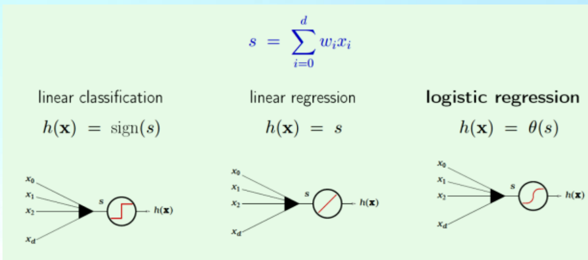
Course ID 80250993 Dates: 9/16-12/23/2021 @ I-205
Tencent/VooV Meeting 521 4678 6257 (passwd 1205)

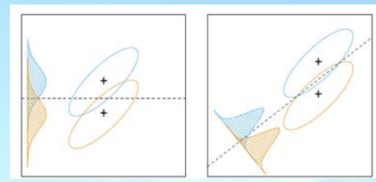
Chapter 7
Statistical Learning Theory

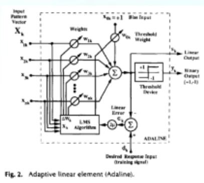
Xuegong Zhang
October 21, 2021

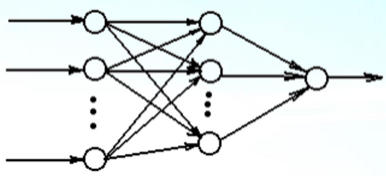
7.1
Introduction

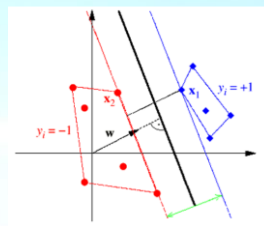
Types of machines we have learned

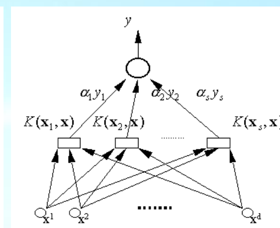








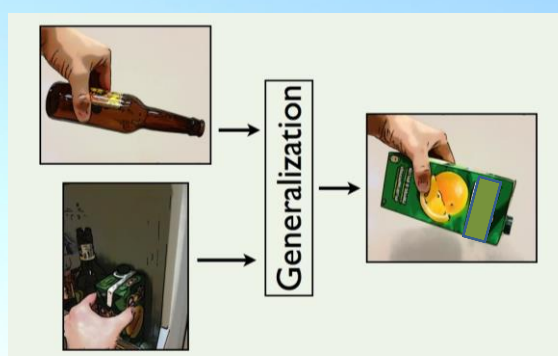




- We have learned quite a few methods, and all seem reasonable. Which one is *better*?

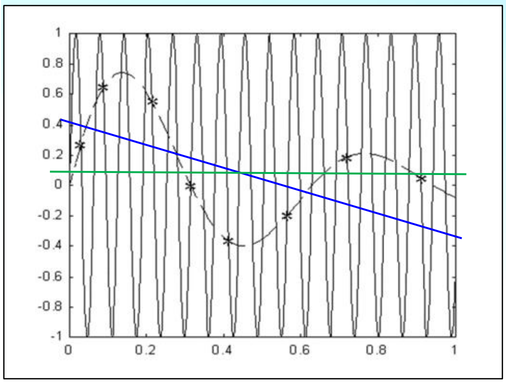
"Better" in what sense?

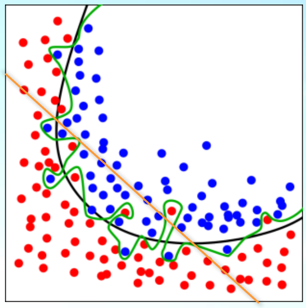
- Generalization!



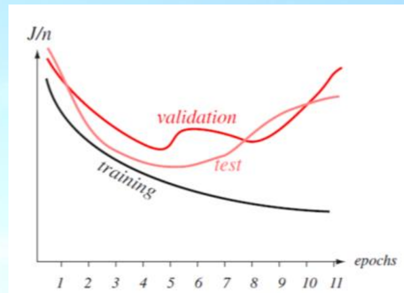
Generalization of LMs

- Under-fitting
- **Over-fitting**





https://en.wikipedia.org/wiki/Overfitting



SVM: Why largest margin is optimal?

- **Generalization**: the expected performance of a machine on future samples after being trained on limited samples
  - The difference between the expected risk and empirical risk
- **Statistical Learning Theory**
  - Large margin

    ➔Low VC dimension

    ➔Low complexity

    ➔High **generalization** ability

$$R(w) \leq R_{\mathrm{emp}}(w) + \Phi\left(\frac{h}{n}\right)$$

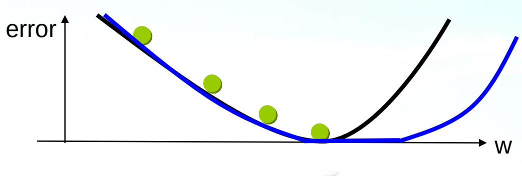
Perceptron

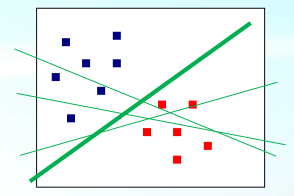
- To train the machine with data $y = \mathrm{sgn}(\sum_{i=1}^{d} w_i x_i + w_0)$
  - Goal: to optimize an objective function $J(\mathbf{w})$
    - The training error (*i.e.*, # of mistakes)
  - Learning algorithm
    - *Gradient-decreasing* $\boldsymbol{w}(t+1) = \boldsymbol{w}(t) - \eta \nabla J$

$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) + \rho_k \sum_{y_j \in \mathrm{Y}^k} y_j$$

Is this the right goal?


V. Vapnik

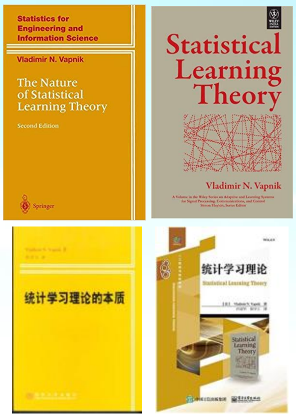




Vapnik: Two schools in the analysis of learning processes

- **Applied analysis** of the learning process
  - To get a good generalization it is sufficient to choose the coefficients of neurons that provide the minimal number of training errors.
  - It is a self-evident **inductive principle**, and does not need justification.
  - The main goal of applied analysis is to find methods for constructing the coefficients simultaneously for all neurons.
- **Theoretical analysis** of the learning process
  - The principle of minimizing the number of training errors is not self-evident and needs to be justified.
  - There could be another **inductive principle** that provides a better level of generalization ability.
  - The main goal of theoretical analysis is to find the inductive principle with the highest level of generalization ability and to construct algorithms that realize this inductive principle.

Key Points of Statistical Learning Theory

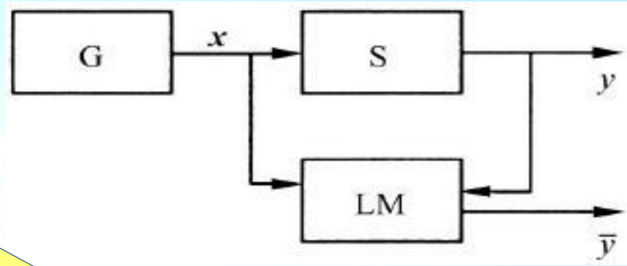


7.2 The Setting of the Machine Learning Problem

Function Estimation Model of the Learning Problem

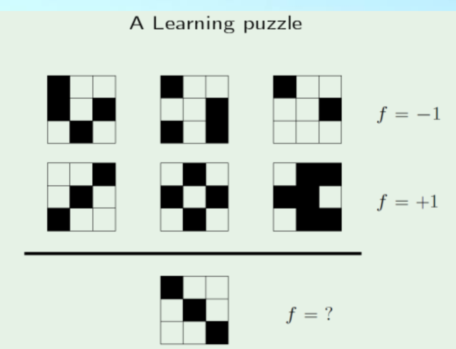
- The general model of learning from examples
  - A generator (G) of random vectors $x \in R^n$, drawn independently from a fixed but unknown probability distribution function (PDF) $F(x)$
  - A supervisor (S) who returns an output value y to every input vector x, according to a conditional distribution function $F(y|x)$, also fixed but unknown
  - A learning machine (LM) capable of implementing a set of functions $f(x, \alpha), \alpha \in \Lambda$, where Λ is a set of parameters
- The problem of learning is that of choosing from a given set of functions $f(x, \alpha), \alpha \in \Lambda$, the one that best approximates the supervisor's response.
- The selection is based on a training set of l i.i.d. observations drawn according to $F(x, y) = F(x)F(y|x)$:

$$(x_1, y_1), \cdots, (x_l, y_l)$$



What is ***the best***?

