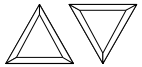
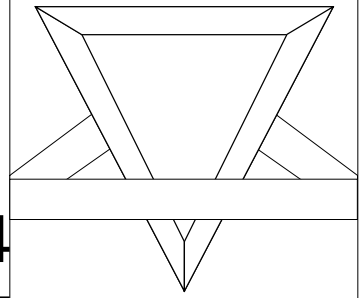
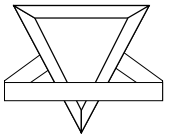
Chapter 4
Statistical Learning Theory



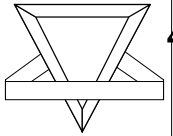


Outline



Introduction

- VC (Vapnik-Chrvonenkis) theory: SLT
  - Conditions for consistency of the ERM inductive principle
  - Bounds on the generalization ability of learning machines based on these conditions
  - Principles for inductive inference from small samples based on these bounds
  - Constructive methods for implementing above inductive principles



4.1 Conditions for Consistency and Convergence of ERM

$$R(\omega) = \int Q(z,\omega) dF(z) \text{ or } R(\omega) = \int Q(\mathbf{z},\omega)\, p(z) dz$$

$$Q(z,\omega) = (y - f(x,\omega))^2$$

$$R_{emp}(\omega) = \sum_{i=1}^{n} Q(z_i,\omega)$$

solution approaches to the learning problem

- estimate unknown c.d.f from data -> find optimal estimate $f(x,\omega_0)$
- seek an estimate providing minimum of the (known) empirical risk (ERM)

Conditions for Consistency and Convergence of ERM

- Consistency
  - Requirement that estimates provided by ERM should converge to the true values as the # of samples grows

$$R(\omega^* \mid n) \to R(\omega_0) \quad when \quad n \to \infty$$

$$\mathrm{Re}_{mp}(\omega^* \mid n) \to R(\omega_0) \quad when \quad n \to \infty$$

  - We can expect

$$\mathrm{Re}_{mp}(\omega^* \mid n) < R(\omega^* \mid n)$$

Conditions for Consistency and Convergence of ERM

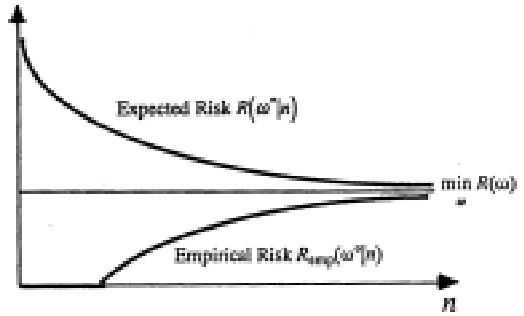


Figure 4.1 Consistency of the ERM.

Conditions for Consistency and Convergence of ERM

- Nontrivial consistency (Vapnik)
  - Consistency should hold for all approximating functions
  - Key Theorem of Learning Theory (Vapnik and Chervonenkis, 1989)

$$\lim_{n\to\infty} P\left[\sup_{\omega} \mid R(\omega) - \mathrm{Re}_{mp}(\omega) \mid > \varepsilon\right] = 0, \quad \forall \varepsilon > 0$$

Conditions for Consistency and Convergence of ERM

- Diversity of a set of functions w.r.t. Zn (dichotomy case)
  - $N(\mathbf{Z}n)$: # of different dichotomies by $Q(z, \omega)$
- $H(\mathbf{Z}n) = \ln N(\mathbf{Z}n)$: random entropy
- $H(n) = E(\ln N(\mathbf{Z}n))$: VC entropy of the set of indicator functions on a sample of size n from $F(z)$
  - provides a measure of the expected diversity of a set of indicator functions with respect to a sample of a given size
  - depends on the set of indicator funcs and on the (unknown) distribution of samples $F(z)$
- $G(n) = \ln \max_{Z_n} N(Z_n)$
  - growth function: distribution-independent
  - provides an upper bound for the (distribution-dependent) entropy

