ML experiments

A. Aim of the study
B. Selection of the response variable
C. Choice of factors and levels
D. Choice of experimental design
E. Performing the experiment
F. Statistical Analysis of the Data
   - Is Algorithm A more accurate than Algorithm B on this dataset?
G. Conclusions and Recommendations

Experimental Design: Replication

Resampling and K-Fold Cross-Validation

- The need for multiple training/validation sets $\{X_i,V_i\}_i$: Training/validation sets of fold i
- K-fold cross-validation: Divide X into k, X_i,i=1,...,K

$$\mathcal{V}_1 = \mathcal{X}_1 \quad \mathcal{T}_1 = \mathcal{X}_2 \cup \mathcal{X}_3 \cup \cdots \cup \mathcal{X}_K$$
$$\mathcal{V}_2 = \mathcal{X}_2 \quad \mathcal{T}_2 = \mathcal{X}_1 \cup \mathcal{X}_3 \cup \cdots \cup \mathcal{X}_K$$
$$\vdots$$
$$\mathcal{V}_K = \mathcal{X}_K \quad \mathcal{T}_K = \mathcal{X}_1 \cup \mathcal{X}_2 \cup \cdots \cup \mathcal{X}_{K-1}$$

- T_i share K-2 parts
- Report the validation error on each Vi,
- Report the average and std of the validation error
- See also the hypothesis testing part below.

5×2 Cross-Validation

- 5 times 2 fold cross-validation (Dietterich, 1998)

$$\mathcal{T}_1 = \mathcal{X}_1^{(1)} \quad \mathcal{V}_1 = \mathcal{X}_1^{(2)}$$
$$\mathcal{T}_2 = \mathcal{X}_1^{(2)} \quad \mathcal{V}_2 = \mathcal{X}_1^{(1)}$$
$$\mathcal{T}_3 = \mathcal{X}_2^{(1)} \quad \mathcal{V}_3 = \mathcal{X}_2^{(2)}$$
$$\mathcal{T}_4 = \mathcal{X}_2^{(2)} \quad \mathcal{V}_4 = \mathcal{X}_2^{(1)}$$
$$\vdots$$
$$\mathcal{T}_9 = \mathcal{X}_5^{(1)} \quad \mathcal{V}_9 = \mathcal{X}_5^{(2)}$$
$$\mathcal{T}_{10} = \mathcal{X}_5^{(2)} \quad \mathcal{V}_{10} = \mathcal{X}_5^{(1)}$$

Leave-One-Out Cross Validation

- Leave-One-Out : Sometimes also called LOO
- Use especially if there are not many data samples and hence can not afford to leave out a lot of examples for validation.
- Do N (no of data samples) folds.
- At fold i (i=1..N), use the ith sample for validation and all the remaining samples for training.

Bootstrapping

Images: rudebaguette.com, lanternlegal.com

- Draw instances from a dataset *with replacement*
- Prob that we do not pick an instance after N draws

$$\left(1-\frac{1}{N}\right)^{N} \approx e^{-1} = 0.368$$

that is, only 36.8% is new!

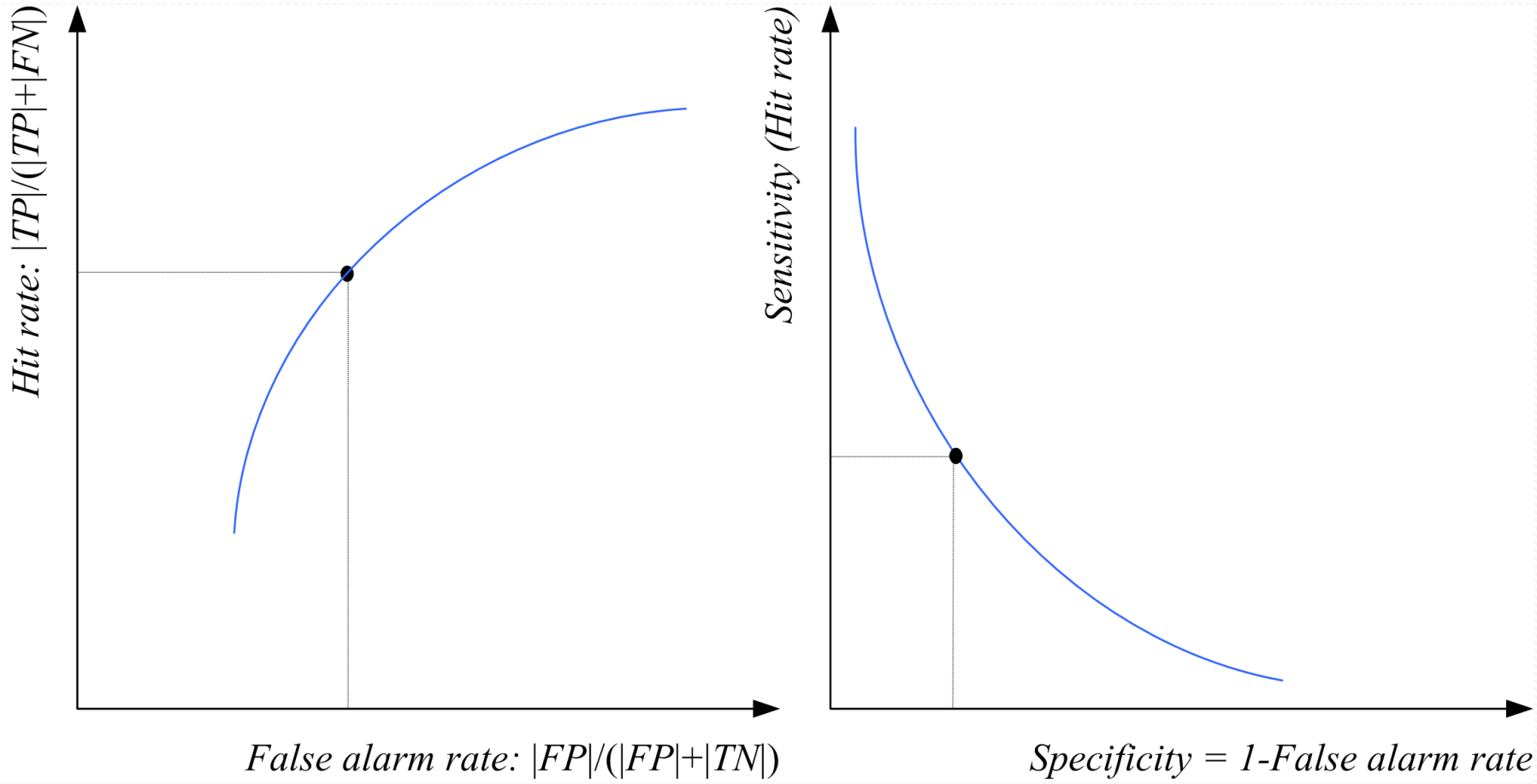
Response Variable: Measuring Classifier Performance

