Bounds on the Generalization

- Risk Functions

. Let $z(\underline{x}, y)$ be an input-output sample pair and Training samples $z_1, z_2, \cdots, z_l$ be generated from the probability density function $p(z)$. Then, the risk function is defined by

$$R(\alpha) = \int Q(z,\alpha) p(z) dz$$

where α is a parameter in Λ (parameter set) and $Q(z,\alpha)$ is a loss function.

. Examples of loss functions:

(1) classification

$$Q(z,\alpha) = \begin{cases} 1 & \text{if } \; y \neq f(x,\alpha) \\ 0 & otherwise \end{cases}$$

where $f(x,\alpha)$ is an estimation function.

(2) regression

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

. Empirical Risk

Empirical risk is defined by

$$R_{emp}(\alpha) = \frac{1}{l} \sum_{i=1}^{l} Q(z_i,\alpha).$$

What is the relationship between $R(\alpha)$ and $R_{emp}(\alpha)$?

- Empirical Risk Minimization (ERM) Principle

. Let

$R_{emp}(\alpha^*|l)$ be the optimal value provided by $Q(z,\alpha^*|l)$ minimizing the empirical risk for l I. I. d. samples and

$R(\alpha^*|l)$ be the true risk for $Q(z,\alpha^*|l)$.

. The ERM principle is consistent if the true risk $R(\alpha^*|l)$ and the empirical risk $R_{emp}(\alpha^*|l)$ converge to the same limit,

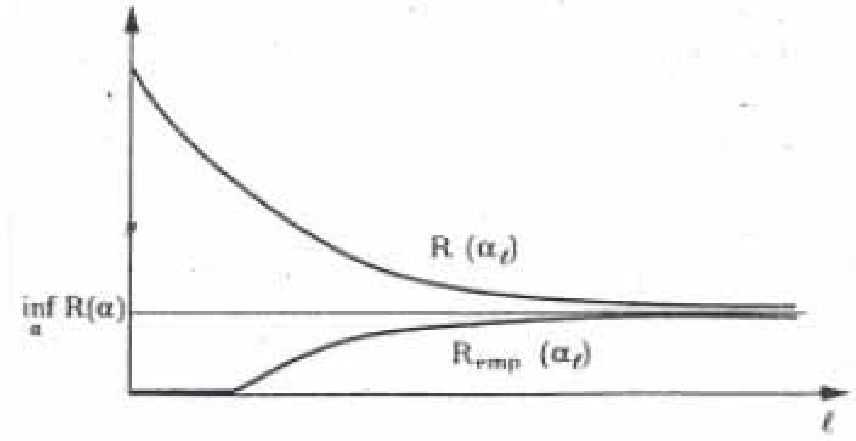
$R(\alpha_0) = \min_\alpha R(\alpha)$ as l goes to infinity, that is,

$\lim_{l\to\infty} R_{emp}(\alpha^*|l) = R(\alpha_0)$ and

$\lim_{l\to\infty} R(\alpha^*|l) = R(\alpha_0)$.

. For the bounded loss functions, the ERM principle is consistent if and only if the empirical risk converges uniformly to the true risk in the following sense:

$$\lim_{l\to\infty} \Pr\left[supr_\alpha |R(\alpha) - R_{emp}(\alpha)| > \epsilon\right] = 0, \quad \forall \epsilon > 0.$$



. The asymptotic rate of convergence is called fast if for any $l > l_0$,

$$\Pr[R(\alpha) - R(\alpha^*) < \epsilon] = e^{-cl\epsilon^2}$$

where c is a positive constant.

. A distribution independent condition (both necessary and sufficient) for the consistency of ERM and fast convergence (Vapnik and Chervonenkis, 1989):

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

where $G(l) = \ln \Pi_H(l) \leqq d(1 + \ln \frac{l}{d})$.

- The Simple Model

. Let us assume that $|\Lambda|$ is finite, that is,

$$Q(z, \alpha_k), \quad k = 1, 2, \cdots, N.$$

For example, each parameter has discrete values within a certain range.

