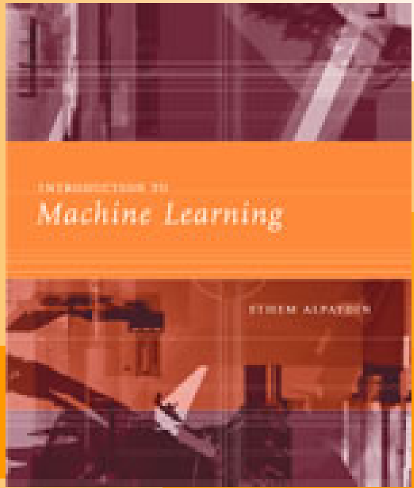


Lecture Slides for

INTRODUCTION TO

Machine Learning

ETHEM ALPAYDIN

alpaydin@boun.edu.tr
http://www.cmpe.boun.edu.tr/~ethem/i2ml

CHAPTER 14:

Assessing and Comparing Classification Algorithms

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

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

Resampling and K-Fold Cross-Validation

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

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

- $\mathcal{T}_i$ share K-2 parts

5×2 Cross-Validation

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

$$
\begin{array}{ll}
\mathcal{T}_1 = \mathcal{X}_1^{(1)} & \mathcal{V}_1 = \mathcal{X}_1^{(2)} \\
\mathcal{T}_2 = \mathcal{X}_1^{(2)} & \mathcal{V}_2 = \mathcal{X}_1^{(1)} \\
\mathcal{T}_3 = \mathcal{X}_2^{(1)} & \mathcal{V}_3 = \mathcal{X}_2^{(2)} \\
\mathcal{T}_4 = \mathcal{X}_2^{(2)} & \mathcal{V}_4 = \mathcal{X}_2^{(1)} \\
\vdots & \\
\mathcal{T}_9 = \mathcal{X}_5^{(1)} & \mathcal{V}_9 = \mathcal{X}_5^{(2)} \\
\mathcal{T}_{10} = \mathcal{X}_5^{(2)} & \mathcal{V}_{10} = \mathcal{X}_5^{(1)}
\end{array}
$$

Bootstrapping

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

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

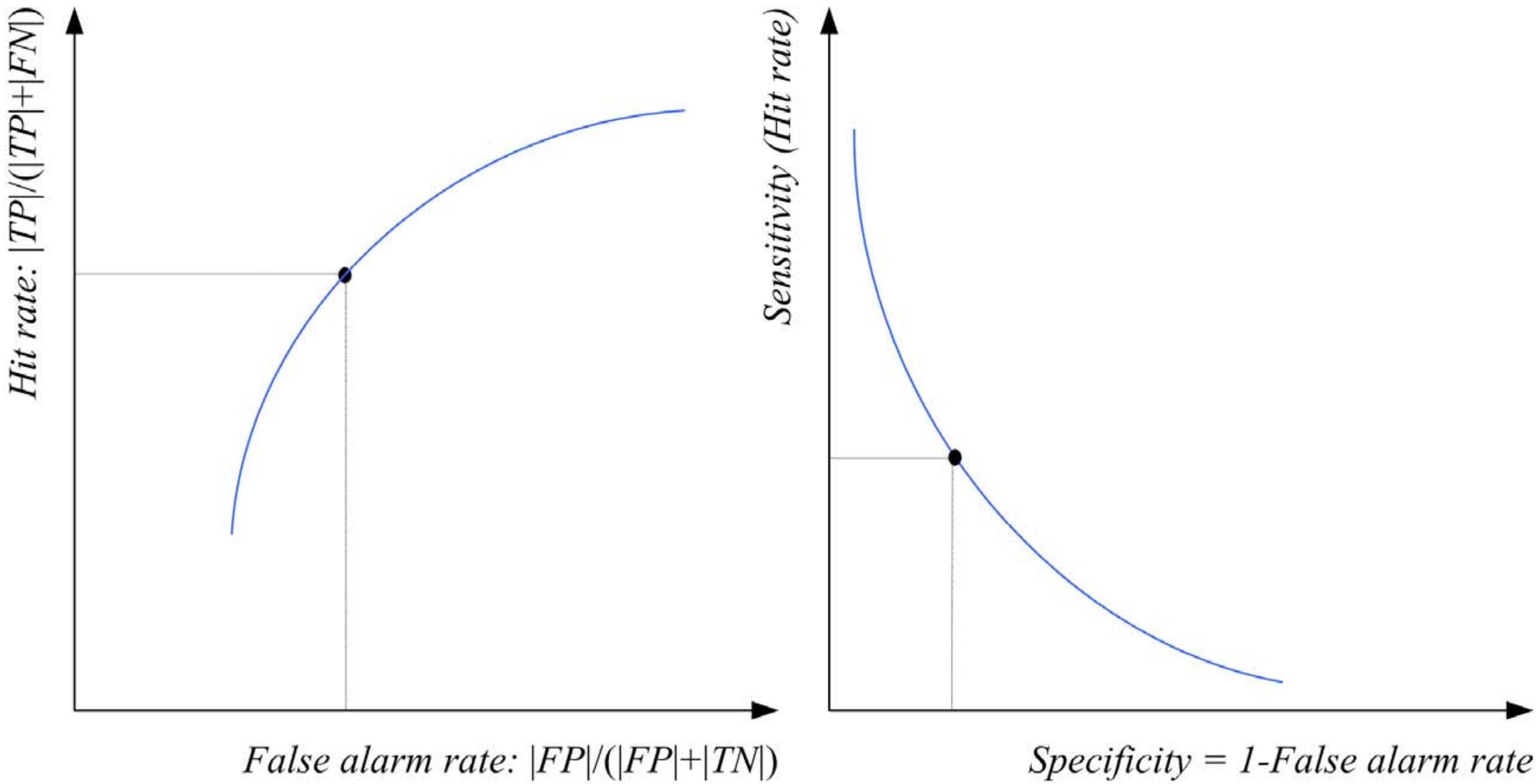
that is, only 36.8% is new!