The feasibility of learning and *values* of a machine

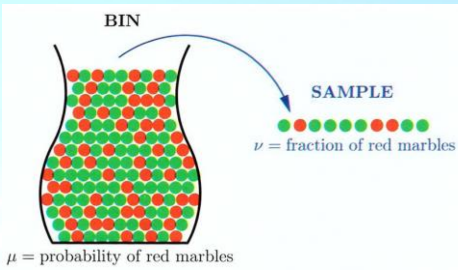


x	y	g	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
0 0 0	○	○	○	○	○	○	○	○	○	○
0 0 1	●	●	●	●	●	●	●	●	●	●
0 1 0	●	●	●	●	●	●	●	●	●	●
0 1 1	○	○	○	○	○	○	○	○	○	○
1 0 0	●	●	●	●	●	●	●	●	●	●
1 0 1		?	○	○	○	○	●	●	●	●
1 1 0		?	○	○	●	●	○	○	●	●
1 1 1		?	○	●	○	●	○	●	○	●

Learning for "unseen samples" is impossible in the deterministic sense.

Abu-Mostafa, Magdon-Ismail, Lin, *Learning from Data*

The feasibility of learning and *values* of a machine



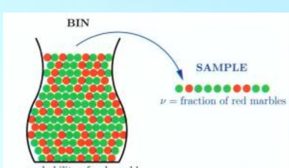
A simpler question:

- Can we predict the color of the next ball?
  - No if we insist on a deterministic answer
  - Yes if we accept a probabilistic answer

Hoeffding Inequality:

$$P(|\nu - \mu| > \epsilon) \leq 2e^{-2\epsilon^2 N}, \forall \epsilon > 0$$

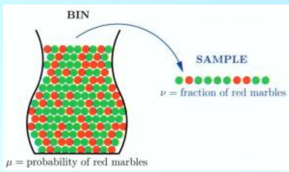
The feasibility of learning and *values* of a machine

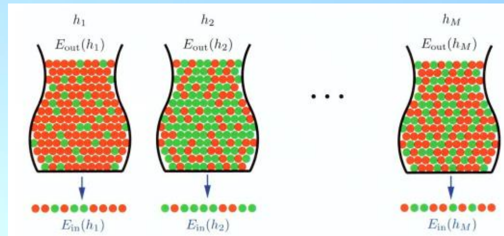


- Hoeffding Inequality:
  $P(|\nu - \mu| > \epsilon) \leq 2e^{-2\epsilon^2 N}, \forall \epsilon > 0$

- ML case (for a fixed hypothesis h):
  $$P(|E_{in}(h) - E_{out}(h)| > \epsilon) \leq 2e^{-2\epsilon^2 N}, \forall \epsilon > 0$$
  - *In-sample error* (training error)
  $$E_{in}(h) = \frac{1}{N}\sum_{i=1}^{N} [\![h(x_i) \neq f(x_i)]\!]$$
  Error count
  - *Out-of-sample error* (expected error on future data)
  $$E_{out}(h) = P\big(h(x) \neq f(x)\big)$$
  Learned function (hypothesis)
  True function (unknown)

The feasibility of learning and *values* of a machine





- Hoeffding Inequality:

$$P(|\nu - \mu| > \epsilon) \le 2e^{-2\epsilon^2 N}, \forall \epsilon > 0$$

- ML case (for a fixed hypothesis h):

$$P(|E_{in}(h) - E_{out}(h)| > \epsilon) \le 2e^{-2\epsilon^2 N}, \forall \epsilon > 0$$

  - *In-sample error* (training error)

$$E_{in}(h) = \frac{1}{N}\sum_{i=1}^{N} [\![h(x_i) \ne f(x_i)]\!]$$

  - *Out-of-sample error* (expected error on future data)

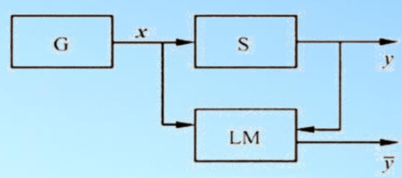
$$E_{out}(h) = P\big(h(x) \ne f(x)\big)$$

- ML case (for M hypotheses):

$$P(|E_{in}(g) - E_{out}(g)| > \epsilon) \le 2Me^{-2\epsilon^2 N}, \forall \epsilon > 0$$

- **Two key questions:**
  - Can we make sure $E_{out}(g)$ is close enough to $E_{in}(g)$?
  - Can we make $E_{in}(g)$ small enough?

The problem of Risk Minimization



Loss function:

$$L(y, f(x, \alpha))$$

Risk Functional 风险泛函:

$$R(\alpha) = \int L(y, f(x, \alpha)) dF(x, y)$$

The goal of learning:

To find the function $f(x, \alpha_0)$ that minimizes the risk functional $R(\alpha)$ over the class of functions $f(x, \alpha), \alpha \in \Lambda$ in the situation when the joint PDF $F(x, y)$ is unknown and the only available information is contained in the training set $(x_1, y_1), \cdots, (x_l, y_l)$.

Three main learning problems

Pattern Recognition

- Indicator function: functions which take only two values: 0 and 1
- Loss function

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

- Risk: probability of classification error

$$R(\alpha) = \int L(y, f(x,\alpha))dF(x,y)$$

- The problem of PR:

To find a function that minimizes the probability of classification error when the probability measure $F(x,y)$ is unknown but the data $(x_1,y_1),\cdots,(x_l,y_l)$ are given.

Regression Estimation

- Set of real functions $f(x,\alpha), \alpha \in \Lambda$
- Loss function

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

- Risk:

$$R(\alpha) = \int L(y, f(x,\alpha))dF(x,y)$$

- Regression Estimation:

To find a function that minimizes the risk functional when the probability measure $F(x,y)$ is unknown but the data $(x_1,y_1),\cdots,(x_l,y_l)$ are given.

Density Estimation

- Set of density functions $f(x,\alpha), \alpha \in \Lambda$
- Loss function

$$L(p(x,\alpha)) = -\log p(x,\alpha)$$

- Risk:

$$R(\alpha) = \int L(f(x,\alpha))dF(x)$$

- Density estimation:

To minimize the risk functional when the corresponding probability measure $F(x)$ is unknown, but i.i.d. data $x_1,\cdots,x_n$ is given.

The general setting of the learning problem

Let's use z to represent the pair (x, y) for convenience.

Let the probability measure $F(z)$ be defined on the space Z. Consider the set of functions $Q(z, \alpha), \alpha \in \Lambda$. The goal is to minimize the risk functional

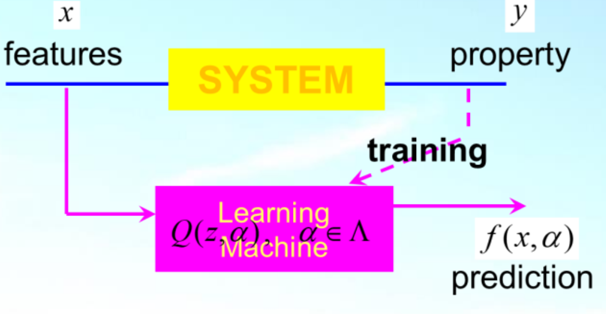
$$R(\alpha) = \int Q(z, \alpha) dF(z), \ \ \alpha \in \Lambda,$$

where the probability measure $F(z)$ is unknown, but an i.i.d. sample

$$z_1, \cdots, z_l$$

is given.

Supervised Learning



$$\min \quad J_P(\boldsymbol{\alpha}) = \sum_{y_j \in \Upsilon^k} (-\boldsymbol{\alpha}^T \boldsymbol{y}_j)$$

$$\min \quad E(\boldsymbol{w}) = \frac{1}{N} \sum_{j=1}^{N} (\boldsymbol{w}^T \boldsymbol{x}_j - y_j)^2$$

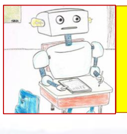
$$\min \quad E(\boldsymbol{w}) = \frac{1}{N} \sum_{j=1}^{N} ln\left(1 + e^{-y_j \boldsymbol{w}^T \boldsymbol{x}_j}\right)$$

$$\min \quad E = \tfrac{1}{2} \textstyle\sum_j (y_j - \hat{y}_j)^2$$

$$\min \quad \Phi(\mathbf{w}, \boldsymbol{\xi}) = \frac{1}{2}(\mathbf{w} \cdot \mathbf{w}) + C\left(\sum_{i=1}^{l} \xi_i\right)$$

All defined on training data

$$\min \quad R(\alpha) = \int Q(z, \alpha) dF(z)$$



Fine, but how can we get the $R(\alpha)$ and minimize it if $F(z)$ is unknown?