Then,

$$\Pr[supr_{1 \leqq k \leqq N}(\int Q(z,\alpha_k)p(z)dz - \frac{1}{l}\sum_{i=1}^{l} Q(z_i,\alpha_k)) > \epsilon]$$
$$\leqq \sum_{i=1}^{N} \Pr[\int Q(z,\alpha_k)p(z)dz - \frac{1}{l}\sum_{i=1}^{l} Q(z_i,\alpha_k) > \epsilon]$$
$$\leqq Ne^{-2\epsilon^2 l}$$

cf. additive Chernoff bound:

$\Pr[p - \hat{p} > \epsilon] < e^{-2\epsilon^2 l}$ or

$\Pr[\hat{p} - p > \epsilon] < e^{-2\epsilon^2 l}$.

. Let $Ne^{-2\epsilon^2 l} = \delta$. Then,

$$\epsilon = \sqrt{\frac{\ln N - \ln\delta}{2l}}.$$

Therefore, with the probability at least $1-\delta$ for all N functions in the set $Q(z,\alpha_k)$, $k = 1, 2, \cdots, N$,

$$\int Q(z,\alpha)p(z)dz - \frac{1}{l}\sum_{i=1}^{l} Q(z_i,\alpha_k) \leqq \sqrt{\frac{\ln N - \ln\delta}{2l}}$$

This implies that

$$R(\alpha_k) \leqq R_{emp}(\alpha_k) + \sqrt{\frac{\ln N - \ln\delta}{2l}}, \quad \forall\, \alpha_k \in \Lambda. \quad \ldots \quad (1)$$

. Let α_0 be the best parameter for $R(\alpha)$. Then,

$$\Pr[\int Q(z,\alpha_0)p(z)dz - \frac{1}{l}\sum_{i=1}^{l} Q(z_i,\alpha_0) > \epsilon] \leqq e^{-2\epsilon^2 l}.$$

This implies that

$$R(\alpha_0) \leqq R_{emp}(\alpha_0) + \sqrt{\frac{-\ln\delta}{2l}}.$$

. multiplicative Chernoff bound:

$$\Pr[\hat{p} < (1-r)p] < e^{-\frac{\gamma^2 pl}{2}}$$ or

$$\Pr[\hat{p} > (1+\gamma)p] < e^{-\frac{\gamma^2 pl}{3}}.$$

Let $\gamma = \frac{\epsilon}{\sqrt{p}}$. Then,

$$\Pr[\frac{p-\hat{p}}{\sqrt{p}} > \epsilon] < e^{-\frac{\epsilon^2 l}{2}}$$ or

$$\Pr[\frac{\hat{p}-p}{\sqrt{p}} > \epsilon] < e^{-\frac{\epsilon^2 l}{3}}.$$

If we apply the multiplicative Chernoff bound,

$$\Pr[supr_{1 \leqq k \leqq N} \frac{R(\alpha_k) - R_{emp}(\alpha_k)}{\sqrt{R(\alpha_k)}} > \epsilon]$$
$$\leqq \sum_{i=1}^{N} \Pr[\frac{R(\alpha_k) - R_{emp}(\alpha_k)}{\sqrt{R(\alpha_k)}} > \epsilon]$$
$$\leqq Ne^{-\frac{\epsilon^2 l}{2}}$$

Let $Ne^{-\frac{\epsilon^2 l}{2}} = \delta$. Then,

$$\epsilon = \sqrt{2\frac{\ln N - \ln\delta}{l}}.$$

Therefore, with the probability at least $1-\delta$ for all N functions in the set $Q(z,\alpha_k)$, $k=1,2,\cdots,N$,

$$\frac{R(\alpha_k)-R_{emp}(\alpha_k)}{\sqrt{R(\alpha_k)}} \leqq \epsilon$$

where

$$\epsilon = \sqrt{2\frac{\ln N - \ln\delta}{l}}\ .$$

That is,

$R(\alpha_k) \leqq R_{emp}(\alpha_k)+\epsilon\sqrt{R(\alpha_k)}$, $\quad \forall\,\alpha_k \in \Lambda$. This implies that

$$R(\alpha_k) \leqq R_{emp}(\alpha_k)+\frac{\epsilon^2}{2}(1+\sqrt{1+\frac{4R_{emp}(\alpha_k)}{\epsilon^2}}), \quad \forall\,\alpha_k \in \Lambda. \quad \ldots \quad (2)$$

Note the second term in the righthand side of the above inequality depends on $R_{emp}(\alpha_k)$.