Conditions for Consistency and Convergence of ERM

$G(n) \le n \ln 2$

- $H_{\text{ann}}(n) = \ln E(N(\mathbf{Z}_n))$: annealed VC entropy
  Using Jensen's inequality

$$\sum_i a_i \ln \chi_i \le \ln\left(\sum_i a_i \chi_i\right)$$

It can be shown that

$H(n) \le H_{ann}(n)$

- $H(n) \le H_{ann}(n) \le G(n) \le n \ln 2$
- Necessary and sufficient condition for consistency of the ERM principle (Vapnik and Chervonenkis, 1968)

$\lim_{n\to\infty} \frac{H(n)}{n} = 0$

Conditions for Consistency and Convergence of ERM

- asymptotic rate of convergence is called *fast* if for any $n>n_0$ the following holds

$$P(R(\varpi) - R(\omega^*) < \varepsilon) = e^{-cn\varepsilon^2}$$

($c>0$ is a positive constant)

Conditions for Consistency and Convergence of ERM

- Sufficient condition for fast rate of convergence

$\lim_{n\to\infty} \frac{H_{ann}(n)}{n} = 0$ (distribution-dependent condition)

- Distribution-independent condition for consistency of ERM and fast convergence

$$\lim_{n\to\infty} \frac{G(n)}{n} = 0$$

4.2 Growth Function and VC-Dimension

▼ VC-dimension

- Growth function is either linear or bounded by a logarithmic function of # of samples *n.*
- The point $n = h$ where the growth starts to slow down is called the VC-dimension.
- If it is finite, the growth function is bounded by

$$G(n) \le h(1 + \ln \frac{n}{h})$$

- Finiteness of *h* provides necessary and sufficient conditions for the fast rate convergence and for distribution-independent consistency of ERM.

VC-Dimension

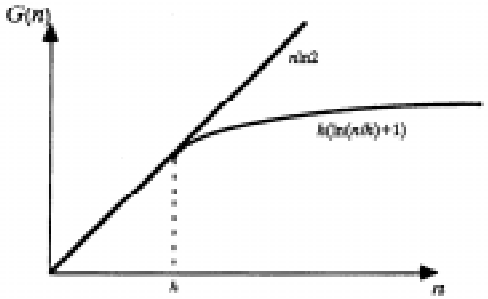


Figure 4.2 Behavior of the Growth Function.

VC-Dimension

- Criterion for demarcation between true and false (inductive) theories (Popper)
  - The necessary condition for the inductive theory to be true is the feasibility of its falsification, i.e., the existence of certain assertions (facts) that cannot be explained by the theory.
  - e.g.) VC-dimension is infinite -> false model
  - *shattering*: *n* samples can be separated by a set of indicator func.s in all 2^n.

VC-Dimension

- VC-dimension of a set of indicator functions
  - VC-dimension *h* <-> if there exist *h* samples that can be shattered by this set of functions but not *h+1*
- VC-dimension of the set of real-valued func.
  - Indicator function

    $A \le Q(\mathbf{z}, \omega) \le B$

    $I(\mathbf{z}, \omega, \beta) = I[Q(\mathbf{z}, \omega) - \beta) > 0]$
  - VC-dimension of real function Q is equal to the VC-dimension of indicator function.

VC-Dimension

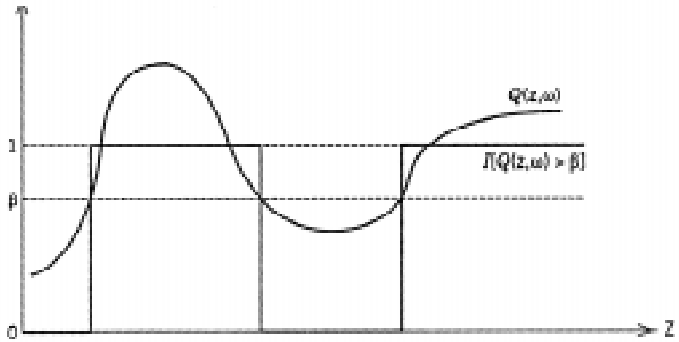


Figure 4.3 VC-dimension of the set of real-valued functions.