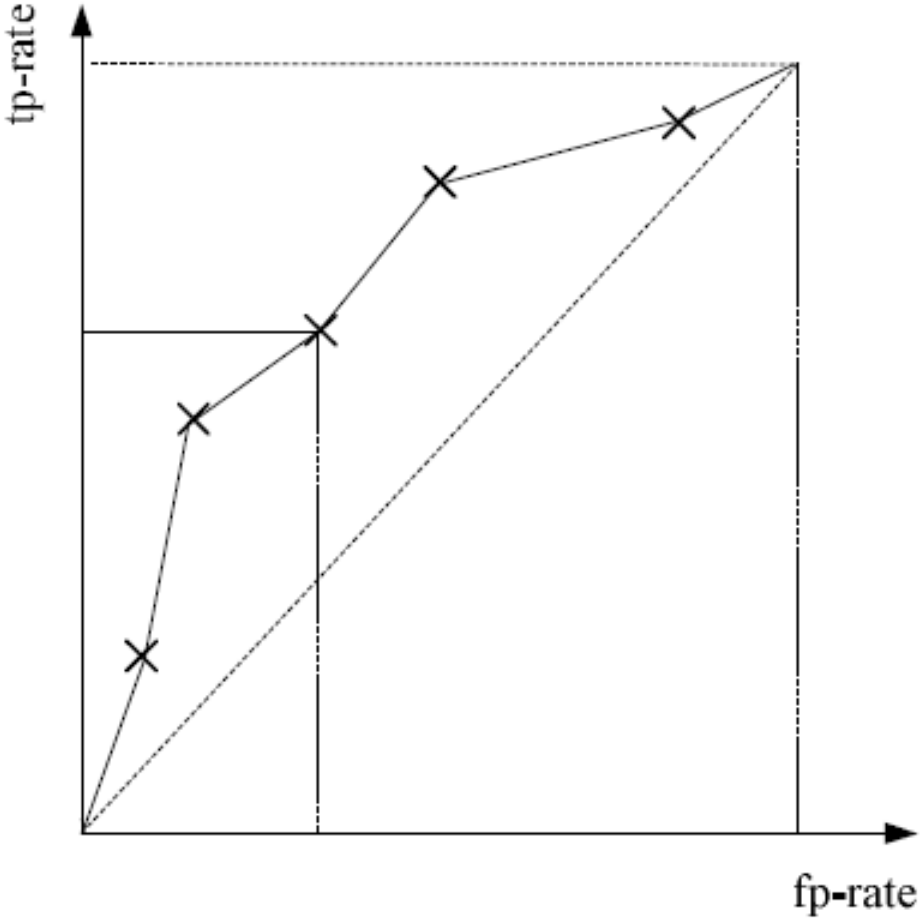
Measuring Error

	Predicted class	
True Class	Yes	No
Yes	TP: True Positive	FN: False Negative
No	FP: False Positive	TN: True Negative

