

Lecture Slides for  
**INTRODUCTION  
TO  
MACHINE  
LEARNING**  
3RD EDITION

ETHEM ALPAYDIN

© The MIT Press, 2014

[alpaydin@boun.edu.tr](mailto:alpaydin@boun.edu.tr)

<http://www.cmpe.boun.edu.tr/~ethem/i2ml3e>

CHAPTER 19:

# DESIGN AND ANALYSIS OF MACHINE LEARNING EXPERIMENTS

# Introduction

3

- 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 ?
- Training/validation/test sets
- Resampling methods:  $K$ -fold cross-validation

# Algorithm Preference

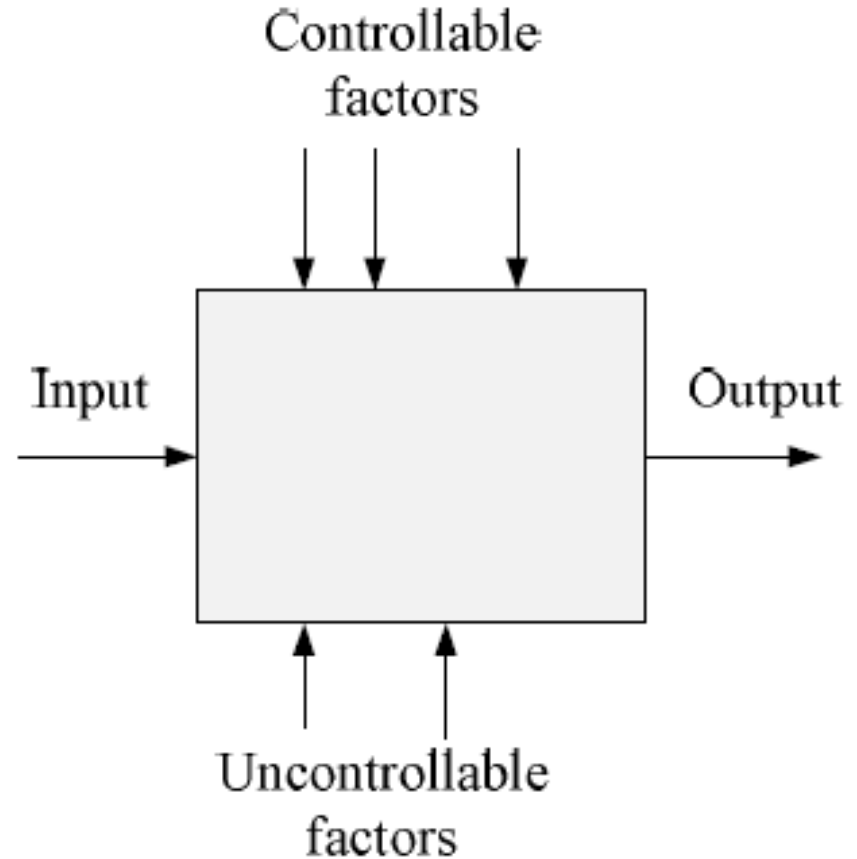
4

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

# Factors and Response

5

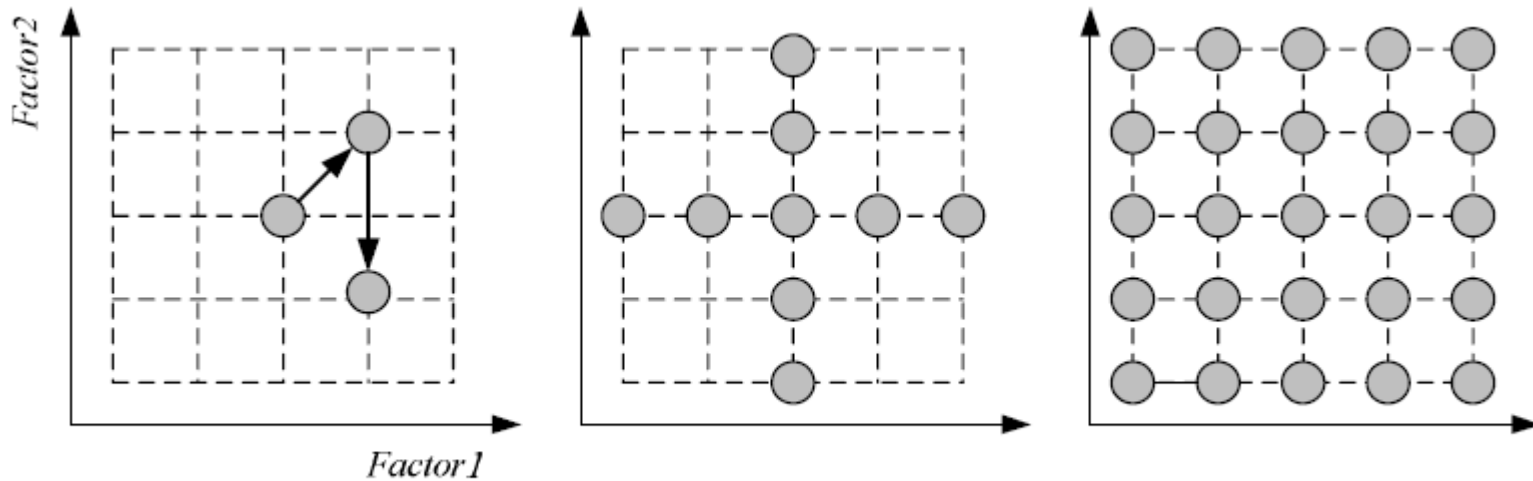
- Response function based on output to be maximized
- Depends on controllable factors
- Uncontrollable factors introduce randomness
- Find the configuration of controllable factors that maximizes response and *minimally affected by uncontrollable factors*



# Strategies of Experimentation

6

How to search the factor space?



(a) Best guess

(b) One factor at a time

(c) Factorial design

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

# Guidelines for ML experiments

7