VC-Dimension

- VC-dimension for Classification and regression problems

Classification

$$Q(\mathbf{z},\omega) = \begin{cases} f(\mathbf{x},\omega) & \text{if } y = 0 \\ 1 - f(\mathbf{x},\omega) & \text{if } y = 1 \end{cases}$$

Regression

$$Q(\mathbf{z},\omega) = (y - f(\mathbf{x},w))^2$$

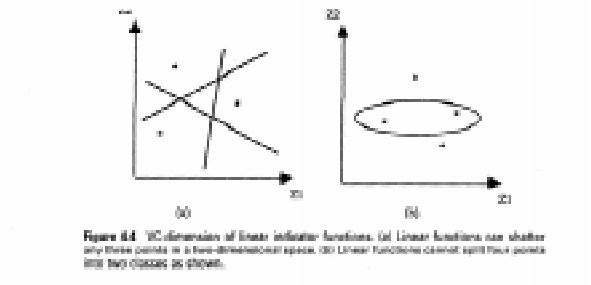
$$h_f \le h \le ch_f$$

$$h \approx h_f$$

VC-Dimension

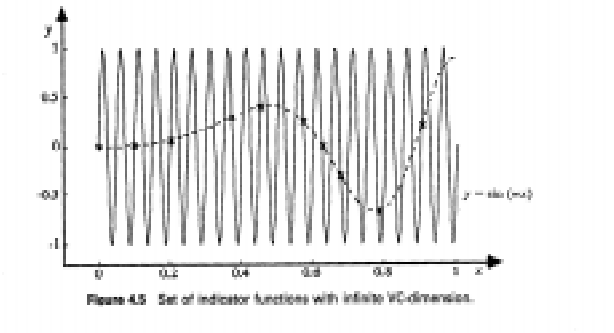
- Examples of calculating VC-dimension
  - VC-Dimension of a set of linear indicator func
    - $h = d + 1$

$$Q(\mathbf{z},\mathbf{w}) = I(\sum_{i=1}^{d} w_i z_i + w_0 > 0)$$



VC-Dimension

- Set of univariate functions with a single parameter
  - $f(x,\omega) = I(\sin wx > 0)$
  - $h = \infty$

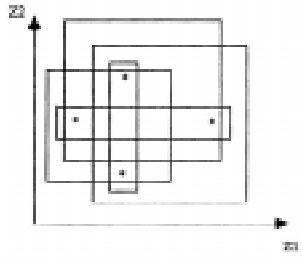


VC-Dimension

- Set of rectangular indicator functions

$$Q(\mathbf{z},\mathbf{c},\mathbf{w}) = 1 \text{ if and only if } |z_i - c_i| \le w_i, \; i = 1,2,...,d$$

  - $h = 2d$



VC-Dimension

- Set of radically symmetric indicator functions

$Q(\mathbf{z},\mathbf{c},r)=1$ if and only if $\|\mathbf{z}-\mathbf{c}\|\le r$

- $h = d+1$

- Set of real-valued "local" functions

$$I(\mathbf{x},\mathbf{c},\alpha)=K\left(\frac{\|\mathbf{x}-\mathbf{c}\|}{\alpha}\right)$$

$$I(\mathbf{x},\mathbf{c},\alpha,\beta)=I\left[K\left(\frac{\|\mathbf{x}-\mathbf{c}\|}{\alpha}\right)-\beta\right]$$ (d+2 free parameter)