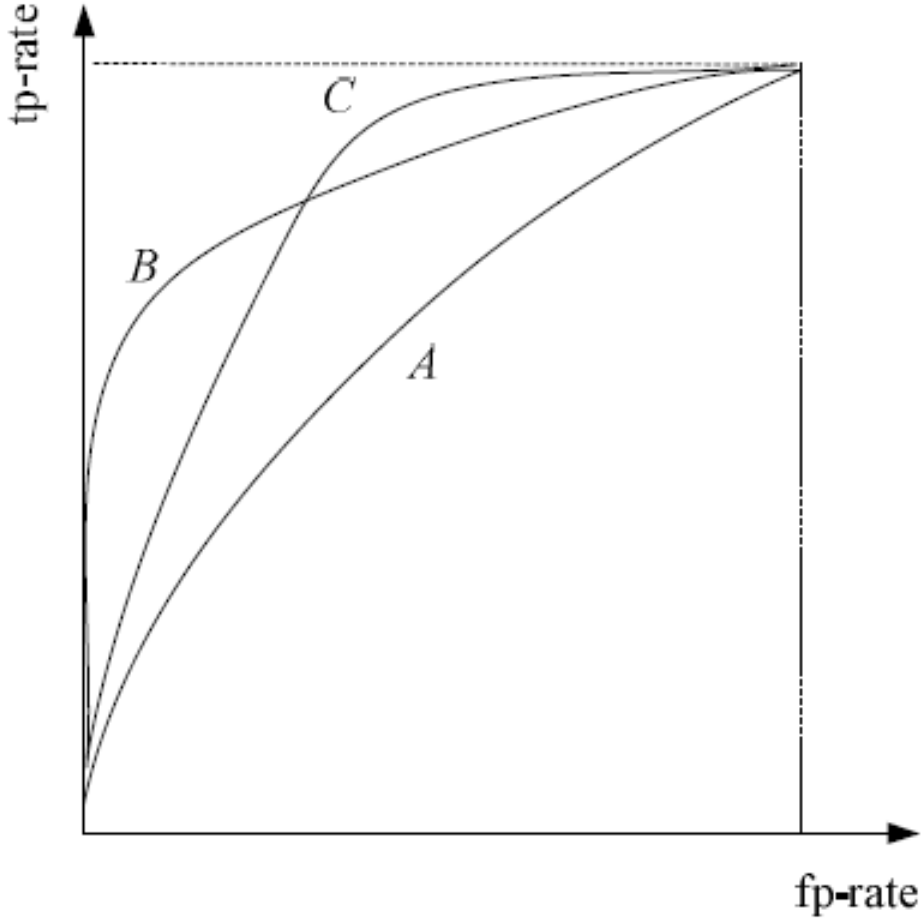
- Error rate = # of errors / # of instances = (FN+FP) / N
- Recall = # of found positives / # of positives
  = TP / (TP+FN) = sensitivity = hit rate
- Precision = # of found positives / # of found
  = TP / (TP+FP)
- Specificity = TN / (TN+FP)
- False alarm rate = FP / (FP+TN) = 1 - Specificity

ROC Curve


Hit rate: |TP|/(|TP|+|FN|)
False alarm rate: |FP|/(|FP|+|TN|)
Sensitivity (Hit rate)
Specificity = 1-False alarm rate



(a) Example ROC curve



(b) Different ROC curves for different classifiers

AUC: Area Under the ROC Curve

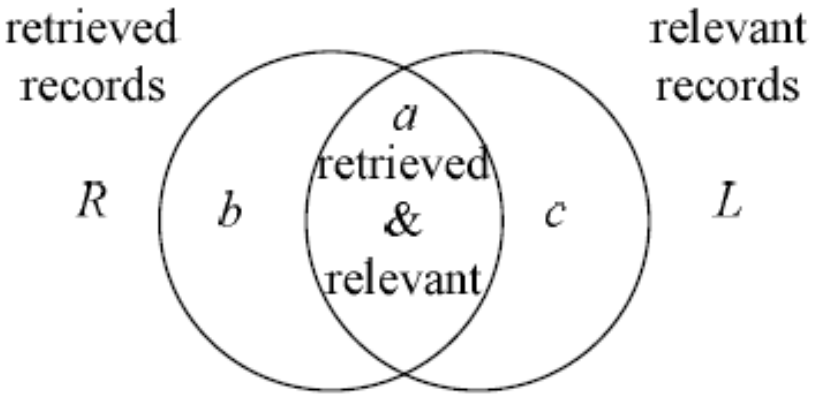
Used to compare classifiers based on all possible operating points (i.e. thresholds decided for positive or negative class).

Computed by taking the area under the ROC curve.

Maximum possible is 1.

The higher the AUC the better the classifier.