- 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
- G. Conclusions and Recommendations

# Resampling and K-Fold Cross-Validation

8

- 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  $k$ ,  $X_i, i=1, \dots, K$

$$V_1 = X_1 \quad T_1 = X_2 \cup X_3 \cup \dots \cup X_K$$

$$V_2 = X_2 \quad T_2 = X_1 \cup X_3 \cup \dots \cup X_K$$

$\vdots$

$$V_K = X_K \quad T_K = X_1 \cup X_2 \cup \dots \cup X_{K-1}$$

- $T_i$  share  $K-2$  parts



# 5×2 Cross-Validation

9

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

$$\mathcal{T}_1 = \mathcal{X}_1^{(1)} \quad \mathcal{V}_1 = \mathcal{X}_1^{(2)}$$

$$\mathcal{T}_2 = \mathcal{X}_1^{(2)} \quad \mathcal{V}_2 = \mathcal{X}_1^{(1)}$$

$$\mathcal{T}_3 = \mathcal{X}_2^{(1)} \quad \mathcal{V}_3 = \mathcal{X}_2^{(2)}$$

$$\mathcal{T}_4 = \mathcal{X}_2^{(2)} \quad \mathcal{V}_4 = \mathcal{X}_2^{(1)}$$

⋮

$$\mathcal{T}_9 = \mathcal{X}_5^{(1)} \quad \mathcal{V}_9 = \mathcal{X}_5^{(2)}$$

$$\mathcal{T}_{10} = \mathcal{X}_5^{(2)} \quad \mathcal{V}_{10} = \mathcal{X}_5^{(1)}$$

# Bootstrapping

10

- 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!