- $h = d+1$

VC-Dimension

- Linear combination of fixed basis functions

$$Q_m(\mathbf{z},\mathbf{w})=\sum_{i=1}^{m} w_i\mathbf{g}_i(\mathbf{z})+w_0$$

  - $h = m+1$

- Linear combination of adaptive basis functions nonlinear in parameters

$$Q_m(\mathbf{z},\mathbf{w},\mathbf{v})=\sum_{i=1}^{m} w_i\mathbf{g}_i(\mathbf{z},\mathbf{v})+w_0$$

  - can be infinite

4.3. Bounds on the Generalization

- The upper bounds on the rate of uniform convergence of the learning processes
  - Evaluate the difference between true risk and the known empirical risk.
  - Constructive distribution-independent bounds form the foundation for a new inductive principle (structural risk minimization) and associated constructive procedures.

Bounds on the Generalization

- Classification
  - With probability 1-η for all Q

$$R(\omega)\le \mathrm{R}_{emp}(\omega)+\frac{\varepsilon}{2}\left(1+\sqrt{1+\frac{4\,\mathrm{R}_{emp}(\omega)}{\varepsilon}}\right)$$

where, $\varepsilon=\varepsilon\left(\frac{\mathrm{n}}{\mathrm{h}},\frac{-\ln\eta}{n}\right)=a_1\frac{h[\ln(a_2 n/h)+1]-\ln(\eta/4)}{n}$

  - If the set of loss functions contains a finite number of elements N

$$\varepsilon=2\frac{\ln N-\ln\eta}{n}$$

Bounds on the Generalization

- With 1-2η probability for the func. that minimizes empirical risk

$$R(\omega^{*}\mid n)-\min_{\omega}\ R(\omega)\le\sqrt{\frac{-\ln\eta}{2n}}+\frac{\varepsilon}{2}\left(1+\sqrt{1+\frac{4}{\varepsilon}}\right)$$

- Confidence level : 1- η
- There is a trade-off between the accuracy provided by the bounds and the degree of confidence

Bounds on the Generalization

▼ Regression

- For 1- η probability, *c* depending on the "tail of the distribution" of the loss functions

$$R(\omega)\le\frac{\text{Re}\ \ mp\ (\omega)}{(1-c\sqrt{\varepsilon})_{+}}$$

- 1 - 2η for that function that minimizes empirical risk

$$\frac{R(\omega^{*}\mid n)-\min_{\omega}R(\omega)}{\min_{\omega}R(\omega)}\le\frac{c\sqrt{\varepsilon}}{1-c\sqrt{\varepsilon}}+O(1/n)$$

Bounds on the Generalization

- If *n* and η are held at particular values, it is possible to determine the value of *h* that leads to the bound approaching

$$\varepsilon(h)=a_1\frac{h[\ln(a_2n/h)+1-\ln(\eta/4)}{n}\ge 1\ \text{ with }\ a_1=1,a_2=1$$

$$\frac{h}{n}\le 0.8\quad\text{for}\ \eta\ge\min\left(\frac{4}{\sqrt{n}},1\right)$$

Bounds on the Generalization

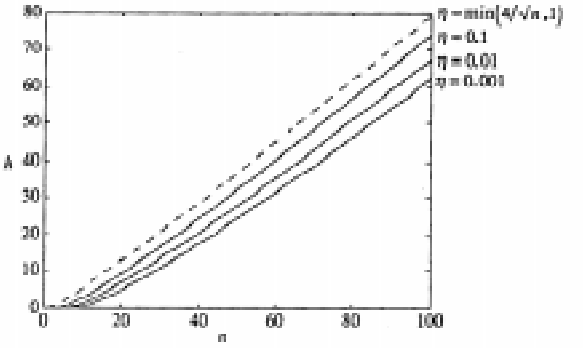


Figure 4.7 Values of *n*, η, and *h* that cause the generalization bound to approach infinity under real-life conditions ($a_1 = 1$, $a_2 = 1$).