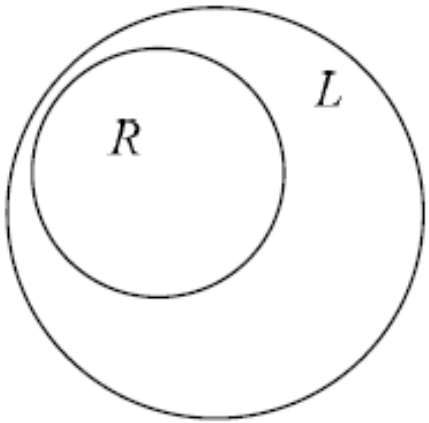
Precision and Recall



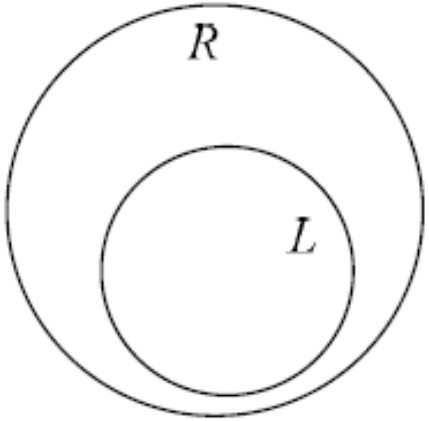
$$\text{Precision: } \frac{a}{a + b}$$

$$\text{Recall: } \frac{a}{a + c}$$

(a) Precision and recall



(b) Precision = 1



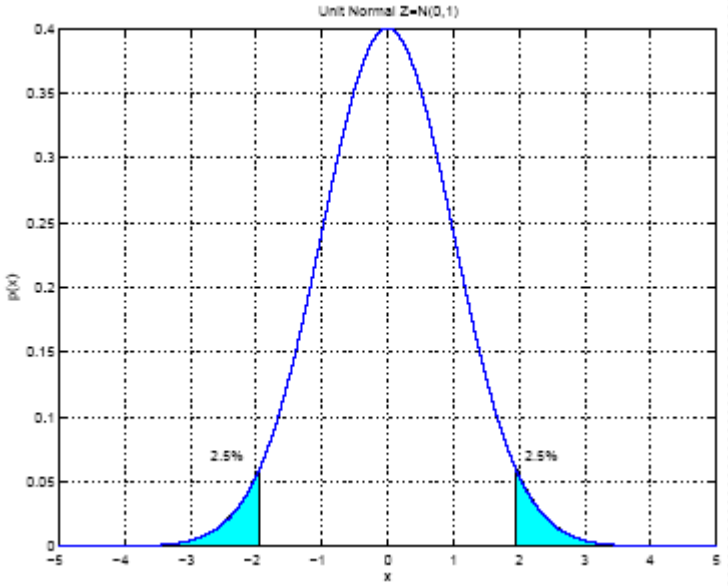
(c) Recall = 1

Assessing a Classification Algorithm's Performance

- For classification error *(same methodolgy can be applied to regression also, if the appropriate parametric form for sampling distribution can be obtained)*
- Same Data Set (Assume distribution on computed errors)
  - Single Algorithm
    - Single Train-Validation Set
      - Binomial Test
      - Approximate Normal Test (need large N)
    - K Train Validation Sets
      - t test
  - Two Learning Algorithms
    - McNemar's Test
    - K-Fold Cross-Validated Paired t Test
    - 5x2 cv Paired t test
    - 5x2 cv Paired F test
  - Multiple Algorithms
    - Analysis of Variance (ANOVA)
- Multiple Datasets (Can not assume the same distribution, non parametric)
  - Two Algorithms
    - Sign Test
    - Wilcoxon Signed Rank Test
  - Multiple Algorithms
    - Kruskal-Wallis (nonparametric version of ANOVA)
    - Tukey's Test (pairwise comparison of ranks)

Interval Estimation (Review)

- $X = \{ x^t \}_t$ where $x^t \sim N(\mu, \sigma^2)$
- $m \sim N(\mu, \sigma^2/N)$



$$\sqrt{N}\frac{(m-\mu)}{\sigma} \sim Z$$ Unit normal

$$P\left\{-1.96 < \sqrt{N}\frac{(m-\mu)}{\sigma} < 1.96\right\} = 0.95$$

$$P\left\{m - 1.96\frac{\sigma}{\sqrt{N}} < \mu < m + 1.96\frac{\sigma}{\sqrt{N}}\right\} = 0.95$$

$$P\left\{m - z_{\alpha/2}\frac{\sigma}{\sqrt{N}} < \mu < m + z_{\alpha/2}\frac{\sigma}{\sqrt{N}}\right\} = 1 - \alpha$$

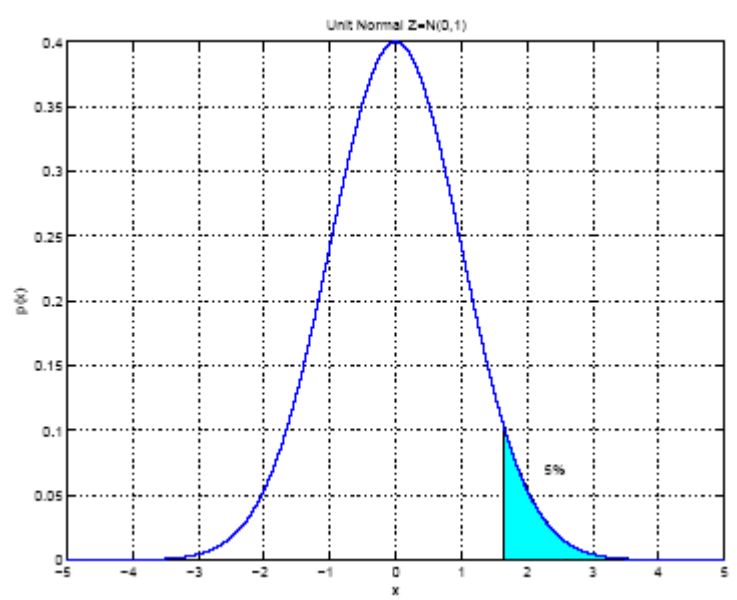
100(1- α) percent (two sided) confidence İnterval for μ