The ERM inductive principle

$R(\alpha) = \int Q(z,\alpha)dF(z)$

Expected risk

Empirical risk functional

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

Empirical Risk Minimization (ERM)

To approximate the function $Q(z,\alpha_0)$ that minimizes the expected risk by the function $Q(z,\alpha_l)$ that minimizes the empirical risk.

Examples:

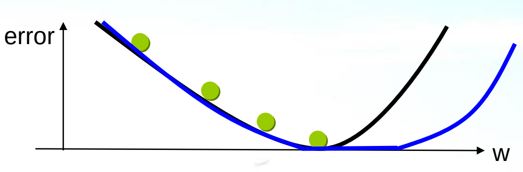
- Linear regression estimation: $\min R_{\text{emp}}(\alpha) = \frac{1}{l}\sum_{i=1}^{l}\left(y_i - f(x_i,\alpha)\right)^2$
- Maximum likelihood estimation of PDF: $\min R_{\text{emp}}(\alpha) = \frac{1}{l}\sum_{i=1}^{l} \ln p(x_i,\alpha)$

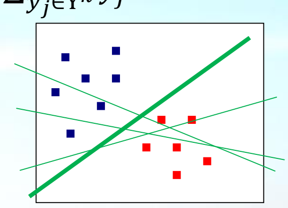
Perceptron

- To train the machine with data $y = \text{sgn}(\sum_{i=1}^{n} w_i x_i + w_0)$
  - Goal: to optimize an objective function $J(\mathbf{w})$
    - The training error (*i.e.*, # of mistakes)
  - Learning algorithm
    - *Gradient-decreasing* $\mathbf{w}(t+1) = \mathbf{w}(t) - \eta\nabla J$

$$\mathbf{w}(t+1) = \mathbf{w}(t) + \rho_k \sum_{y_j \in Y^k} y_j$$

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





Questions about ERM

What we managed to do

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

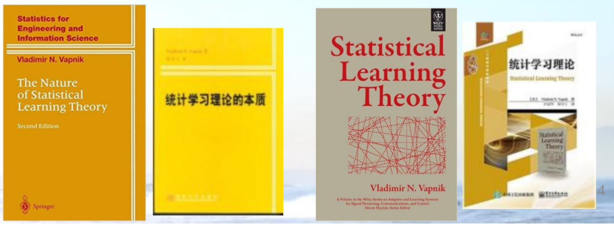
What we wanted to do

$$\min \ R(\alpha) = \int Q(z,\alpha)dF(z), \quad \alpha \in \Lambda$$

- Will empirical risk $R_{\mathrm{emp}}(\alpha)$ converge to expected risk $R(\alpha)$ when $l \to \infty$?
  - Will the solution $Q(x, \alpha_l)$ converge to the solution $Q(x, \alpha_0)$?
- For limited number of samples:
  - How well $Q(x, \alpha_l)$ approximates $Q(x, \alpha_0)$?
  - What is the $R(\alpha_l)$ for the α_l that minimizes the empirical risk?
  - When there are multiple α_l that makes $R_{\mathrm{emp}}(\alpha_l) = 0$, which one is the best?

Four parts of Statistical Learning Theory

- What are the (necessary & sufficient) conditions for consistency of an ERM learning process?
- How fast is the rate of convergence of the learning process?
- How can one control the rate of convergence (the generalization ability) of the learning process?
- How can one construct algorithms that can control the generalization ability?



10-second break

7.3 Consistency of ERM Learning Processes

$\min R_{emp}(\alpha) = \frac{1}{l}\sum_{i=1}^{l} Q(z_i,\alpha) \Rightarrow \min R(\alpha) = \int Q(z,\alpha)dF(z), \quad \alpha \in \Lambda$

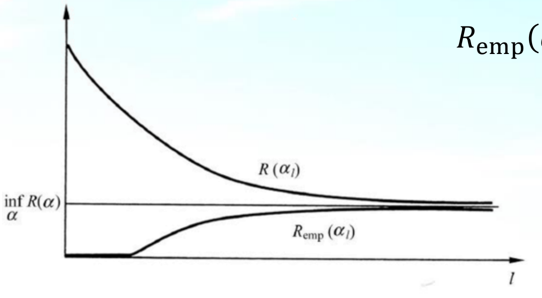
Consistency

Definition: We say that the principle (method) of ERM is *consistent* for the set of functions $Q(z,\alpha), \alpha \in \Lambda$ and for the PDF $F(z)$ if the following two sequences converge in probability to the same limit

$$R(\alpha_l) \xrightarrow[l\to\infty]{P} \inf_{\alpha\in\Lambda} R(\alpha)$$

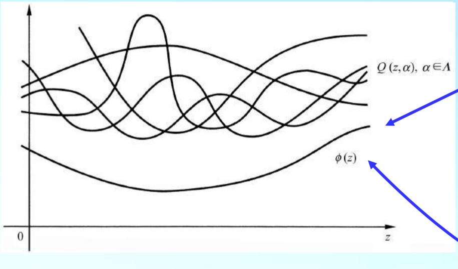
$$R_{\text{emp}}(\alpha_l) \xrightarrow[l\to\infty]{P} \inf_{\alpha\in\Lambda} R(\alpha)$$

inf: *Infimum*, the greatest lower bound of a set



The learning process is consistent if both expected risks $R(\alpha_l)$ and empirical risks $R_{emp}(\alpha_l)$ converge to the minimal possible value of the risk, $\inf_{\alpha\in\Lambda} R(\alpha)$.

Consistency can be trivial for some function sets：



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

$$R(\alpha_l) \xrightarrow[l\to\infty]{P} \inf_{\alpha\in\Lambda} R(\alpha)$$

$$R(\alpha_l) \xrightarrow[l\to\infty]{P} \inf_{\alpha\in\Lambda} R(\alpha)$$

Suppose we have established that for some set of functions $Q(z,\alpha)$, $\alpha \in \Lambda$, the ERM method is not consistent. Consider an extended set of functions that includes this set of functions and one additional function, $\phi(z)$. Suppose that the additional function satisfies the inequality

$$\inf_{\alpha\in\Lambda} Q(z,\alpha) > \phi(z), \quad \forall z.$$

It is clear (Fig. 2.2) that for the extended set of functions (containing $\phi(z)$) the ERM method will be consistent. Indeed, for any distribution function and for any number of observations, the minimum of the empirical risk will be attained on the function $\phi(z)$ that also gives the minimum of the expected risk.

Strict (Nontrivial) Consistency

Definition:

We say that the ERM method is *nontrivially consistent* for the set of functions $Q(z,\alpha)$, $\alpha \in \Lambda$ and for the PDF $F(z)$ if for any nonempty subset $\Lambda(c), c \in (-\infty, \infty)$ of this set of functions defined as

$$\Lambda(c) = \left\{ \alpha : \int Q(z,\alpha) dF(z) > c, \alpha \in \Lambda \right\}$$

the convergence $\inf\limits_{\alpha \in \Lambda(c)} R_{\mathrm{emp}}(\alpha) \underset{l \to \infty}{\overset{P}{\to}} \inf\limits_{\alpha \in \Lambda(c)} R(\alpha)$ is valid.

In the other words, the ERM is nontrivially consistent if it provides convergence for the subset of functions that remain after the functions with smallest values of the risks are excluded from this set.

The Key Theorem of Learning Theory

Theorem [Vapnik and Chervonenkis, 1989]

Let $Q(z,\alpha), \alpha \in \Lambda$ be a set of functions that satisfy the condition $A \leq \int Q(z,\alpha)(z) \leq B$, i.e., the function set has bounded loss function $R(\alpha)$, then for ERM principle to be consistent, it is *necessary and sufficient* that the empirical risk $R_{\mathrm{emp}}(\alpha)$ converge uniformly to the actual risk $R(\alpha)$ over the set $Q(z,\alpha), \alpha \in \Lambda$ in the following sense:

$$\lim_{l \to \infty} P\left\{ \sup_{\alpha} \left(R(\alpha) - R_{\mathrm{emp}}(\alpha) \right) > \varepsilon \right\} = 0, \qquad \forall \varepsilon > 0$$

sup: *supremum*, the least upper bound of a set

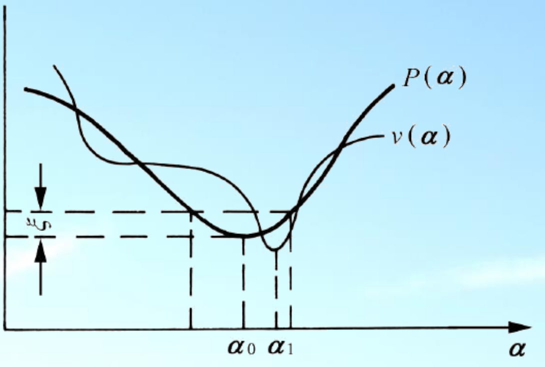
uniform one-sided convergence

determined by the "worst" function of the set of functions → Worst case analysis


(Left to right) J. Rissanen, V. Vapnik, A. Gammerman, A. Chervonenkis, C. Wallace and R. Solomonoff

D

[2]The following fact confirms the importance of this theorem. Toward the end of the 1980s and ther beginning of the 1990s several alternative approaches to learning theory were attempted based on the idea that statistical learning theory is a theory of "worst-case analysis.". In these approaches authors expressed a hope to develop a learning theory for "real-case analysis." According to the key theorem, this type of theory for the ERM principle is impossible.