4.4 Structural Risk Minimization

- If $\frac{h}{n}$ is small, other factors must be minimized.
- First term in (4.22) depends on a particular func. from the set of functions.
- Second term depends mainly on the VC-dimension of the set of functions.
- Structural risk minimization (SRM) provides a formal mechanism for choosing an optimal model complexity for a finite sample.

SRM

- Under SRM the set S of loss functions Q(**z**,**w**), **w**$\in \Omega$ has a *structure*, that is, it consists of the nested subsets $S_k = \{Q(\mathbf{z},\mathbf{w}), w \in \Omega_k\}$ such that $S_1 \subset S_2 \subset ... \subset S_k \subset ...$ where $h_1 < h_2 ... < h_k ...$
- Solving a learning problem with finite data
  - requires a priori specification of a structure on a set of approximating functions then
  - 1. selecting an element of a structure(having optimal complexity)
  - 2. esitmating the model from the element

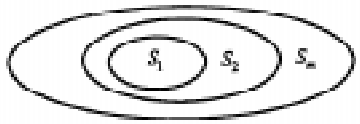


Figure 4.8 Structure on a set of functions.

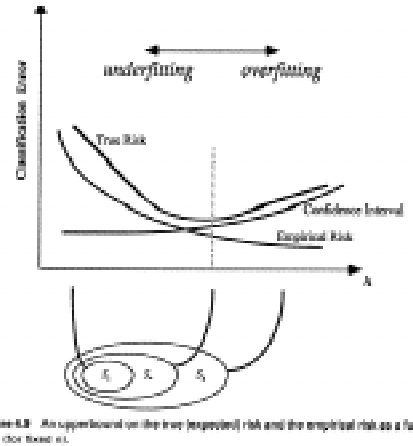


Figure 4.9 An upper bound on the true (expected) risk and the empirical risk as a function of h (for fixed n).

SRM

- 1. Dictionary representation

$$f_m(\mathbf{x},\mathbf{w},\mathbf{V}) = \sum_{i=0}^{m} w_i \mathbf{g}(\mathbf{x},\mathbf{v}_i)$$

$$f_1 \subset f_2 \subset ... \subset f_k ...$$

for example,

$$f_m(x,\mathbf{w}) = \sum_{i=0}^{m} w_i x^i$$

SRM

- 2. Penalization formulation

  $S_k = \{f(\mathrm{x}, \mathrm{w}), \|\mathrm{w}\|^2 \le c_k\}$ where $c_1 < c_2 < c_3 \ldots$

  $R_{pen}(\omega, \lambda_k) = R_{emp}(\omega) + \lambda_k \|\mathrm{w}\|^2$

  with an appropriately chosen Lagrange multiplier λ_k

  such that $\lambda_1 > \lambda_2 > \lambda_3 \ldots$

- 3. Input preprocessing

  $z = K(x, \beta)$

  $S_k = \{f(K(\mathbf{x}, \beta), \mathbf{w}), \beta \le c_k\}$ where $c_1 > c_2 > c_3 \ldots$

SRM

- 4. Initial conditions for training algorithm

  $S_k = \{A : f(\mathbf{x}, \mathbf{w}), \|\mathbf{w}^0\| \le c_k\}$

4.5 Case Study : Comparison of Methods for Model Selection

- Objective : to choose the model complexity optimally for a given training sample
- Practical application of either SRM or penalization requires two tasks :
  - Estimation of model parameters (via minimization of the penalized empirical risk)
  - Estimation of the prediction risk
- Two major approaches for estimating prediction risk :
  - Analytic methods
  - Resampling or data-driven methods

Case Study

- Training samples

  $y = \sin^2(2\pi x) + \varepsilon, \; x \in [0,1]$

  - sample size : 10, 20, 30, 100
  - noise : defined in terms of SNR as the ratio of the standard deviation of the true output values for given input samples over the standard deviation of the gaussian noise
- Approximating functions
  - class of polynomials of degree *m*
  - set of functions are linear in parameters : solving linear least squares