If $R_{emp}(\alpha_k)=0$, the second term becomes

$$\epsilon^2 = \frac{2(\ln N - \ln\delta)}{l}.$$

This gives more tight bound than (1).

- Generalization Bounds for the Finite VC Dimension

If $|H|$ is infinite, we need to consider generalization bounds using the VC dimension of H.

Lemma:

For the finite $VCD(H) = d$ and $l \geqq 8/\epsilon$, the following inequality holds:

$$\Pr\left[supr_{\alpha \in \Lambda} R(\alpha) - R_{emp}(\alpha) > \epsilon\right] \leqq 2\Pi_H(2l)2^{-\frac{\epsilon l}{2}}.$$

Theorem:

For the finite $VCD(H) = d$ and $l \geqq 8/\epsilon$,

with the probability at least $1-\delta$

$$R(\alpha) \leqq R_{emp}(\alpha) + \epsilon, \quad \forall \alpha \in \Lambda$$

where

$$\epsilon = \frac{2}{\ln 2} \frac{d\left(1 + \ln\frac{2l}{d}\right) - \ln\frac{\delta}{2}}{l}.$$

(proof)

From the previous lemma,

$$\Pr\left[supr_{\alpha \in \Lambda} R(\alpha) - R_{emp}(\alpha) > \epsilon\right] \leqq 2\Pi_H(2l)2^{-\frac{\epsilon l}{2}}.$$

For PAC learning, let

$$2\Pi_H(2l)2^{-\frac{\epsilon l}{2}} = \delta.$$

Then,

$$2\left(\frac{e2l}{d}\right)^d 2^{-\frac{\epsilon l}{2}} = \delta$$

-> $d\left(1+\ln\frac{2l}{d}\right)-\frac{\epsilon l}{2}\ln 2 = \ln\frac{\delta}{2}$

-> $\epsilon = \frac{2}{\ln 2}\frac{d\left(1+\ln\frac{2l}{d}\right)-\ln\frac{\delta}{2}}{l}$

Therefore, with the probability at least $1-\delta$

$$R(\alpha) \leqq R_{emp}(\alpha)+\epsilon, \quad \forall\alpha \in \Lambda.$$

. An alternative bound (Vapnik, 1998)

The following inequality is derived using the multiplicative Chernoff bound:

$$\Pr[supr_{\alpha\in\Lambda}\frac{R(\alpha)-R_{emp}(\alpha)}{\sqrt{R(\alpha)}} > \epsilon] < 4\Pi_H(2l)e^{-\frac{\epsilon^2 l}{4}}.$$

With the probability at least $1-\delta$,

$$R(\alpha) \leqq R_{emp}(\alpha)+\frac{\epsilon^2}{2}(1+\sqrt{1+\frac{4R_{emp}(\alpha)}{\epsilon^2}}), \quad \forall\alpha \in \Lambda$$

where

$$\epsilon^2 = 4\frac{d(1+\ln\frac{2l}{d})-\ln\frac{\delta}{4}}{l}.$$

. Example: linear discriminant function

$$h(\underline{x}) = w_0 + \sum_{i=1}^{n} w_i x_i$$

Let $n = 2$, $l = 100$, and $\delta = 0.05$.