Measuring Error

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

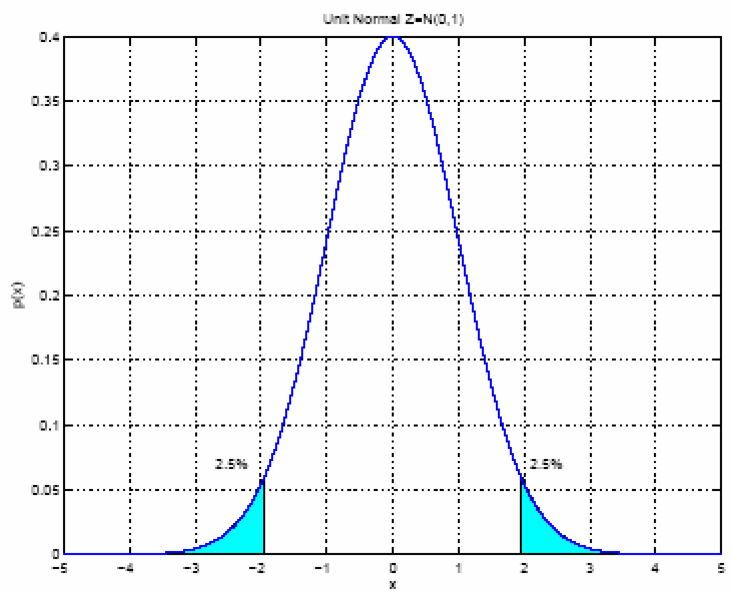
- Error rate = # of errors / # of instances = (FN+FP) / N
- Recall = # of found positives / # of positives
  = TP / (TP+FN) = sensitivity = hit rate
- Precision = # of found positives / # of found
  = TP / (TP+FP)
- Specificity = TN / (TN+FP)
- False alarm rate = FP / (FP+TN) = 1 - Specificity

ROC Curve


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

Interval Estimation

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



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

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

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

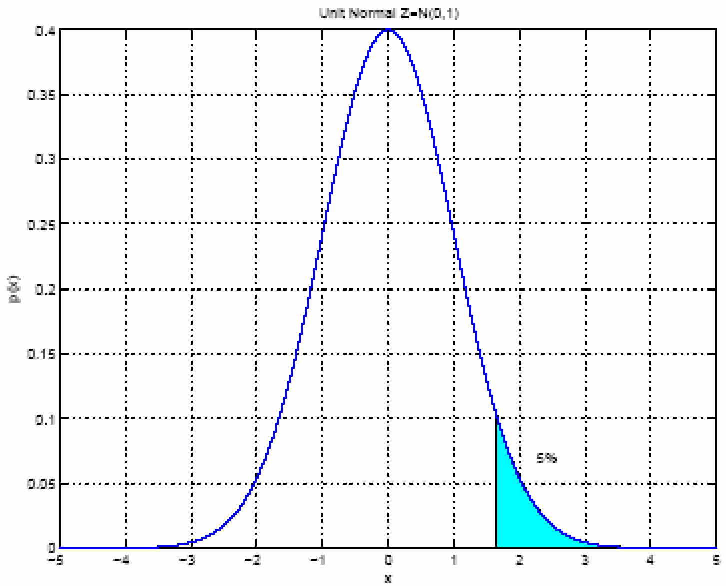
$$P\left\{m - z_{\alpha/2}\frac{\sigma}{\sqrt{N}} < \mu < m + z_{\alpha/2}\frac{\sigma}{\sqrt{N}}\right\} = 1-\alpha$$

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

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

or

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



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

When σ^2 is not known:

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

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

Hypothesis Testing

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

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

  Accept H_0 with level of significance α if μ_0 is in the $100(1-\alpha)$ 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 $\dfrac{\sqrt{N}}{\sigma}(m - \mu_0) \in (-\infty, z_\alpha)$

- Variance unknown: Use t, instead of z.