# Performance Measures

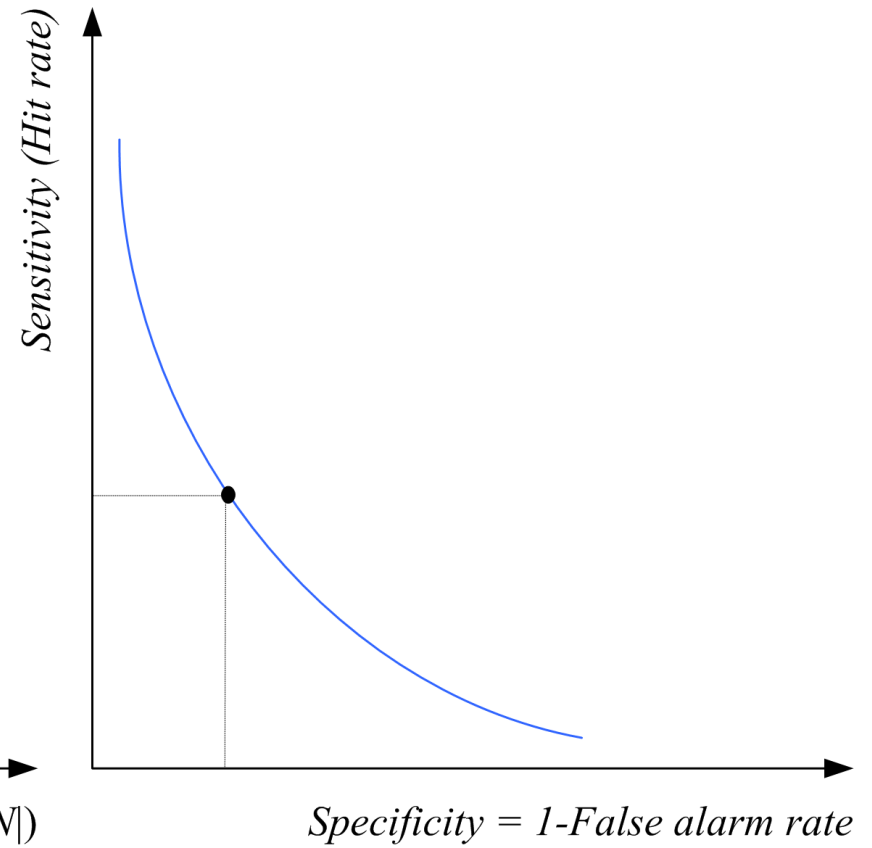
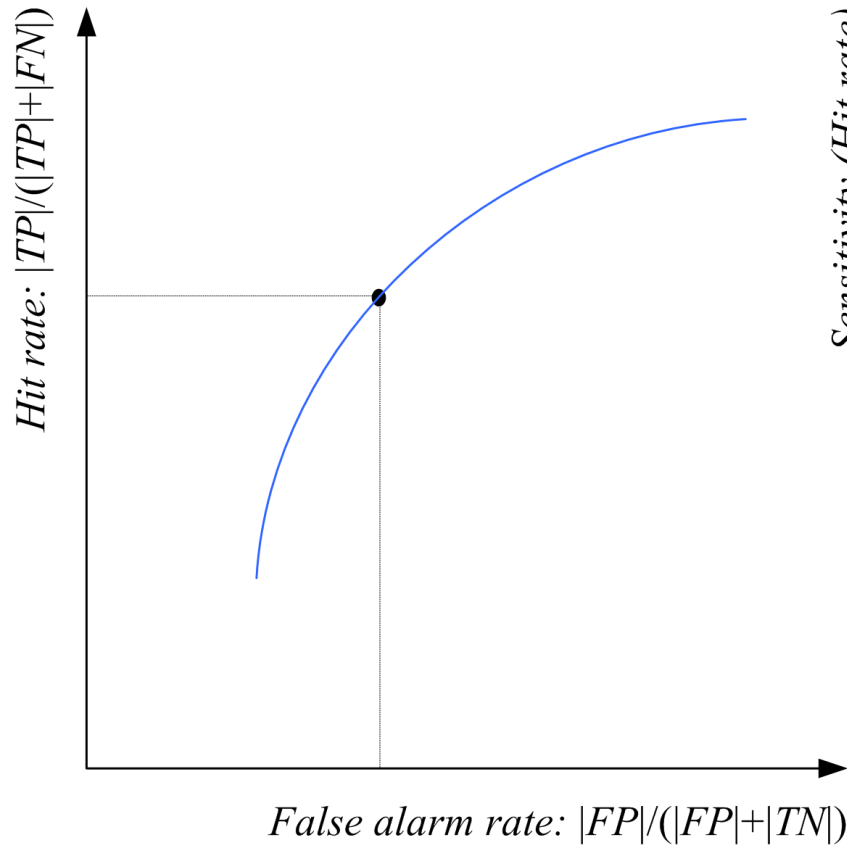
11