Then, $VCD(H) = 3$, and

$$\epsilon = \frac{2}{\ln 2}\frac{1}{100}\left(3\left(1 + \ln\frac{2 \cdot 100}{3}\right) - \ln\frac{0.05}{2}\right) \approx 0.56.$$

If we use Vapnik's bound,

$$\epsilon^2 = \frac{4}{100}\left(3\left(1 + \ln\frac{2 \cdot 100}{3}\right) - \ln\frac{0.05}{4}\right) \approx 0.80.$$

- Generalization Bounds for Regression

. An indicator function for the set of real-valued functions is defined:

consider a set of real-valued loss functions such that

$$A \leqq Q(z,\alpha) \leqq B.$$

an indicator function is defined by

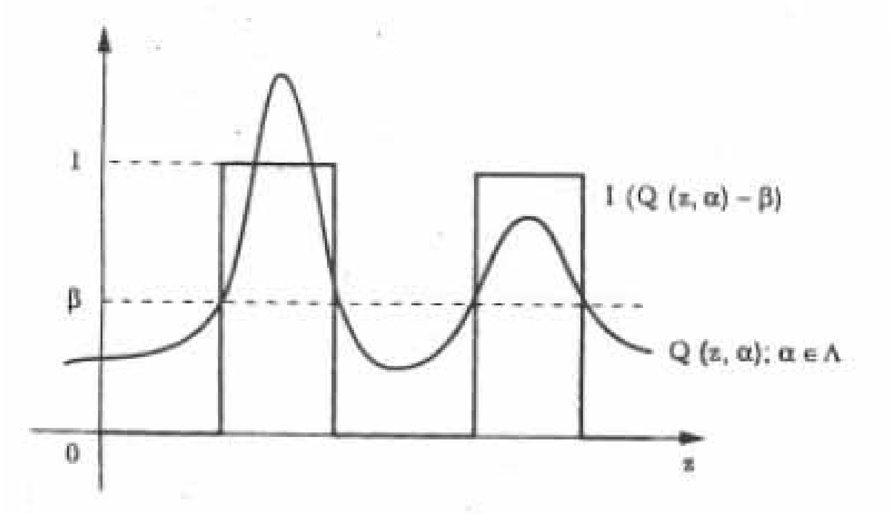
$$I(z,\alpha,\beta) = I(Q(z,\alpha) - \beta)$$

where

$A \leqq \beta \leqq B$ and

$$I(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & otherwise \end{cases}$$

. An indicator function $I(Q(z,\alpha)-\beta)$:



. Theorem: VC dimension of loss functions (Vapnik, 1995)
Let $Q(z,\alpha)=(y-f(x,\alpha))^2$. Then,
the VC dimension d of $Q(z,\alpha)$ is bounded by

$$d_f \leqq d \leqq c d_f$$

where c is a positive constant and h_f is the VC dimension of $f(x,\alpha)$.

. The VC dimension of $Q(z,\alpha)$ is by definition equal to the VC dimension of the set of indicator functions with parameters α and β.

. Lemma: For the bounded loss function $Q(z,\alpha)$, the following inequality holds:

$$\Pr\left[supr_{\alpha\in\Lambda}\frac{R(\alpha)-R_{emp}(\alpha)}{\sqrt{R(\alpha)}}<\epsilon\right]<4\Pi_H(2l)e^{-\frac{\epsilon^2 l}{4(B-A)}}.$$

. Theorem (Vapnik, 1998): With the probability at least $1-\delta$, the following inequality holds:

$$R(\alpha)\leqq R_{emp}(\alpha)+\frac{\epsilon^2}{2}(1+\sqrt{1+\frac{R_{emp}(\alpha)}{\epsilon^2}})$$

where

$$\epsilon^2=4(B-A)\frac{d(1+\ln\frac{2l}{d})-\ln\frac{\delta}{4}}{l}.$$

. generalization bounds for non-negative (not necessarily bounded) loss functions

Theorem (Vapnik, 1998): Suppose we non-negative (not necessarily bounded) loss function $Q(z,\alpha)$ and $p>2$. Then, with the probability at least $1-\delta$,

$$R(\alpha)\leqq\frac{R_{emp}(\alpha)}{(1-\tau^* a(p)\epsilon)_+}$$

where

$$supr_{\alpha\in\Lambda}\frac{\sqrt[p]{\int Q^p(z,\alpha)p(z)dz}}{\int Q(z,\alpha)p(z)dz}=\tau_p<\tau^*,$$

$a(p) = \sqrt[p]{\frac{1}{2}\left(\frac{p-1}{p-1}\right)^{p-1}}$, $\epsilon = \sqrt{4\frac{G(2l) - \ln\frac{\delta}{4}}{l}}$, and

$(u)_+ = \max(u,0)$.