Average-case $\lim_{l\to\infty} P\left\{\left|\left(R(\alpha)-R_{\mathrm{emp}}(\alpha)\right)\right|>\varepsilon\right\}=0,\ \ \forall\varepsilon>0$

Worst-case $\lim_{l\to\infty} P\left\{\sup_{\alpha}\left(R(\alpha)-R_{\mathrm{emp}}(\alpha)\right)>\varepsilon\right\}=0,\ \ \forall\varepsilon>0$

Conditions of Uniform Two-sided Convergence

Two stochastic processes (empirical processes)

Two-sided process $$\xi^l=\sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)dF(z)-\frac{1}{l}\sum_{i=1}^{l}Q(z_i,\alpha)\right|,\quad l=1,2,\cdots$$

Uniform two-sided convergence $$\lim_{l\to\infty}P\left\{\xi^l>\varepsilon\right\}=0,\quad \forall\varepsilon>0$$

One-sided process $$\xi_+^l=\sup_{\alpha\in\Lambda}\left(\int Q(z,\alpha)dF(z)-\frac{1}{l}\sum_{i=1}^{l}Q(z_i,\alpha)\right)_+,\quad l=1,2,\cdots$$

Uniform one-sided convergence $$\lim_{l\to\infty}P\left\{\xi_+^l>\varepsilon\right\}=0,\quad \forall\varepsilon>0$$

The Law of Large Numbers and it's generalization

$$\xi^l = \sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)dF(z) - \frac{1}{l}\sum_{i=1}^{l} Q(z_i,\alpha)\right|, \quad l = 1,2,\cdots$$

1. **The law of large numbers**: The sequence of the means of random variables ξ^l converges to zero as l increases.

   If the function set $Q(z,\alpha), \alpha \in \Lambda$ has only one element, $\xi^l \xrightarrow[l\to\infty]{} 0$

2. **The law of large numbers in an N-dimensional vector space** $\xi^l \xrightarrow[l\to\infty]{P} 0$

   The sequence of random variables ξ^l converges to zero in probability if the set of functions $Q(z,\alpha), \alpha \in \Lambda$ contains a finite number N of elements.

The Law of Large Numbers and it's generalization

$$\xi^l = \sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)dF(z) - \frac{1}{l}\sum_{i=1}^{l} Q(z_i,\alpha)\right|, \quad l = 1,2,\cdots$$

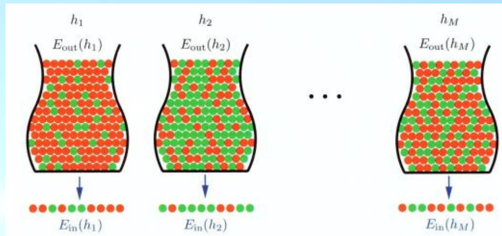
3. **The law of large numbers in the functional space**?

   The sequence of ξ^l for a set with infinite number of elements *does not necessarily converges to zero*.

   → It depends on <u>the properties of the set of functions</u> $Q(z,\alpha), \alpha \in \Lambda$.

   ? The necessary and sufficient condition of uniform two-sided convergence.

Recall: for finite number of hypotheses



- ML case (for M hypotheses):

$$P(|E_{in}(g) - E_{out}(g)| > \epsilon) \leq 2Me^{-2\epsilon^2 N}, \forall \epsilon > 0$$

What if the function set has infinite number of functions?

Xuegong Zhang Abu-Mostafa, Magdon-Ismail, Lin, *Learning from Data* 35

单选题 1分 设置

4-min break

- A Yes, I’m back to my computer.
- B Sorry, not yet.

提交

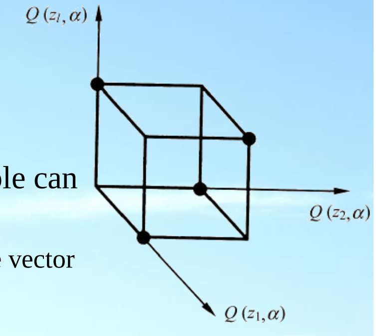
Xuegong Zhang 36

7.4 Capacity of Function Sets

Entropy of the Set of Indicator Functions

Let $Q(z,\alpha), \alpha \in \Lambda$ be a set of indicator functions.

Consider a sample $z_1, \cdots, z_l$



➢ Diversity of the function set on the given data

- $N^{\wedge}(z_1, \cdots, z_l)$: number of different separations of a given sample can be done using functions from the set.
  - Number of vertices of the l-dimensional cube that can be obtained by the vector $q(\alpha) = \big(Q(z_1,\alpha), \cdots, Q(z_l,\alpha)\big), \alpha \in \Lambda$
- Random Entropy $H^{\Lambda}(z_1, \cdots, z_l) = \ln N^{\Lambda}(z_1, \cdots, z_l)$
- VC Entropy: $H^{\Lambda}(l) = E \ln N^{\Lambda}(z_1, \cdots, z_l)$
  - The entropy of the set of indicator functions on samples of size l

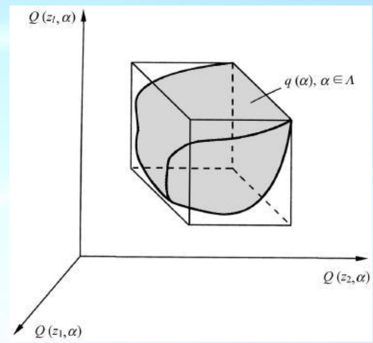
Entropy of the Set of Real Functions

Consider a set of bounded loss functions $A \leq Q(z,\alpha) \leq B, \alpha \in \Lambda$. Using this set and the given sample $z_1, \cdots, z_l$, one can construct the set of l-dimensional vectors

$$\{q(\alpha) = \big(Q(z_1,\alpha), \cdots, Q(z_l,\alpha)\big), \alpha \in \Lambda\}$$

Let $N = N^\wedge(\varepsilon; z_1, \cdots, z_l)$ be the number of elements of the "minimal ε-net" of this set of vectors $\{q(\alpha), \alpha \in \Lambda\}$.

- Random Entropy $H^\Lambda(\varepsilon; z_1, \cdots, z_l) = \ln N^\Lambda(\varepsilon; z_1, \cdots, z_l)$
- VC Entropy: $H^\Lambda(\varepsilon; l) = E \ln N^\Lambda(\varepsilon; z_1, \cdots, z_l)$



A generalization of the entropy for indicator functions

$$N^\Lambda(\varepsilon; z_1, \cdots, z_l) = N^\Lambda(z_1, \cdots, z_l)$$
$$H^\Lambda(\varepsilon; z_1, \cdots, z_l) = H^\Lambda(z_1, \cdots, z_l)$$
$$H^\Lambda(\varepsilon; l) = H^\Lambda(l)$$

Conditions for uniform two-sided convergence

Theorem [Vapnik and Chervonenkis, 1981]

For uniform two-sided convergence

$$\lim_{l\to\infty} P\left\{\sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)dF(z) - \frac{1}{l}\sum_{i=1}^{l} Q(z_i,\alpha)\right| > \varepsilon\right\} = 0, \quad \forall \varepsilon > 0$$

It is necessary and sufficient that $\lim_{l\to\infty} \frac{H^\Lambda(\varepsilon, l)}{l} = 0, \quad \forall \varepsilon > 0$

Corollary [Vapnik and Chervonenkis, 1968, 1971]

For the set of indicator functions, the necessary & sufficient condition is

$$\lim_{l\to\infty} \frac{H^\Lambda(l)}{l} = 0$$

Take-home message:

- For a sample with limited size, the capacity of the function set should not be too strong.

Necessary & Sufficient Conditions of Uniform One-sided Convergence

$$\lim_{l\to\infty} P\left\{\sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)dF(z)-\frac{1}{l}\sum_{i=1}^{l}Q(z_i,\alpha)\right|>\varepsilon\right\}=0,\quad \forall\varepsilon>0$$

$$\Updownarrow$$

$$\lim_{l\to\infty} P\left\{\left[\sup_{\alpha}\left(R(\alpha)-R_{emp}(\alpha)\right)>\varepsilon\right]\ or\left[\sup_{\alpha}\left(R_{emp}(\alpha)-R(\alpha)\right)>\varepsilon\right]\right\}=0$$

- So the necessary & sufficient condition for two-sided convergence is sufficient for one-sided convergence, but not necessary.