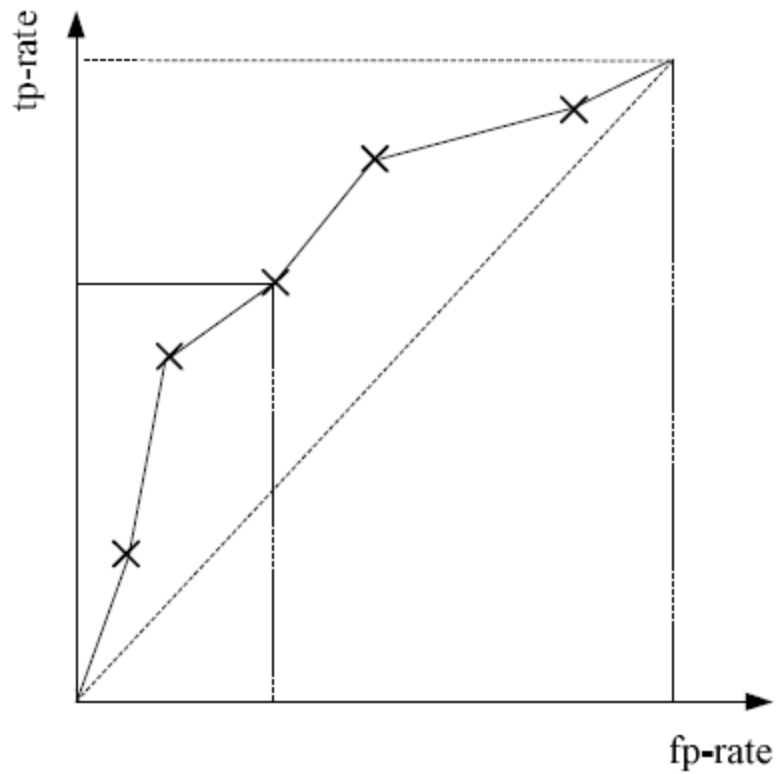
	Predicted class	
True Class	Yes	No
Yes	TP: True Positive	FN: False Negative
No	FP: False Positive	TN: True Negative