. Estimating the bounds of VC dimension from samples

Let

$k = \tau^* a(p)$.

Then, for most practical regression, we can safely assume that

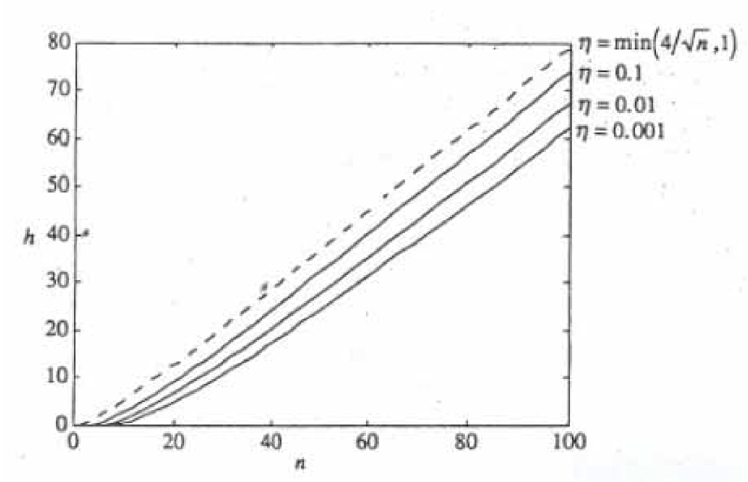
$k = 1$

and ϵ should be less than 1.

One way is to estimate ϵ from samples after learning, that is, estimating parameters a_1 and a_2 in

$$\epsilon = a_1\frac{d(1+\ln\frac{a_2 l}{d}) - \ln\frac{\delta}{4}}{l}.$$

For example, $\epsilon \geqq 1$ with $a_1 = 1$ and $a_2 = 1$.



If $l = 50$ and $\delta = 0.1$, $d < 32$.

- Summary of Generalization Bounds

(1) classification:

For all $\alpha \in \Lambda$, the following inequality holds:

$$R(\alpha) \leqq R_{emp}(\alpha) + \epsilon$$

where

$$\epsilon = O(\frac{G(l)}{l}) \text{ and } G(l) = d(1 + \ln\frac{l}{d}).$$

(2) regression with bounded loss functions:

For all $\alpha \in \Lambda$, the following inequality holds:

$$R(\alpha) \leqq R_{emp}(\alpha) + \epsilon.$$

(3) regression with unbounded loss functions:

For all $\alpha \in \Lambda$, the following inequality holds:

$$R(\alpha) \leqq \frac{R_{emp}(\alpha)}{(1-\sqrt{\epsilon})_+}.$$

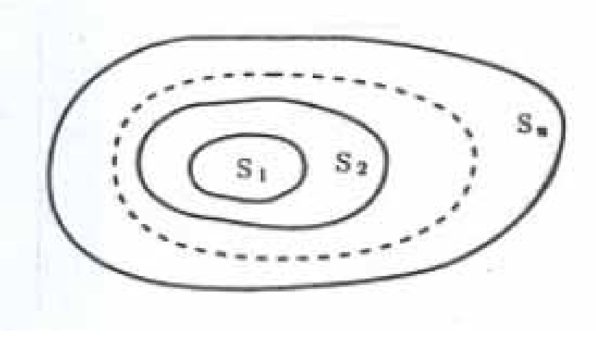
- Structural Risk Minimization (SRM) Principle

. The structure Σ on a set S of loss functions $Q(z,\alpha)$ is defined by the set of nested subset of functions

$$S_1 \subset S_2 \subset \cdots \subset S_k \subset \cdots$$

where

$$S_k = \{Q(z,\alpha) | \alpha \in \Lambda_k\}.$$



. admissible structures: the structures satisfying the following properties:

(1) any element S_k of structure Σ has a finite VCD d_k.