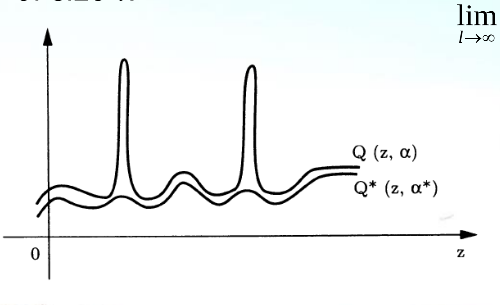
Theorem [Vapnik and Chervonenkis, 1989]

In order for uniform one-sided convergence of empirical means to their expectations to hold for the set of totally bounded functions $Q(z,\alpha),\alpha\in\Lambda$, it is **necessary and sufficient** that for any positive δ,η, and ε there exist a set of functions $Q^*(z,\alpha^*),\alpha^*\in\Lambda^*$ satisfying

$$Q(z,\alpha)-Q^*(z,\alpha^*)\geq 0,\quad \forall z,$$
$$\int(Q(z,\alpha)-Q^*(z,\alpha^*))dF(z)\leq\delta.$$

such that the following holds for the ε-entropy of the set $Q^*(z,\alpha^*),\alpha^*\in\Lambda^*$ on samples of size l:

$$\lim_{l\to\infty}\frac{H^{\Lambda^*}(\varepsilon,l)}{l}<\eta$$



- According to the Key Theorem, this is necessary and sufficient for consistency of ERM method.

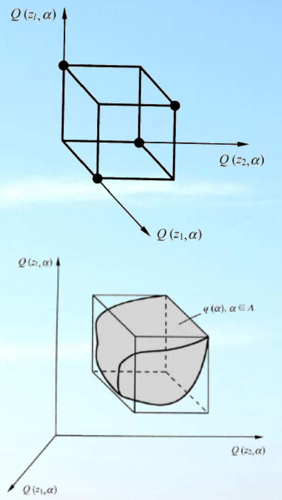
Use of the *Theory of Nonfalsifiability* to show why the ERM method is not consistent if the condition is violated.

Three Property Measures of Function Sets

Consider the set of indicator functions $Q(z,\alpha), \alpha \in \Lambda$.

- VC Entropy $H^{\Lambda}(l) = \mathrm{E}\ln N^{\Lambda}(z_1,\cdots,z_l)$
- Annealed Entropy $H_{ann}^{\Lambda}(l) = \ln \mathrm{E}\, N^{\Lambda}(z_1,\cdots,z_l)$
- Growth Function $G^{\Lambda}(l) = \ln \sup\limits_{z_1,\cdots,z_l} N^{\Lambda}(z_1,\cdots,z_l)$

They have the relation $H^{\Lambda}(l) \le H_{ann}^{\Lambda}(l) \le G^{\Lambda}(l)$



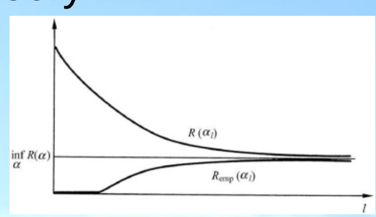
Xuegong Zhang 43

Three Milestones in Learning Theory

Consider the set of indicator functions $Q(z,\alpha), \alpha \in \Lambda$.

- VC Entropy

$$\lim_{l\to\infty} \frac{H^{\Lambda}(l)}{l} = 0 \quad \Rightarrow \quad \text{ERM consistency}$$



- Annealed Entropy

$$\lim_{l\to\infty} \frac{H_{\mathrm{ann}}^{\Lambda}(l)}{l} = 0 \quad \Rightarrow \quad \text{ERM Fast convergence}$$

$$P\left\{ R(\alpha_l) - R(\alpha_0) > \varepsilon \right\} < e^{-c\varepsilon^2 l}$$

- Growth Function

$$\lim_{l\to\infty} \frac{G^{\Lambda}(l)}{l} = 0 \quad \Leftrightarrow \quad \text{Conditions for consistency of ERM for any probability measure}$$

Xuegong Zhang 44

Four parts of Statistical Learning Theory

- What are the (necessary & sufficient) conditions for consistency of an ERM learning process?
- How fast is the rate of convergence of the learning process?
- How can one control the rate of convergence (the generalization ability) of the learning process?
- How can one construct algorithms that can control the generalization ability?

Xuegong Zhang 45

- What does "consistency" mean in applications?
  - ✓ Min Empirical Risk $\underset{l\to\infty}{\Longleftrightarrow}$ Min Expected Risk
- When is a sample large enough?
  - – a question about the convergence rate

D

Xuegong Zhang 46

10-second break

7.5
Bounds of Generalization
and VC Dimension

$\lim_{l\to\infty} P\left\{\sup_{\alpha\in\Lambda}\left(R(\alpha)-R_{emp}(\alpha)\right)>\varepsilon\right\}=0,\quad \forall\varepsilon>0$ $\quad P\{\,R(\alpha_l)-R(\alpha_0)>\varepsilon\}<e^{-c\varepsilon^2 l}$

The Basic Inequalities [Vapnik and Chervonenkis, 1968, 1971; Vapnik, 1979, 1996]

Let $Q(z,\alpha), \alpha\in\Lambda$ be a set of indicator functions, $H^{\wedge}(l)$ the corresponding VC entropy, $H_{ann}^{\wedge}(l)$ the annealed entropy and $G^{\wedge}(l)$ the growth function, the following bounds on the rate of uniform convergence hold true:

$$P\left\{\sup_{\alpha\in\Lambda}\left|R(\alpha)-R_{\text{emp}}(\alpha)\right|>\varepsilon\right\}\le 4\exp\left\{\left(\frac{H_{\text{ann}}^{\Lambda}(2l)}{l}-\varepsilon^2\right)l\right\}$$

$$P\left\{\sup_{\alpha\in\Lambda}\frac{R(\alpha)-R_{\text{emp}}(\alpha)}{\sqrt{R(\alpha)}}>\varepsilon\right\}\le 4\exp\left\{\left(\frac{H_{\text{ann}}^{\Lambda}(2l)}{l}-\frac{\varepsilon^2}{4}\right)l\right\}$$

The bounds are nontrivial if $\lim_{l\to\infty}\frac{H_{\text{ann}}^{\wedge}(l)}{l}=0$.

The 2nd milestone.

Main Distribution-Independent Bounds

Since $H_{\text{ann}}^{\Lambda}(l)\le G^{\Lambda}(l)$,

for any distribution $F(z)$, the following inequalities hold

$$P\left\{\sup_{\alpha\in\Lambda}\left|R(\alpha)-R_{\text{emp}}(\alpha)\right|>\varepsilon\right\}\le 4\exp\left\{\left(\frac{G^{\Lambda}(2l)}{l}-\varepsilon^2\right)l\right\}$$

$$P\left\{\sup_{\alpha\in\Lambda}\frac{R(\alpha)-R_{\text{emp}}(\alpha)}{\sqrt{R(\alpha)}}>\varepsilon\right\}\le 4\exp\left\{\left(\frac{G^{\Lambda}(2l)}{l}-\frac{\varepsilon^2}{4}\right)l\right\}$$

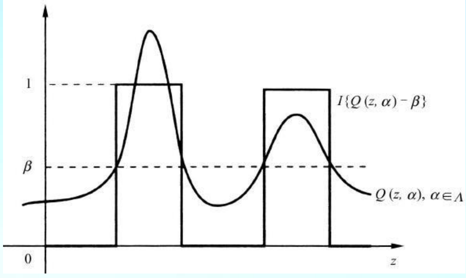
The bounds are nontrivial if $\lim_{l\to\infty}\frac{G^{\Lambda}(l)}{l}=0$.

--- necessary & sufficient for distribution-free uniform convergence.

The 3rd milestone.

- If this condition is violated, then there exist probability measures $F(z)$ on Z for which uniform convergence does not take place.

Generalization for the set of real functions



Construct a set of indicators of the set of real functions $Q(z,\alpha), \alpha \in \Lambda$:

$$I(z,\alpha,\beta) = \theta\{Q(z,\alpha) - \beta\}, \alpha \in \Lambda, \beta \in \mathrm{B}$$

where $\theta(\,)$ is the step function, and use the properties of the set of indicators to study the set of real functions.