- 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

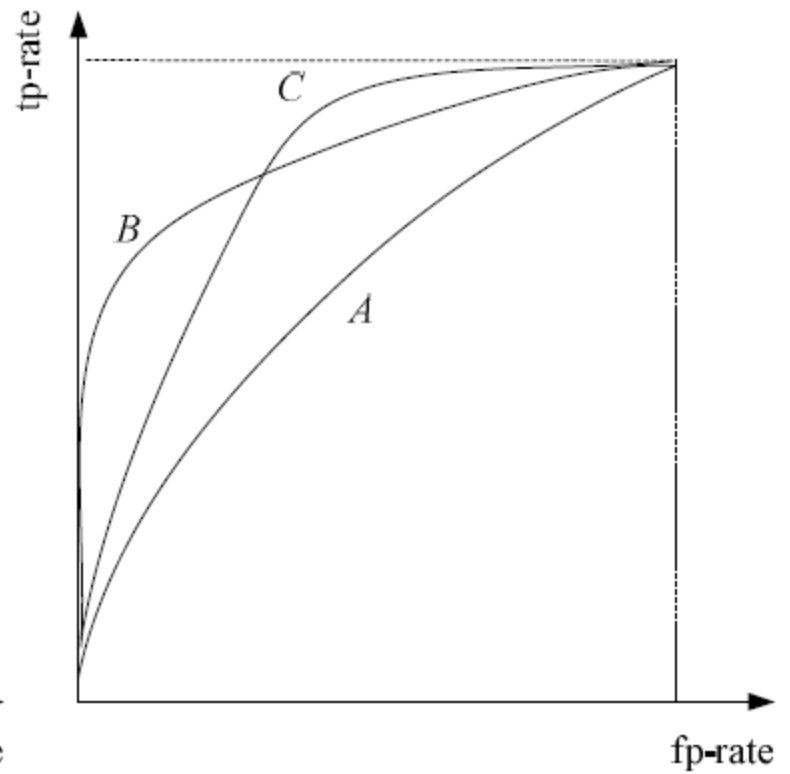
# ROC Curve

12





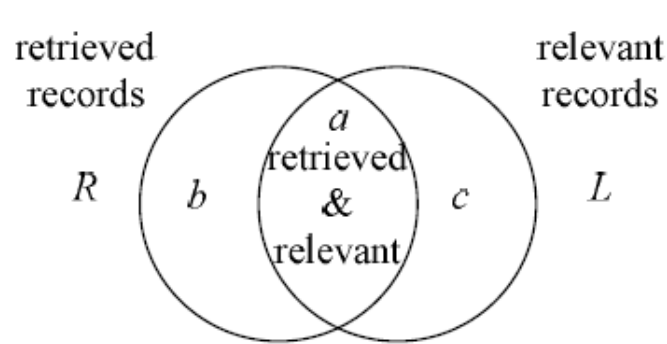
(a) Example ROC curve



(b) Different ROC curves for different classifiers

# Precision and Recall

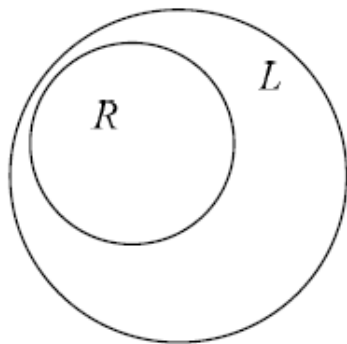
14



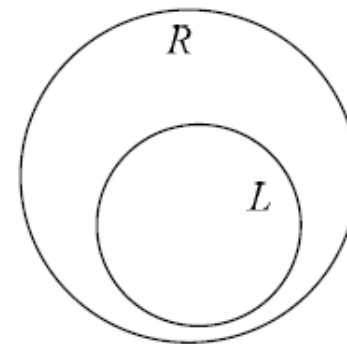
$$\text{Precision: } \frac{a}{a + b}$$

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

(a) Precision and recall



(b) Precision = 1



(c) Recall = 1

# Interval Estimation

15

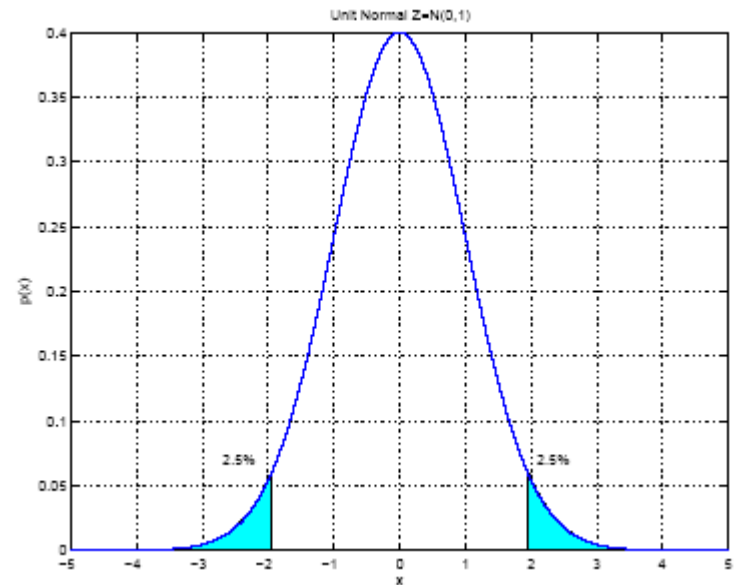
- $X = \{x^t\}_t$  where  $x^t \sim N(\mu, \sigma^2)$
- $m \sim N(\mu, \sigma^2/N)$