(2) any element S_k of structure Σ contains either

(a) a set of totally bounded functions

$0 \leqq Q(z,\alpha) \leqq B_k$, $\alpha \in \Lambda_k$ or

(b) a set of non-negative functions $Q(z,\alpha)$, $\alpha \in \Lambda_k$ satisfying the inequality

$$supr_{\alpha \in \Lambda_k} \frac{\sqrt[p]{E[Q^p(z,\alpha)}}{E[Q(z,\alpha)]} \leqq \tau_k < \infty.$$

. Example:

(1) the sequence of VCD h_k for the element S_k

$h_1 \leqq h_2 \leqq \cdots \leqq h_k \leqq \cdots$

(2)-a the sequence of the bound B_k for the element S_k

$B_1 \leqq B_2 \leqq \cdots \leqq B_k \leqq \cdots$

(2)-b the sequence of the bound τ_k for the element S_k

$\tau_1 \leqq \tau_2 \leqq \cdots \leqq \tau_k \leqq \cdots$

. Example:

(1) bounded loss functions: $0 \leqq Q(z,\alpha) \leqq B_k$, $\quad \alpha \in \Lambda_k$

$$R(\alpha_l^k) \leqq R_{emp}(\alpha_l^k) + B_k \epsilon_k(l) \left(1 + \sqrt{1 + \frac{4R_{emp}(\alpha_l^k)}{B_k \epsilon_k(l)}} \right)$$

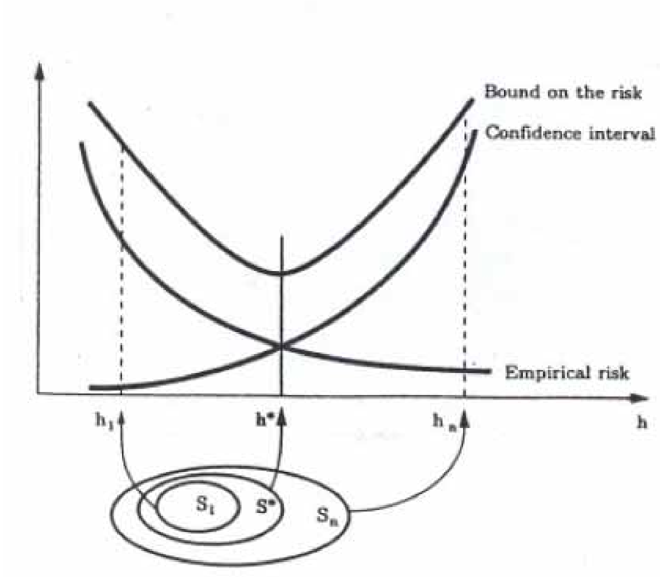
(2) non-negative loss functions: $0 \leqq Q(z,\alpha)$, $\quad \alpha \in \Lambda_k$

$$R(\alpha_l^k) \leqq \frac{R_{emp}(\alpha_l^k)}{(1 - a(p)\tau_k \sqrt{\epsilon_k(l)}\,)_+}$$

where

$$\epsilon_k(l) = 4 \frac{h_k(1 + \ln \frac{2l}{h_k}) - \ln \frac{\delta}{4}}{l}.$$

. For a given set of samples $z_1, \cdots, z_l$, the SRM method chooses the element S_k of the structure for which the smallest bound on the risk is achieved.



. Asymptotic analysis of the SRM principle

A law determining, for any given l, the number $n = n(l)$ of the elements S_n of the structure Σ in which we will minimize the empirical risk.

Theorem (Vapnik, 1998): The SRM method provides approximations $Q(z, \alpha_l^{n(l)})$ for which the sequence of risks $R(\alpha_l^{n(l)})$ converges to the smallest risk

$$R(\alpha_0) = INF_{\alpha \in \Lambda} \int Q(z,\alpha) p(z) dz$$

with the asymptotic rate of convergence

$$V(l) = r_{n(l)} + T_{n(l)} \sqrt{\frac{h_{n(l)} \ln l}{l}} \quad \text{and} \quad r_{n(l)} = R(\alpha_0^{n(l)}) - R(\alpha_0)$$

if the law $n = n(l)$ satisfies

$$\lim_{l \to \infty} \frac{T_{n(l)}^2 h_{n(l)} \ln l}{l} = 0$$

where

(a) $T_n = B_n$ for a structure with totally bounded functions and

(b) $T_n = \tau_n$ for a structure with non-negative functions.