Then we have the following for real functions:

- VC Entropy $H^{\Lambda,\mathrm{B}}(l)$
- Annealed Entropy $H_{ann}^{\Lambda,\mathrm{B}}(l)$
- Growth Function $G^{\Lambda,\mathrm{B}}(l)$

Basic Inequalities for Real Functions

(1) For totally bounded functions $A \leq Q(z,\alpha) \leq B,\ \alpha \in \Lambda$

$$P\left\{\sup_{\alpha\in\Lambda}\left|R(\alpha) - R_{\text{emp}}(\alpha)\right| > \varepsilon\right\} \leq 4\exp\left\{\left(\frac{H_{\text{ann}}^{\Lambda,\mathrm{B}}(2l)}{l} - \frac{\varepsilon^2}{(B-A)^2}\right)l\right\}$$

(2) For totally bounded nonnegative functions $0 \leq Q(z,\alpha) \leq B,\ \alpha \in \Lambda$

$$P\left\{\sup_{\alpha\in\Lambda}\frac{R(\alpha) - R_{\text{emp}}(\alpha)}{\sqrt{R(\alpha)}} > \varepsilon\right\} \leq 4\exp\left\{\left(\frac{H_{\text{ann}}^{\Lambda,\mathrm{B}}(2l)}{l} - \frac{\varepsilon^2}{4B}\right)l\right\}$$

(3) For nonnegative functions with "constrained energy*" $0 \leq Q(z,\alpha),\ \alpha \in \Lambda$

$$P\left\{\sup_{\alpha\in\Lambda}\frac{\int Q(z,\alpha)dF(z) - \frac{1}{l}\sum_{i=1}^{l}Q(z_i,\alpha)}{\sqrt[p]{\int Q^p(z,\alpha)dF(z)}} > a(p)\varepsilon\right\} \leq 4\exp\left\{\left(\frac{H_{\text{ann}}^{\Lambda,\mathrm{B}}(2l)}{l} - \frac{\varepsilon^2}{4}\right)l\right\}$$

where $a(p) = \sqrt[p]{\frac{1}{2}\left(\frac{p-1}{p-2}\right)^{p-1}}$.

* The pth moment of $Q(z,\alpha)$ $m_p(\alpha) = \sqrt[p]{\int Q^p(z,\alpha)dF(z)}$ exists.

Main Distribution-Independent Bounds

$$H_{ann}^{\Lambda,B}(l) \le G^{\Lambda,B}(l)$$

(1) For totally bounded functions $A \le Q(z,\alpha) \le B$, $\alpha \in \Lambda$

$$P\left\{\sup_{\alpha\in\Lambda}\left|R(\alpha) - R_{\text{emp}}(\alpha)\right| > \varepsilon\right\} \le 4\exp\left\{\left(\frac{G^{\Lambda,B}(2l)}{l} - \frac{\varepsilon^2}{(B-A)^2}\right)l\right\}$$

(2) For totally bounded nonnegative functions $0 \le Q(z,\alpha) \le B$, $\alpha \in \Lambda$

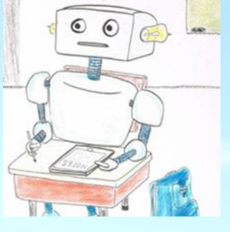
$$P\left\{\sup_{\alpha\in\Lambda}\frac{R(\alpha) - R_{\text{emp}}(\alpha)}{\sqrt{R(\alpha)}} > \varepsilon\right\} \le 4\exp\left\{\left(\frac{G^{\Lambda,B}(2l)}{l} - \frac{\varepsilon^2}{4B}\right)l\right\}$$

(3) For nonnegative functions with "constrained energy*" $0 \le Q(z,\alpha)$, $\alpha \in \Lambda$

$$P\left\{\sup_{\alpha\in\Lambda}\frac{\int Q(z,\alpha)dF(z) - \frac{1}{l}\sum_{i=1}^{l} Q(z_i,\alpha)}{\sqrt[p]{\int Q^p(z,\alpha)dF(z)}} > a(p)\varepsilon\right\} \le 4\exp\left\{\left(\frac{G^{\Lambda,B}(2l)}{l} - \frac{\varepsilon^2}{4}\right)l\right\}$$

where $a(p) = \sqrt[p]{\frac{1}{2}\left(\frac{p-1}{p-2}\right)^{p-1}}$.

* The pth moment of $Q(z,\alpha)$ $m_p(\alpha) = \sqrt[p]{\int Q^p(z,\alpha)dF(z)}$ exists.



What are all those for?

Bounds on the Generalization Ability of Learning Machines

What we managed to do

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

What we wanted to do

$$\min \; R(\alpha) = \int Q(z,\alpha)\,dF(z), \quad \alpha \in \Lambda$$

What we are really interested in:

- What is the true risk $R(\alpha_l)$ obtained by the $Q(z,\alpha_l)$ that minimizes $R_{\text{emp}}(\alpha)$?
- How close is this risk to the minimal $\inf_{\alpha} R(\alpha)$, $\alpha \in \Lambda$, for the given set of functions?

Case 1. The set of totally bounded functions

Let $A \le Q(z,\alpha) \le B$, $\alpha \in \Lambda$ be a set of totally bounded functions. Then:

(A) The following inequalities hold with probability at least $1-\eta$ simultaneously for all functions of $Q(z,\alpha), \alpha \in \Lambda$ (including the function that minimizes the empirical risk):

$$R_{emp}(\alpha) - \frac{(B-A)}{2}\sqrt{\mathcal{E}} \le R(\alpha) \le R_{emp}(\alpha) + \frac{(B-A)}{2}\sqrt{\mathcal{E}}$$

(B) The following inequality holds with probability at least $1-2\eta$ for the function $Q(z,\alpha_l)$ that minimizes the empirical risk:

$$R(\alpha_l) - \inf_{\alpha\in\Lambda} R(\alpha) \le (B-A)\sqrt{\frac{-\ln\eta}{2l}} + \frac{(B-A)}{2}\sqrt{\mathcal{E}}$$

$$\mathcal{E} = 4\,\frac{G^{\Lambda,B}(2l) - \ln\left(\frac{\eta}{4}\right)}{l}$$

The bounds on the generalization ability is only conceptual, not constructive.

Because the annealed entropy and growth function are just conceptual and cannot be calculated.

The structure of the growth function

$G^{\Lambda}(l) = \ln \sup_{z_1,\cdots,z_l} N^{\Lambda}(z_1,\cdots,z_l)$

Theorem [Vapnik and Chervonenkis, 1968, 1971]

Any growth function either satisfies the equality $G^{\Lambda}(l) = l\ln 2$, or is bounded by the inequality

$$G^{\Lambda}(l) \leq h(\ln\frac{l}{h} + 1)$$

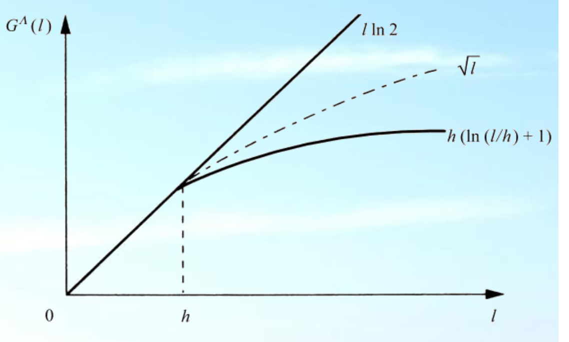
where h is an integer such that when $l = h$,

$$G^{\Lambda}(l) = l\ln 2$$

and

$$G^{\wedge}(h+1) < (h+1)\ln 2.$$

- VC Dimension of the set of indicator functions: infinite or h



$$\frac{H^{\Lambda}(l)}{l} \leq \frac{H^{\Lambda}_{ann}(l)}{l} \leq \frac{G^{\Lambda}(l)}{l} \leq \frac{h(\ln\frac{l}{h}+1)}{l}, \quad (l > h)$$

- Finiteness of the VC dimension is a necessary and sufficient condition for distribution-independent consistency of the ERM learning machines[Vapnik and Chervonenkis, 1974].
- A finite VC dimension implies a fast rate of convergence.

VC Dimension of a Set of Indicator Functions

VC Dimension of a set of indicator functions [Vapnik and Chervonenkis, 1968, 1971]

The VC dimension of a set of indicator functions $Q(z,\alpha),\ \alpha \in \Lambda$ is the maximum number h of vectors $z_1,\cdots,z_h$ that can be separated into two classes in all 2^h possible ways using functions of the set.

(i.e., the max number of vectors that can be *shattered* by the set of functions)

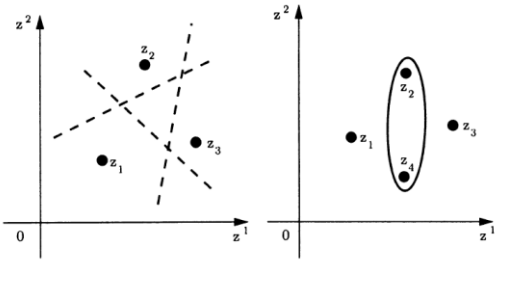


FIGURE 3.3. The VC dimension of the lines in the plane is equal to 3, since they can shatter three vectors, but not four: The vectors z_2, z_4 cannot be separated by a line from the vectors z_1, z_3.