Case Study

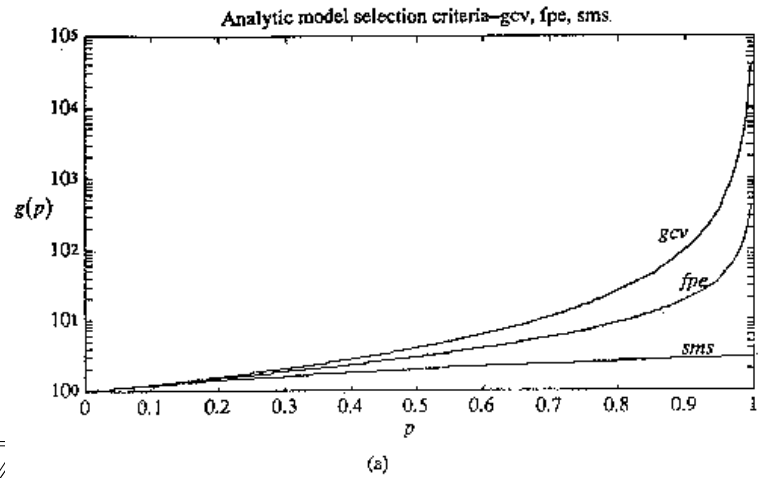
- Model selection
  - Choosing an optimal polynomial degree *m* for a given training sample
  - Comparison set
    - Final prediction error (fpe)
    - Schwartz criteria (sc)
    - Generalized cross-validation (gcv)
    - Shibata's model selector (sms)
    - Leave-one-out cross-validation (cv)
    - Vapnik's measure (vm)

$$g(p,n)=\left(1-\sqrt{p-p\ln p+\frac{\ln n}{2n}}\right)_{+}^{-1}$$

Case Study

- Comparison strategy
  - 1000 times repetition for given small-size training set
  - Standard box plot notation describing empirical distribution
- Experimental results
  - Vapnik's measure for model selection shows superior overall performance
- Summary of comparison results
  - Small size training data may cause no guaranteed performance, so that measures like Vapnik's are required.
  - Vapnik's measure guarantees the best worst-case estimates.

Case Study



Case Study

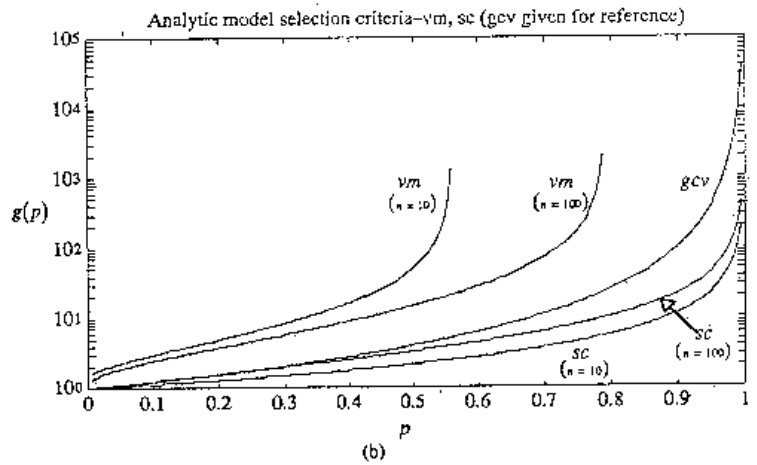
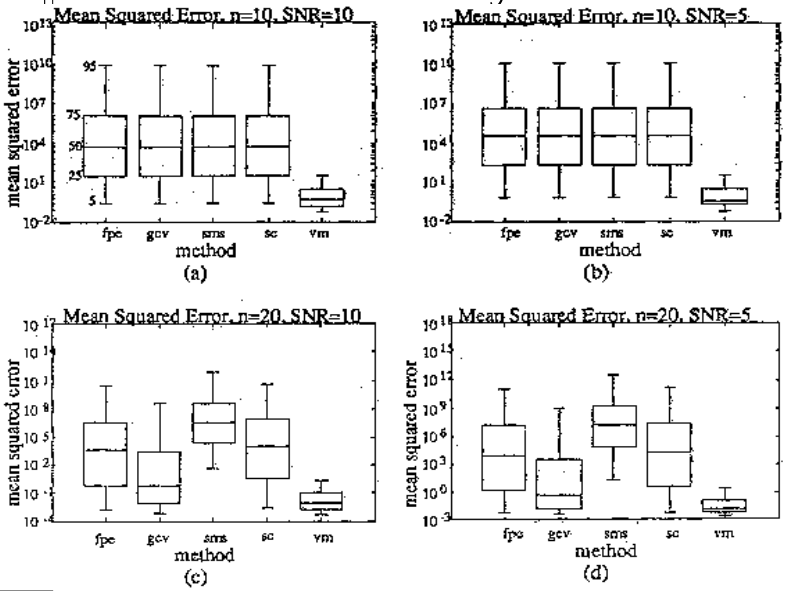