100(1- α) percent (one sided) confidence interval

$$P\left\{\sqrt{N}\frac{(m-\mu)}{\sigma}<1.64\right\}=0.95$$

$$P\left\{m-1.64\frac{\sigma}{\sqrt{N}}<\mu\right\}=0.95$$

$$P\left\{m-z_{\alpha}\frac{\sigma}{\sqrt{N}}<\mu\right\}=1-\alpha$$



When σ^2 is not known:

$$S^2=\sum_t\left(x^t-m\right)^2/(N-1) \qquad \frac{\sqrt{N}(m-\mu)}{S}\sim t_{N-1}$$

$$P\left\{m-t_{\alpha/2,N-1}\frac{S}{\sqrt{N}}<\mu<m+t_{\alpha/2,N-1}\frac{S}{\sqrt{N}}\right\}=1-\alpha$$

Hypothesis Testing

- Reject a null hypothesis if not supported by the sample with enough confidence
- X = $\{ x^t \}_t$ where $x^t \sim N (\mu, \sigma^2)$

$H_0: \mu = \mu_0$ vs. $H_1: \mu \neq \mu_0$

Accept H_0 with level of significance α if μ_0 is in the

100(1- α) confidence interval

$$\frac{\sqrt{N}(m-\mu_0)}{\sigma} \in (-z_{\alpha/2}, z_{\alpha/2})$$

Two-sided test

	Decision	
Truth	Accept	Reject
True	Correct	Type I error
False	Type II error	Correct (Power)

- One-sided test: $H_0: \mu \leq \mu_0$ vs. $H_1: \mu > \mu_0$

  Accept if $$\frac{\sqrt{N}(m-\mu_0)}{\sigma} \in (-\infty, z_\alpha)$$

- Variance unknown: Use t, instead of z

  Accept $H_0: \mu = \mu_0$ if $$\frac{\sqrt{N}(m-\mu_0)}{S} \in (-t_{\alpha/2,N-1}, t_{\alpha/2,N-1})$$

Example 12.4, p.350

Cross-validation

The following table gives a possible result of evaluating three learning algorithms on a data set with 10-fold cross-validation:

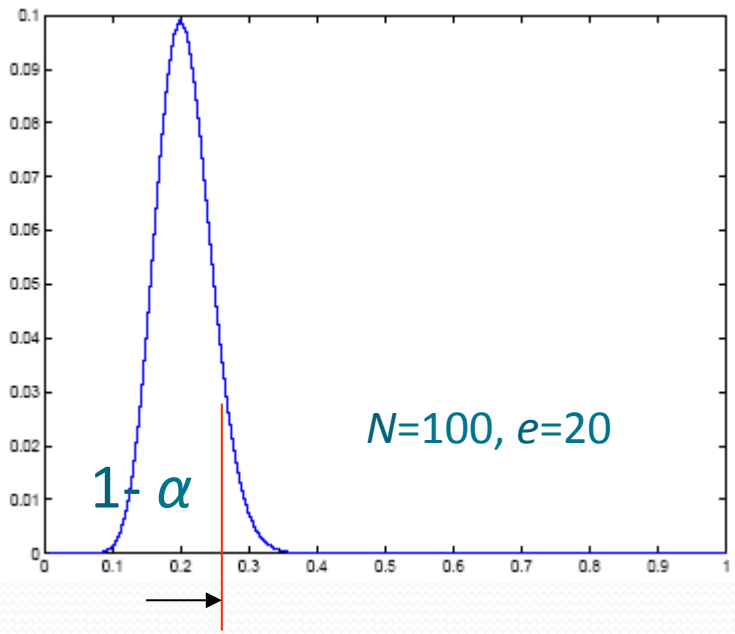
Fold	*Naive Bayes*	*Decision tree*	*Nearest neighbour*
1	0.6809	0.7524	0.7164
2	0.7017	0.8964	0.8883
3	0.7012	0.6803	0.8410
4	0.6913	0.9102	0.6825
5	0.6333	0.7758	0.7599
6	0.6415	0.8154	0.8479
7	0.7216	0.6224	0.7012
8	0.7214	0.7585	0.4959
9	0.6578	0.9380	0.9279
10	0.7865	0.7524	0.7455
avg	0.6937	0.7902	0.7606
stdev	0.0448	0.1014	0.1248

The last two lines give the average and standard deviation over all ten folds. Clearly the decision tree achieves the best result, but should we completely discard nearest neighbour?

Assessing Error: $H_0: p \leq p_0$ vs. $H_1: p > p_0$

- Single training/validation set: Binomial Test

  If error prob is p_0, prob that there are e errors or less in N validation trials is



$$P\{X \leq e\} = \sum_{j=1}^{e} \binom{N}{j} p_0^{\,j} \left(1 - p_0^{\,j}\right)^{N-j}$$