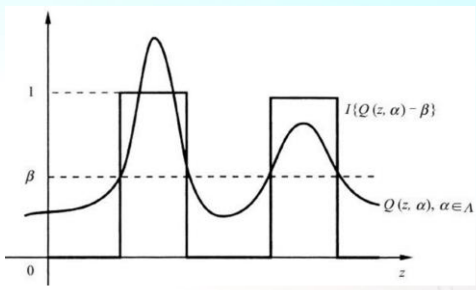
VC Dimension of a Set of Real Functions

VC Dimension of a set of real functions [Vapnik, 1979]

Let $A \leq Q(z,\alpha) \leq B,\ \alpha \in \Lambda$ is a set of real functions bounded by A and B (can be $\pm\infty$). Consider the set of indicators

$$I(z,\alpha,\beta) = \theta\{Q(z,\alpha) - \beta\}, \alpha \in \Lambda, \beta \in (\mathrm{A}, \mathrm{B}).$$

The VC dimension of a set of real functions is defined to be the VC dimension of the set of corresponding indicators.



Examples

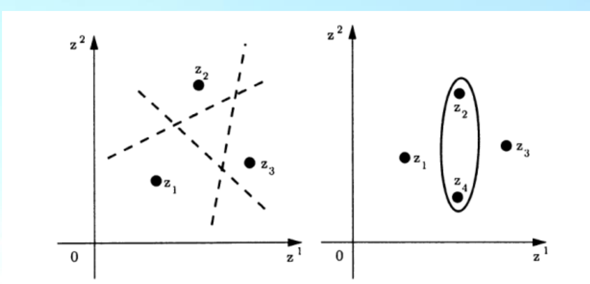


FIGURE 3.3. The VC dimension of the lines in the plane is equal to 3, since they can shatter three vectors, but not four: The vectors z_2, z_4 cannot be separated by a line from the vectors z_1, z_3.

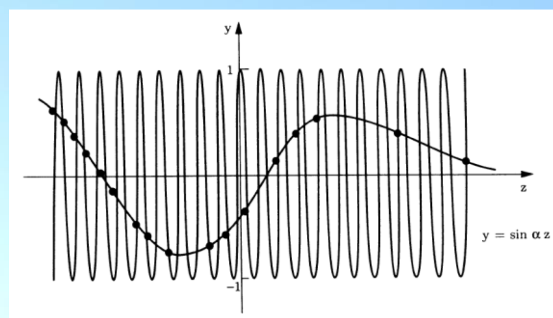


FIGURE 3.4. Using a high-frequency function $sin(\alpha z)$, one can approximate well the value of any function $-1 \leq f(z) \leq 1$ at ℓ appropriately chosen points.

- The set of linear functions in d-dimensional space has VC dimension of d+1.
  - Not true for other functions.
- VC dimension of a set of sin() functions is infinite.

Constructive Distribution-Independent Bounds

$$R_{emp}(\alpha) - \frac{(B-A)}{2}\sqrt{\mathcal{E}} \le R(\alpha) \le R_{emp}(\alpha) + \frac{(B-A)}{2}\sqrt{\mathcal{E}} \qquad \mathcal{E} = 4\frac{G^{\Lambda,B}(2l) - \ln\left(\frac{\eta}{4}\right)}{l}$$

$$G^{\Lambda}(l) \le h\left(\ln\frac{l}{h} + 1\right), l > h$$

$$\mathcal{E} = 4\frac{h\left(\ln\frac{2l}{h} + 1\right) - \ln(\eta/4)}{l}$$

If the set of functions contains a finite number of N elements, then

$$\mathcal{E} = 2\frac{\ln N - \ln\eta}{l}$$

Xuegong Zhang 63

Case 1. The set of totally bounded functions

Let $A \le Q(z,\alpha) \le B, \alpha \in \Lambda$ be a set of totally bounded functions. Then:

(A) The following inequalities hold with probability at least $1-\eta$ simultaneously for all functions of $Q(z,\alpha), \alpha \in \Lambda$ (including the function that minimizes the empirical risk):

$$R_{emp}(\alpha) - \frac{(B-A)}{2}\sqrt{\mathcal{E}} \le R(\alpha) \le R_{emp}(\alpha) + \frac{(B-A)}{2}\sqrt{\mathcal{E}}$$

(B) The following inequality holds with probability at least $1-2\eta$ for the function $Q(z,\alpha_l)$ that minimizes the empirical risk:

$$R(\alpha_l) - \inf_{\alpha\in\Lambda} R(\alpha) \le (B-A)\sqrt{\frac{-\ln\eta}{2l}} + \frac{(B-A)}{2}\sqrt{\mathcal{E}}$$

$$\mathcal{E} = 4\frac{h\left(\ln\frac{2l}{h} + 1\right) - \ln(\eta/4)}{l}$$

Xuegong Zhang 64

Take-home Message:

$$R(\alpha) \le R_{emp}(\alpha) + \frac{(B-A)}{2}\sqrt{\mathcal{E}} \qquad \mathcal{E} = 4\frac{h\left(\ln\frac{2l}{h}+1\right)-\ln(\eta/4)}{l}$$

Specifically, for a set of indicator functions, we have

$$R(\alpha) \le R_{emp}(\alpha) + \sqrt{\left(\frac{h(\ln(2l/h)+1)-\ln(\eta/4)}{l}\right)}$$

or

$$R(\alpha) \le R_{emp}(\alpha) + \Phi(h/l)$$

$\Phi(\cdot)$ is a monotonic function

Note: The bound is only valid for finite VC dimension, and can be lose when VC dimension is very high.

Four parts of Statistical Learning Theory

- What are the (necessary & sufficient) conditions for consistency of an ERM learning process?
- How fast is the rate of convergence of the learning process?
- How can one control the rate of convergence (the generalization ability) of the learning process?
- How can one construct algorithms that can control the generalization ability?

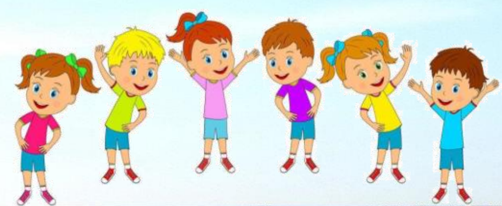
单选题 1分

设置

5-min break

A Yes, I’m back to my computer.

B Sorry, not yet.



提交

清华大学

7.6
Controlling the Generalization Ability of Learning Processes

Note: The theory for controlling the generalization ability of learning machines is devoted to situations with *a small sample* of training instances, say $\frac{l}{h} < 20$.

SRM (Structural Risk Minimization) Inductive Principle

$$R(\alpha) \le R_{emp}(\alpha) + \sqrt{\left(\frac{h(\ln(2l/h)+1) - \ln(\eta/4)}{l}\right)}$$

$$R(w) \le R_{emp}(w) + \Phi(h/l)$$

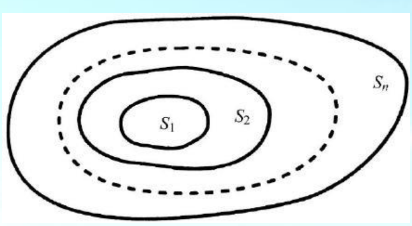
We need to minimize both terms.

Empirical Risk
It depends on a specific function of the set, decided by the learning process.

Confidence Interval
It depends on the VC dimension of the function set, therefore on the design of the machine.