Figure 4.10 Various analytical model selection penalization functions: (*a*) Generalized cross-validation (gcv), final prediction error (fpe), and Shibata's model selector (sms). (*b*) Vapnik's measure (vm) and Schwartz criteria (sc) for sample sizes indicated. The parameter p is equal to h/n.

Case Study



Case Study

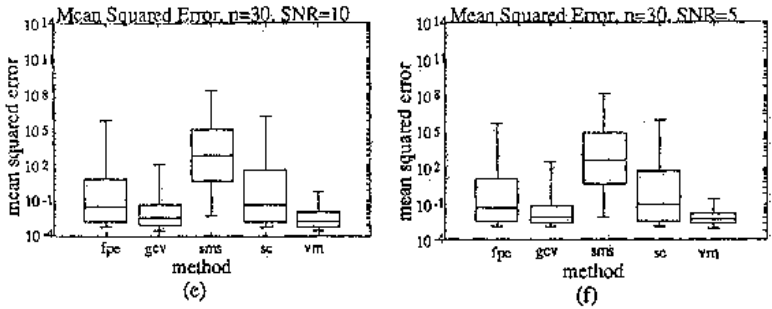
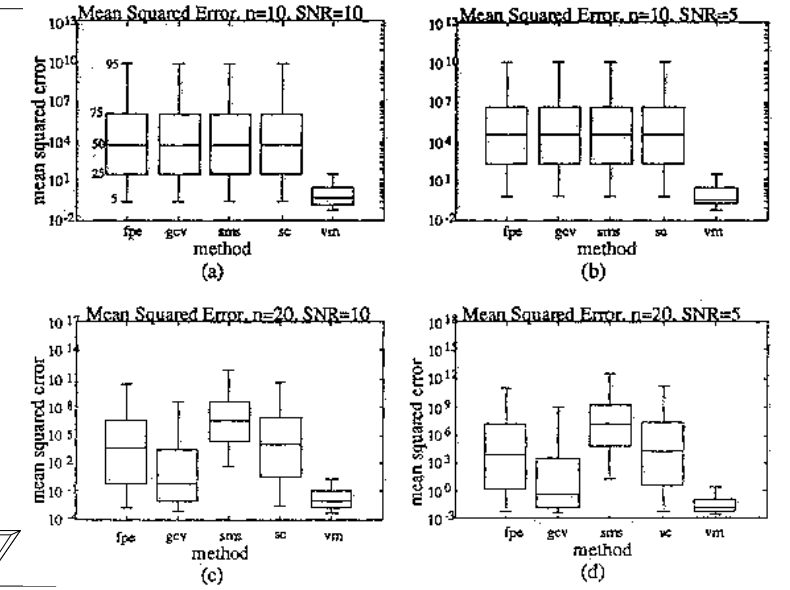


Figure 4.11 Mean squared error for each model selection approach. Training set size and noise level are indicated.