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

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

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

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



100(1 -  $\alpha$ ) percent  
confidence interval

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

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

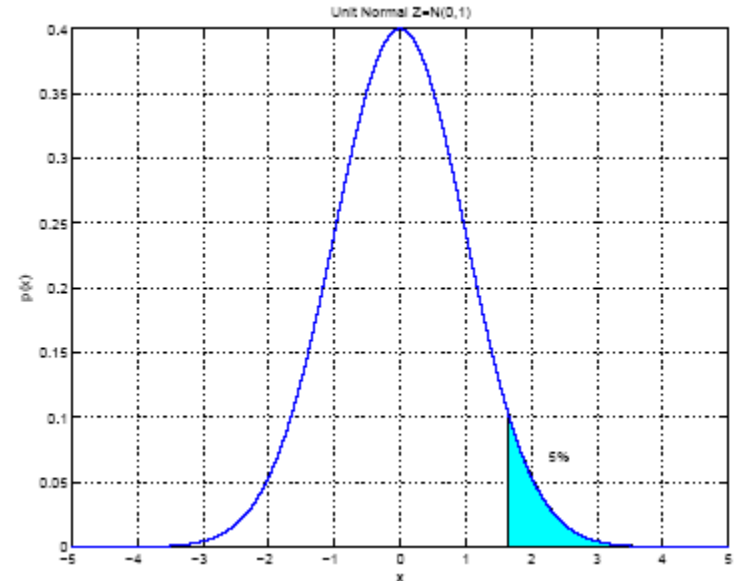
$$P\left\{m - z_{\alpha}\frac{\sigma}{\sqrt{N}} < \mu\right\} = 1 - \alpha$$

When  $\sigma^2$  is not known:

$$s^2 = \sum_t (x^t - m)^2 / (N-1) \quad \frac{\sqrt{N}(m-\mu)}{s} \sim t_{N-1}$$

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

100(1 -  $\alpha$ ) percent one-sided confidence interval





# Hypothesis Testing

17

- Reject a null hypothesis if not supported by the sample with enough confidence

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

$$H_0: \mu = \mu_0 \text{ vs. } H_1: \mu \neq \mu_0$$

Accept  $H_0$  with level of significance  $\alpha$  if  $\mu_0$  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

	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

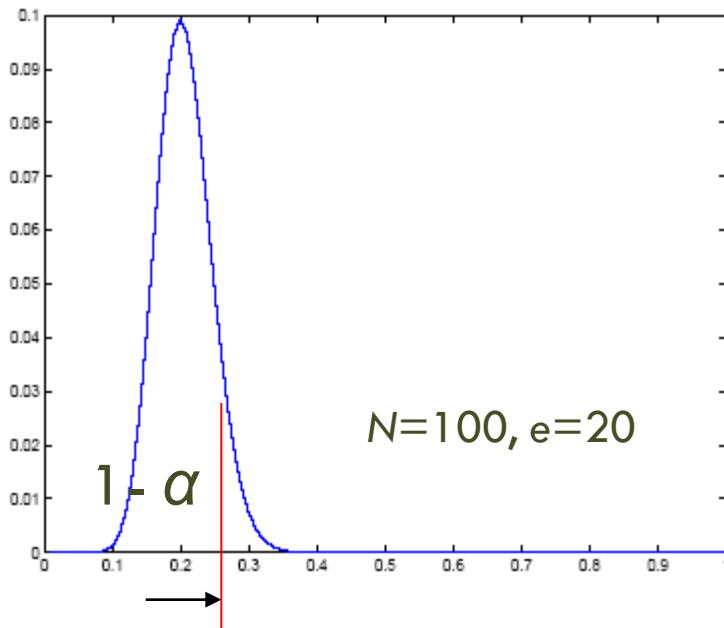
$$\frac{\sqrt{N}(m - \mu_0)}{S} \in (-t_{\alpha/2, N-1}, t_{\alpha/2, N-1})$$