  Accept H_0: $\mu = \mu_0$ if

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

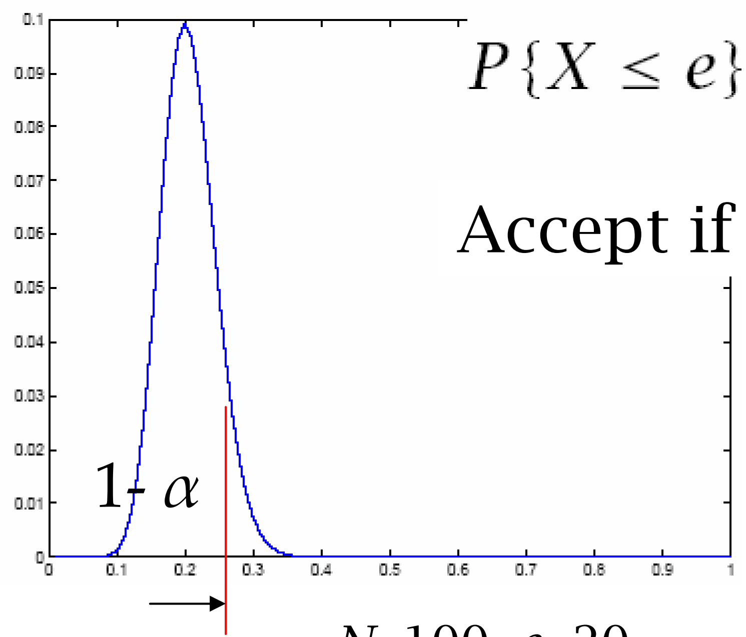
Assessing Error: $H_0: p \leq p_0$ vs. $H_1: p > p_0$

- 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- α

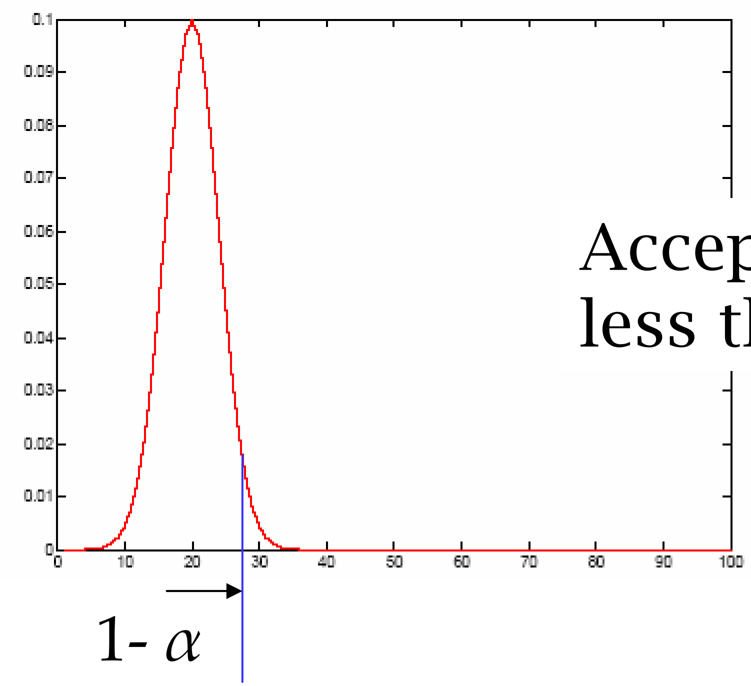


N=100, e=20

Normal Approximation to the Binomial

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

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



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

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

$$\frac{(|e_{01} - e_{10}| - 1)^2}{e_{01} + e_{10}} \sim \mathcal{X}_1^2$$

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

K-Fold CV Paired t Test

- 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

$$H_0 \quad : \quad \mu = 0 \qquad H_1 \quad : \quad \mu \neq 0$$

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

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

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

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

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

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

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

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

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

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

- Errors of L algorithms on K folds

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

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

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

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

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

$$\sum_j \frac{(m_j - m)^2}{\sigma^2/K} \sim \mathcal{X}^2_{L-1} \qquad SS_b \equiv K \sum_j (m_j - m)^2$$

So, when H_0 is true, we have

$$\frac{SS_b}{\sigma^2} \sim \mathcal{X}^2_{L-1}$$

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

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

$$SS_w \equiv \sum_j \sum_i (X_{ij}-m_j)^2$$

$$(K-1)\frac{S_j^2}{\sigma^2} \sim \mathcal{X}_{K-1}^2 \qquad \frac{SS_w}{\sigma^2} \sim \mathcal{X}_{L(K-1)}^2$$

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

H_0: $\mu_1 = \mu_2 = \mu_3 \ldots = \mu_L$ if $< F_{\alpha,L-1,L(K-1)}$

Other Tests

- Range test (Newman-Keuls): $\underline{1\,4\,5}\ \underline{2\,3}$
- Nonparametric tests (Sign test, Kruskal-Wallis)
- Contrasts: Check if 1 and 2 differ from 3,4, and 5
- Multiple comparisons require Bonferroni correction
  If there are m tests, to have an overall significance of α, each test should have a significance of α/m.
- Regression: CLT states that the sum of iid variables from *any* distribution is approximately normal and the preceding methods can be used.
- Other loss functions ?