Accept if this prob is less than 1- α

H_0: $\mu = \mu_0$ vs. H_1: $\mu \neq \mu_0$

Accept H_0 with level of significance α if μ_0 is in the $100(1-\alpha)$ confidence interval

$$\frac{\sqrt{N}(m-\mu_0)}{\sigma} \in (-z_{\alpha/2}, z_{\alpha/2})$$

Two-sided test

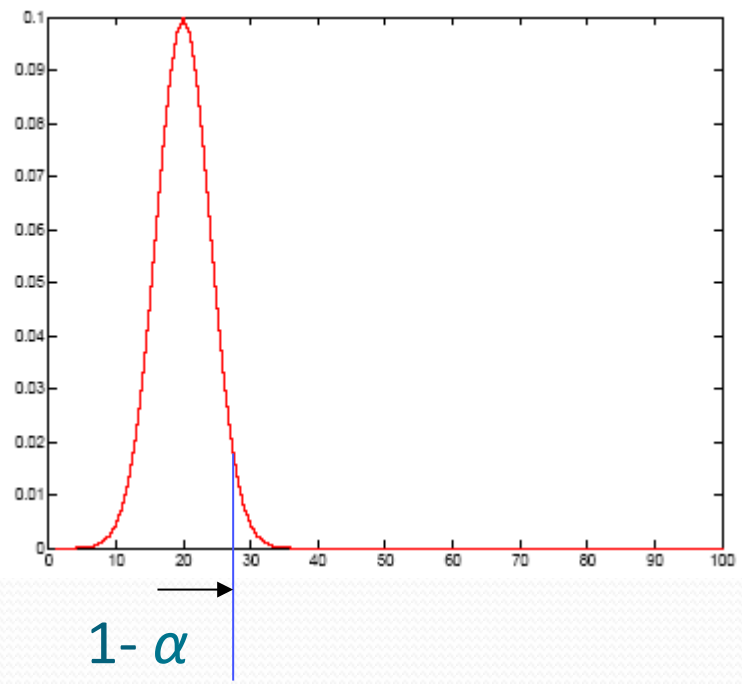
Null hypothesis and p-value

We can, however, use similar reasoning to test a particular *null hypothesis* we have about a.

- For example, suppose our null hypothesis is that the true accuracy is 0.5 and that the standard deviation derived from the binomial distribution is therefore $\sqrt{0.5(1-0.5)/100} = 0.05$.
- Given our estimate of 0.80, we then calculate the *p-value*, which is the probability of obtaining a measurement of 0.80 or higher given the null hypothesis.
- The p-value is then compared with a pre-defined significance level, say $\alpha = 0.05$: this corresponds to a confidence of 95%.
- The null hypothesis is rejected if the p-value is smaller than α; in our case this applies since $p = 1.9732 \cdot 10^{-9}$.

Normal Approximation to the Binomial

- Number of errors X is approx N with mean Np_0 and var $Np_0(1-p_0)$



$$\frac{X - Np_0}{\sqrt{Np_0(1-p_0)}} \sim Z$$

Accept if this prob for $X = e$ is less than $z_{1-\alpha}$

$$p_1 \quad \mathcal{V}_1 = \mathcal{X}_1 \quad \mathcal{T}_1 = \mathcal{X}_2 \cup \mathcal{X}_3 \cup \cdots \cup \mathcal{X}_K$$
$$p_2 \quad \mathcal{V}_2 = \mathcal{X}_2 \quad \mathcal{T}_2 = \mathcal{X}_1 \cup \mathcal{X}_3 \cup \cdots \cup \mathcal{X}_K$$
$$\vdots$$
$$p_K \quad \mathcal{V}_K = \mathcal{X}_K \quad \mathcal{T}_K = \mathcal{X}_1 \cup \mathcal{X}_2 \cup \cdots \cup \mathcal{X}_{K-1}$$

t Test

- Multiple training/validation sets
- $x^t_i = 1$ if instance t misclassified on fold i
- Error rate of fold i: $p_i = \frac{\sum_{t=1}^{N} x_i^t}{N}$
- With m and s^2 average and var of p_i , we accept p_0 or less error if

$$\frac{\sqrt{K}(m - p_0)}{s} \sim t_{K-1}$$

is less than $t_{\alpha,K\text{-}1}$

Comparing Classifiers:

$H_0: \mu_0 = \mu_1$ vs. $H_1: \mu_0 \neq \mu_1$

- Single training/validation set: McNemar's Test

e_{00}: Number of examples misclassified by both	e_{01}: Number of examples misclassified by 1 but not 2
e_{10}: Number of examples misclassified by 2 but not 1	e_{11}: Number of examples correctly classified by both

- Under H_0, we expect $e_{01} = e_{10} = (e_{01} + e_{10})/2$

$$\frac{\left(|e_{01} - e_{10}| - 1\right)^2}{e_{01} + e_{10}} \sim \chi_1^2$$

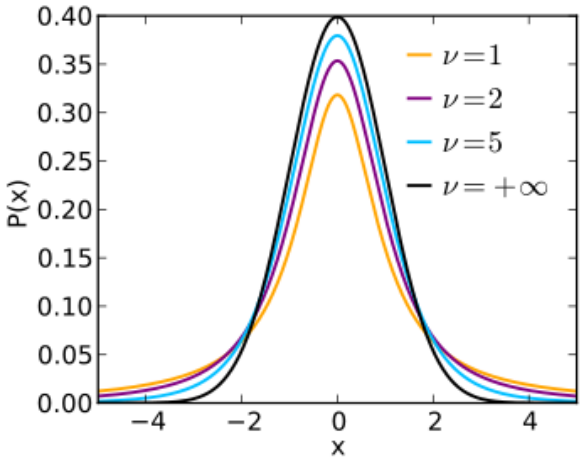
Accept if $< X^2_{\alpha,1}$

Because we are summing squares of normals
Includes Edward's Correction for continuity