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

19

## □ 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



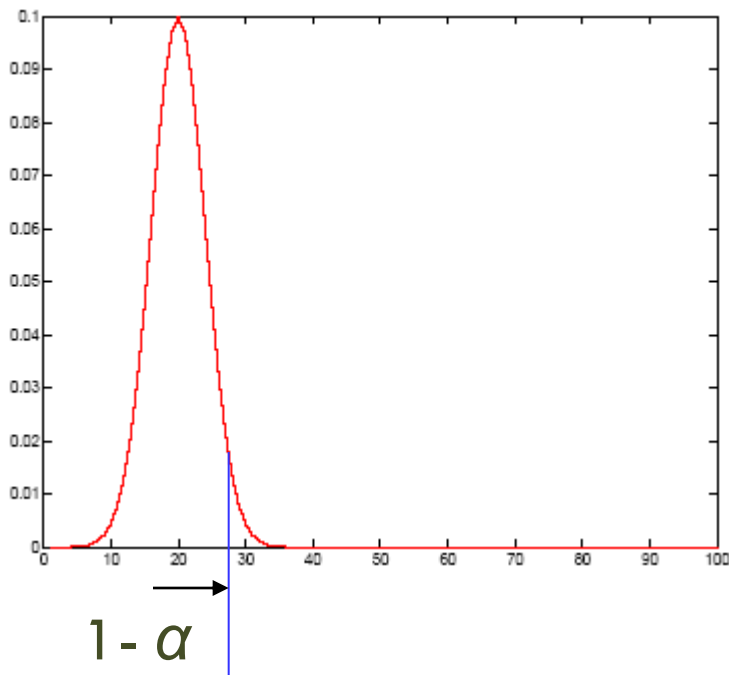
$$P\{X \leq e\} = \sum_{j=1}^e \binom{N}{j} p_0^j (1 - p_0)^{N-j}$$

Accept if this prob is less than  $1 - \alpha$

# Normal Approximation to the Binomial

20

- 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  $z_{1-\alpha}$

# Paired $t$ Test

21

- Multiple training/validation sets
- $x_i^t = 1$  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  $t_{\alpha, K-1}$

# Comparing Classifiers: $H_0: \mu_0 = \mu_1$ vs. $H_1: \mu_0 \neq \mu_1$

22

- 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  $e_{01} = e_{10} = (e_{01} + e_{10})/2$

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

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

# K-Fold CV Paired $t$ Test

23

- 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

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

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

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

# 5×2 cv Paired $t$ Test

24

- 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  $i=1, \dots, 5$

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

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

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  $< t_{\alpha,5}$



# 5×2 cv Paired $F$ Test

25

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

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

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

26

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

- Errors of  $L$  algorithms on  $K$  folds

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

- We construct two estimators to  $\sigma^2$  .

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 :

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

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

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

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

$$\sum_j \frac{(m_j - m)^2}{\sigma^2 / K} \sim \chi_{L-1}^2 \quad SSb \equiv K \sum_j (m_j - m)^2$$

So when  $H_0$  is true, we have

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

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

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

$$SSW \equiv \sum_j \sum_i (x_{ij} - m_j)^2$$

$$(K-1) \frac{S_j^2}{\sigma^2} \sim \chi_{K-1}^2 \quad \frac{SSW}{\sigma^2} \sim \chi_{L(K-1)}^2$$

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

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

# ANOVA table

29

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

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

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

# Comparison over Multiple Datasets

30

- 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

# Multivariate Tests

31

- Instead of testing using a single performance measure, e.g., error, use multiple measures for better discrimination, e.g., [fp-rate,fn-rate]
- Compare  $p$ -dimensional distributions
- Parametric case: Assume  $p$ -variate Gaussians

$$H_0 : \mu_1 = \mu_2 \text{ vs. } H_1 : \mu_1 \neq \mu_2$$

# Multivariate Pairwise Comparison

32

- Paired differences:  $d_i = x_{1i} - x_{2i}$

$$H_0 : \boldsymbol{\mu}_d = \mathbf{0} \text{ vs. } H_1 : \boldsymbol{\mu}_d \neq \mathbf{0}$$

- Hotelling's multivariate  $T^2$  test

$$T'^2 = K \mathbf{m}^T \mathbf{S}^{-1} \mathbf{m}$$

- For  $p=1$ , reduces to paired  $t$  test



# Multivariate ANOVA

33

## □ Comparison of $L > 2$ algorithms

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_L \text{ vs.}$$

$$H_1 : \boldsymbol{\mu}_r \neq \boldsymbol{\mu}_s \text{ for at least one pair } r, s$$

$$\mathbf{H} = K \sum_{j=1}^L (\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^T$$

$$\mathbf{E} = \sum_{j=1}^L \sum_{i=1}^K (\mathbf{x}_{ij} - \mathbf{m}_j)(\mathbf{x}_{ij} - \mathbf{m}_j)^T$$

$$\Lambda' = \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}|}$$

is Wilks's  $\Lambda$  distributed with  $p, L(K-1), L-1$  degrees of freedom