. Example:

Let $Q(z,\alpha)$, $\alpha \in \Lambda$ be non-negative loss function for $p=2$ and $\tau_n < \tau^* < \infty$.

Consider a structure for which $h_n = n$ and let

$$r_n = \left(\frac{1}{n}\right)^c.$$

Determine n such that the asymptotic rate of convergence reaches its maximum.

Here, the asymptotic rate of convergence is given by

$$V(l) = r_n + T_n\sqrt{\frac{h_n \ln l}{l}} = \left(\frac{1}{n}\right)^c + T_n\sqrt{\frac{n \ln l}{l}}\ .$$

Find n such that $\dfrac{\partial V}{\partial n} = 0$. That is,

$$-cn^{-c-1} + \frac{T_n}{2}\left(\frac{n\ln l}{l}\right)^{-1/2}\frac{\ln l}{l} = 0.$$

-> $\quad n = \left(\dfrac{T_n}{2c}\right)^{-2/(2c+1)}\left(\dfrac{\ln l}{l}\right)^{-1/(2c+1)}$

-> $\quad n \propto \left[\dfrac{l}{\ln l}\right]^{1/(2c+1)}$

where $[a]$ is the integer part of a.

In this case, the asymptotic rate of convergence is given by

$$V(l) \propto \left(\frac{\ln l}{l}\right)^{c/(2c+1)}.$$

- Optimization of Regression Models (Model Selection Methods)

. The form of the estimate of risk functions:

$$\hat{R}(f_n) = R_{emp}(f_n)\,T(n,l)$$

where $T(n,l)$ represents the complexity term associated with the hypothesis space with n parameters and l samples.

. Akaike Information Criteria (AIC) and
Bayesian Information Criteria (BIC) are given as follows:

$$T_{AIC}(n,l) = \frac{1+n/l}{1-n/l}$$

$$T_{BIC}(n,l) = 1 + \frac{\ln l}{2}\frac{n/l}{(1-n/l)}$$

Here, AIC and BIC model selection criteria come from the asymptotic analysis for linear models using mean square error. That is, AIC and BIC are good when we use linear regression models with large number of samples.

What if we use nonlinear models with small number of samples?
-> one of candidates is using the model selection criteria
based on VC dimension.

. VC dimension based information criteria (Cherkassky et al., 1999)

The risk estimate is bounded by

$$\hat{R}(f_n) = R_{emp}(f_n)\left(1 - c\sqrt{\frac{d(1+\ln\frac{l}{d}) - \ln\delta}{l}}\right)_{+}^{-1}$$

where c is a constant dependent upon norm and tails of the loss function distribution.

-> applicable to nonlinear models even for small number of samples.

-> hard to estimate the VC dimension of nonlinear models

. Comments on model selection methods

1) AIC and BIC are good model selection criteria when we use linear regression models for large number of samples.

2) VC dimension based model selection criteria are good for nonlinear models even for small number of samples.

3) In practice, if target functions are of simple forms and the regression models are trained for small number of samples, VC dimension based model selection criteria shows better performance than AIC or BIC.

4) However, the accurate estimation of VC dimension for nonlinear models is difficult. An alternative criteria which can be estimated from samples such as the modulus of continuity information criteria (MCIC) is required.

References

. M. Anthony and N. Biggs, "Computational Learning Theory," Cambridge University Press, 1992, chapters 3, 7, and 8.

. V. Vapnik, "Statistical Learning Theory," John Wiley and Sons, 1998, chapters 3, 4, 5, and 6.

. V. Cherkassky and F. Mulier, "Learning from Data," John Wiley and Sons, 1998, chapter 4.