Case Study



Case Study

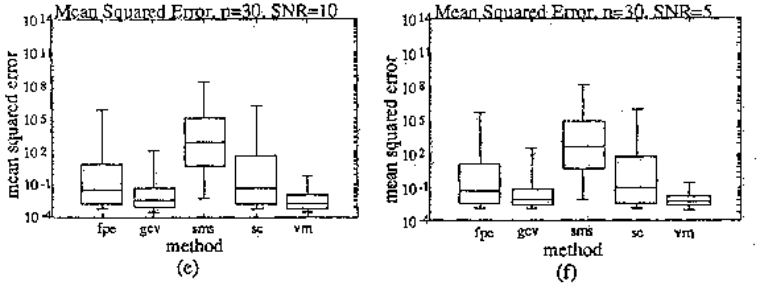
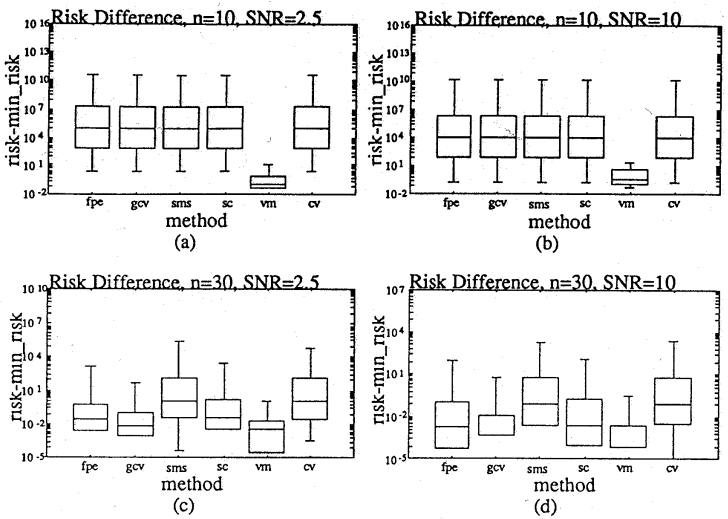


Figure 4.11 Mean squared error for each model selection approach. Training set size and noise level are indicated.

Case Study



Case Study

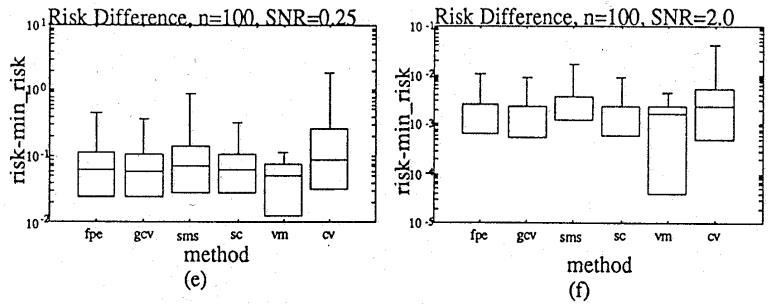


Figure 4.13 Risk difference for each model selection approach. Training set size and noise level are indicated.

4.6 Summary

- SLT framework can be used in three ways:
  - For the interpretation and critical evaluation of empirical learning methods developed in statistics and neural networks.
  - For developing new constructive learning procedures based on SLT.
  - For developing new inductive principles, such as transductive inference and local risk minimization.
- Comments on SLT framework
  - SLT sometimes doesn't seems to conform with real and complex problems and we cannot expect SLT to provide immediate and clear solutions to practical problems. With all these difficulties, all learning methods must be consistent in SLT senses in order to be a reliable one.