K-Fold CV Paired *t* Test

$$
\begin{array}{cccc}
p_1^2 & p_1^1 & \mathcal{V}_1 = \mathcal{X}_1 & \mathcal{T}_1 = \mathcal{X}_2 \cup \mathcal{X}_3 \cup \cdots \cup \mathcal{X}_K \\
p_2^2 & p_2^1 & \mathcal{V}_2 = \mathcal{X}_2 & \mathcal{T}_2 = \mathcal{X}_1 \cup \mathcal{X}_3 \cup \cdots \cup \mathcal{X}_K \\
\vdots & \vdots & \vdots & \\
p_K^2 & p_K^1 & \mathcal{V}_K = \mathcal{X}_K & \mathcal{T}_K = \mathcal{X}_1 \cup \mathcal{X}_2 \cup \cdots \cup \mathcal{X}_K
\end{array}
$$

- Use *K*-fold cv to get *K* training/validation folds
- p_i^1, p_i^2: Errors of classifiers 1 and 2 on fold *i*
- $p_i = p_i^1 - p_i^2$: Paired difference on fold *i*
- The null hypothesis is whether p_i has mean 0



$$H_0 : \mu = 0 \quad \text{vs.} \quad H_0 : \mu \neq 0$$

$$m = \frac{\sum_{i=1}^{K} p_i}{K} \qquad s^2 = \frac{\sum_{i=1}^{K} (p_i - m)^2}{K-1}$$

$$\frac{\sqrt{K}(m-0)}{s} = \frac{\sqrt{K} \cdot m}{s} \sim t_{K-1} \quad \text{Accept if in } \left(-t_{\alpha/2,K-1}, t_{\alpha/2,K-1}\right)$$

Significance testing in cross-validation: the paired t-test

- For a pair of algorithms we calculate the difference in accuracy on each fold; this difference is normally distributed if the two accuracies are. Our null hypothesis is that the true difference is 0, so that any differences in performance are attributed to chance. We calculate a p-value using the normal distribution, and reject the null hypothesis if the p-value is below our significance level α.
- The one complication is that we don't have access to the true standard deviation in the differences, which therefore needs to be estimated. This introduces additional uncertainty into the process, which means that the sampling distribution is bell-shaped like the normal distribution but slightly more heavy-tailed. This distribution is referred to as the *t-distribution*.
- The extent to which the t-distribution is more heavy-tailed than the normal distribution is regulated by the number of *degrees of freedom*: in our case this is equal to 1 less than the number of folds (since the final fold is completely determined by the other ones).

Example 12.6, p.353

Paired t-test

The numbers show pairwise differences in each fold. The null hypothesis in each case is that the differences come from a normal distribution with mean 0 and unknown standard deviation.

Fold	*NB–DT*	*NB–NN*	*DT–NN*
1	-0.0715	-0.0355	0.0361
2	-0.1947	-0.1866	0.0081
3	0.0209	-0.1398	-0.1607
4	-0.2189	0.0088	0.2277
5	-0.1424	-0.1265	0.0159
6	-0.1739	-0.2065	-0.0325
7	0.0992	0.0204	-0.0788
8	-0.0371	0.2255	0.2626
9	-0.2802	-0.2700	0.0102
10	0.0341	0.0410	0.0069
avg	-0.0965	-0.0669	0.0295
stdev	0.1246	0.1473	0.1278
p-value	**0.0369**	**0.1848**	**0.4833**

The p-value in the last line of the table is calculated by means of the t-distribution with $k-1=9$ degrees of freedom, and only the difference between the naive Bayes and decision tree algorithms is found significant at $\alpha=0.05$.

5×2 cv Paired *t* Test

- Use 5×2 cv to get 2 folds of 5 tra/val replications (Dietterich, 1998)
- $p_i^{(j)}$: difference btw errors of 1 and 2 on fold j=1, 2 of replication i=1,...,5

$$\bar{p}_i = \left(p_i^{(1)} + p_i^{(2)}\right)/2 \qquad s_i^2 = \left(p_i^{(1)} - \bar{p}_i\right)^2 + \left(p_i^{(2)} - \bar{p}_i\right)^2$$

$$\frac{p_1^{(1)}}{\sqrt{\sum_{i=1}^{5} s_i^2 / 5}} \sim t_5$$

Could use here any other $p_i^{(j)}$

Two-sided test: Accept H_0: $\mu_0 = \mu_1$ if in $(-t_{\alpha/2,5}, t_{\alpha/2,5})$

One-sided test: Accept H_0: $\mu_0 \leq \mu_1$ if $< t_{\alpha,5}$

Compare all values of $p_i^{(j)}$

5×2 cv Paired *F* Test

$$\frac{\sum_{i=1}^{5}\sum_{j=1}^{2}\left(p_i^{(j)}\right)^2}{2\sum_{i=1}^{5}s_i^2}\sim F_{10,5}$$

Because we are taking ratios of two X^2 r.v.s

Two-sided test: Accept H_0: $\mu_0 = \mu_1$ if $< F_{\alpha,10,5}$

Comparing $L>2$ Algorithms: Analysis of Variance (Anova)

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_L$$

- Errors of L algorithms on K folds

$$X_{ij} \sim \mathcal{N}\left(\mu_j, \sigma^2\right), j = 1,\ldots,L,\ i = 1,\ldots,K$$

- We construct two estimators to σ^2 .

  One is valid if H_0 is true, the other is always valid.

  We reject H_0 if the two estimators disagree.

If H_0 is true:

$$m_j = \sum_{i=1}^{K} \frac{X_{ij}}{K} \sim \mathcal{N}\left(\mu, \sigma^2/K\right)$$

$$m = \frac{\sum_{j=1}^{L} m_j}{L} \qquad S^2 = \frac{\sum_j \left(m_j - m\right)^2}{L-1}$$

Thus an estimator of σ^2 is $K \cdot S^2$, namely,

$$\hat{\sigma}^2 = K \sum_{j=1}^{L} \frac{\left(m_j - m\right)^2}{L-1}$$

$$\sum_j \frac{\left(m_j - m\right)^2}{\sigma^2/K} \sim \chi^2_{L-1} \qquad SSb \equiv K \sum_j \left(m_j - m\right)^2$$

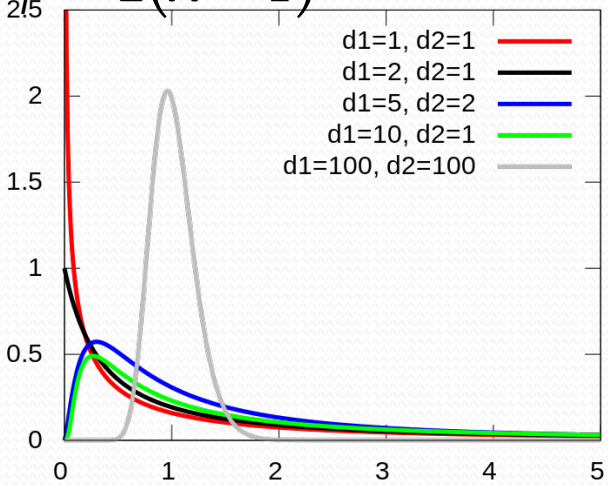
So when H_0 is true, we have

$$\frac{SSb}{\sigma^2} \sim \chi^2_{L-1}$$

Regardless of H_0 our second estimator to σ^2 is the average of group variances S_j^2:

$$S_j^2 = \frac{\sum_{i=1}^{K}\left(X_{ij} - m_j\right)^2}{K-1} \qquad \hat{\sigma}^2 = \sum_{j=1}^{L}\frac{S_j^2}{L} = \sum_j\sum_i \frac{\left(X_{ij} - m_j\right)^2}{L(K-1)}$$

$$SSw \equiv \sum_j\sum_i \left(X_{ij} - m_j\right)^2$$



$$(K-1)\frac{S_j^2}{\sigma^2} \sim \chi^2_{K-1} \qquad \frac{SSw}{\sigma^2} \sim \chi^2_{L(K-1)}$$

$$\left(\frac{SSb/\sigma^2}{L-1}\right) \Big/ \left(\frac{SSw/\sigma^2}{L(K-1)}\right) = \frac{SSb/(L-1)}{SSw/(L(K-1))} \sim F_{L-1,L(K-1)}$$

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_L \text{ if } < F_{\alpha,L-1,L(K-1)}$$

ANOVA table

Source of variation	Sum of squares	Degrees of freedom	Mean square	F_0
Between groups	$SS_b \equiv K\sum_j (m_j - m)^2$	$L-1$	$MS_b = \frac{SS_b}{L-1}$	$\frac{MS_b}{MS_w}$
Within groups	$SS_w \equiv \sum_j \sum_i (X_{ij} - m_j)^2$	$L(K-1)$	$MS_w = \frac{SS_w}{L(K-1)}$	
Total	$SS_T \equiv \sum_j \sum_i (X_{ij} - m)^2$	$L \cdot K - 1$		

If ANOVA rejects, we do pairwise posthoc tests

$$H_0 : \mu_i = \mu_j \text{ vs } H_1 : \mu_i \neq \mu_j$$

$$t = \frac{m_i - m_j}{\sqrt{2}\sigma_w} \sim t_{L(K-1)}$$

Where $mi \sim N(\mu i, \sigma_w^2 = MS_w/K)$

Comparison over Multiple Datasets

- Comparing two algorithms:

  Sign test: Count how many times *A* beats *B* over *N* datasets, and check if this could have been by chance if A and B did have the same error rate

- Comparing multiple algorithms

  Kruskal-Wallis test: Calculate the average rank of all algorithms on N datasets, and check if these could have been by chance if they all had equal error

  If KW rejects, we do pairwise posthoc tests to find which ones have significant rank difference

- See the Tutorial by Padraic Cunningham at
- http://www.ecmlpkdd2009.net/wp-content/uploads/2009/08/evaluation-in-machine-learning.pdf

- Peter Flach Machine Learning slides, Ch 12.:
- https://docs.google.com/file/d/0B8ya4ynGkqSxdXRoVmJFNmJEQTQ/edit