SRM: Structural Risk Minimization

min $R(w) \le R_{emp}(w) + \Phi(h/l)$



$$Q(z,\alpha), \quad \alpha \in \Lambda$$

$$\Downarrow$$

$$S_k = \{Q(z,\alpha), \alpha \in \Lambda_k\}$$
$$S_1 \subset S_2 \subset \cdots \subset S_n \cdots$$
$$h_1 \le h_2 \le \cdots \le h_n \cdots$$

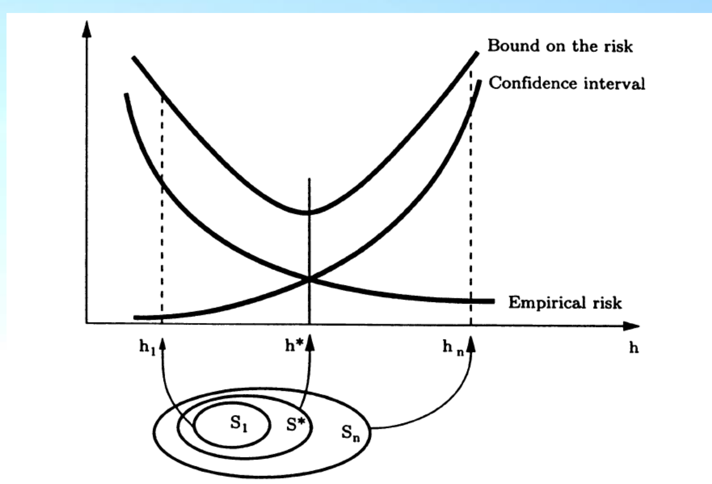
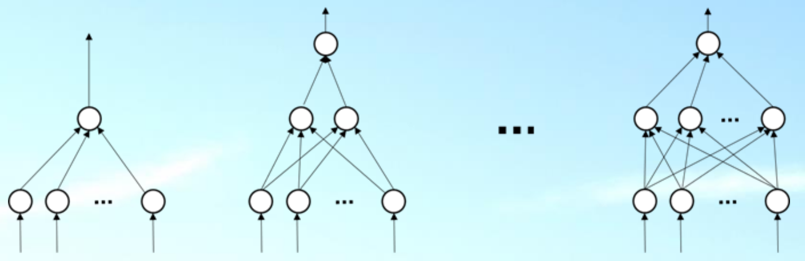


FIGURE 4.2. The bound on the risk is the sum of the empirical risk and the confidence interval. The empirical risk decreases with the index of the element of the structure, while the confidence interval increases. The smallest bound of the risk is achieved on some appropriate element of the structure.

Examples of Structures of Set of Functions



Structure by weight decay in MLP training (regularization)

$$S_p = \{ f(x,w), \|w\| \le C_p \} \qquad C_1 < C_2 < \cdots < C_n$$

$$E(w, \gamma_p) = \frac{1}{l}\sum_{i=1}^{l} L(y_i, f(x_i, w)) + \gamma_p \|w\|^2$$

Regularization

Four parts of Statistical Learning Theory

- What are the (necessary & sufficient) conditions for consistency of an ERM learning process?
- How fast is the rate of convergence of the learning process?
- How can one control the rate of convergence (the generalization ability) of the learning process?
- How can one construct algorithms that can control the generalization ability?

Why largest margin is optimal?

- **Generalization**: the expected performance of a machine on future samples after being trained on limited samples
  - The difference between the expected risk and empirical risk

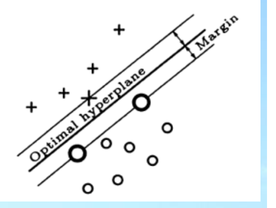
- **Statistical Learning Theory**
  - Large margin
    ➔Low VC dimension
    ➔Low complexity
    ➔High **generalization** ability

$$R(w) \leq R_{emp}(w) + \Phi\left(\frac{h}{l}\right)$$

7.7 Generalization Ability of SVM

VC Dimension of the Optimal Hyperplane

Δ-Margin Separating Hyperplane



We call a hyperplane $(w^* \cdot x) - b = 0, |w^*| = 1$

a Δ-margin separating hyperplane if it classifies vectors as follows:

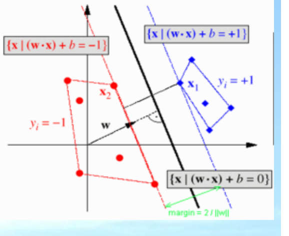
$$y = \begin{cases} 1 & if\ (w^* \cdot x) - b \geq \Delta \\ -1 & if\ (w^* \cdot x) - b \leq -\Delta \end{cases}$$

Theorem

Let d-dimensional vectors $x \in X$ belong to a sphere of radius R. Then the set of Δ-margin separating hyperplanes has VC dimension h bounded by the inequality

$$h \leq \min\left(\left[\frac{R^2}{\Delta^2}\right], d\right) + 1$$

VC dimension of a SVM



The set of canonical separating hyperplane in d-dimensional space with $\|w\| \leq A$ has the VC dimension bound of

$$h \leq \min([R^2 A^2], d) + 1$$

where R is the radius of a sphere containing the data.

- VC dimension can be much smaller after controlling the margin.

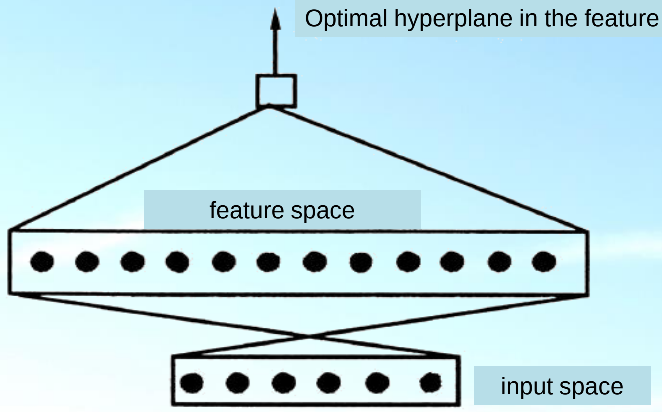
N+1

$$R(w) \leq R_{\text{emp}}(w) + \Phi(h/l)$$

Empirical Risk
It depends on a specific function of the set, decided by the learning process.

Confidence Interval
It depends on the VC dimension of the function set, therefore on the design of the machine.

Recap:



Idea:

- Mapping the original input space into a new feature space by nonlinear transformation, and finding the optimal hyperplane in the feature space

What problems we may meet with this approach?

(1) **Can we handle the computation in feature space?**

(technical issue)

- e.g, a 4th or 5th order polynomial function on 200d data (like 16x16 pixels) need feature spaces of ~billion dimension.

(curse of dimensionality)

(2) Will the "optimal hyperplane" still optimal or good in the feature space?

(conceptual issue)

- Can it still work or can it generalize?

Yes, it can, because we control the margin.

Controlling generalization in high dimensional spaces

$R(w) \leq R_{emp}(w) + \Phi\left(\frac{h}{l}\right),\ \ h \leq \min([R^2 A^2], d) + 1$

Theorem. If training sets containing l examples are separated by the maximal margin hyperplanes, then the expectation (over training sets) of the probability of test error is bounded as

$$\mathrm{E}\, P_{error} \leq \mathrm{E} \min\left(\frac{m}{l}, \frac{[R^2|w|^2]}{l}, \frac{d}{l}\right)$$

where m is the number of support vectors, R is the radius of a sphere containing the data and $|w|^{-2}$ is the value of the margin, and d is the dimensionality of the input space.

Controlling generalization in high dimensional spaces

$$\mathrm{E}\, P_{error} \leq \mathrm{E} \min\left(\frac{m}{l}, \frac{[R^2|w|^2]}{l}, \frac{d}{l}\right)$$

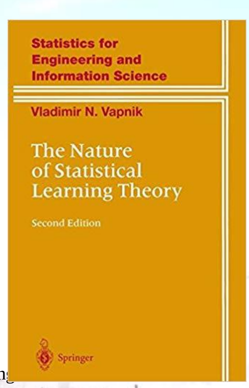
Three reasons why optimal hyperplanes can generalize:

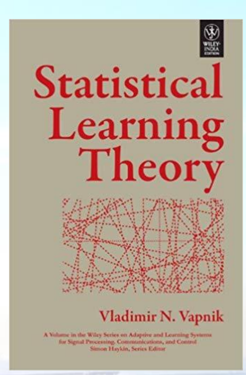
- Because the expectation of the data compression is large
- Because the expectation of the margin is large
- Because the input space is small and sample size is large

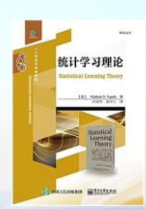
Discussions

- Key features of SVM
  - Large margin kernel machines
  - Good for high-dimensional small-sample problems
  - Open questions
    - Choice of kernels
    - Computation cost when sample size is big
    - …
- Discussions on SLT
  - Vapnik: "*Nothing is more practical than a good theory*."
  - Worst-case analysis
  - Small sample scenarios
  - Open questions
    - Estimation of VC dimensions
    - Admissible structures of a set of functions
    - Expanding the theory to big-data scenarios
    - …

Extensive self-study needed if you find yourself attracted by the theory.







7.8
Ill-posed Problems and Regularization

Ill-posed problems

- Inverse problem

$$Az = u$$

  - Using observed $(\boldsymbol{z}, \boldsymbol{u})$ to solve for A^{-1}, to predict $\boldsymbol{z}$ from $\boldsymbol{u}$ in the future
  - Well-posed: if the solution exist, unique and stable
  - Ill-posed: if small change in $\tilde{\boldsymbol{u}}$ cause big changes in $\boldsymbol{z} = A^{-1}\tilde{\boldsymbol{u}}$

Regularization for Ill-posed Problems

- Ill-posed problem

$$Af = F, \qquad f \in \mathcal{F}$$

  – Small noise $\|F - F_\delta\| < \delta$ causing large change in solution
  – Minimizing $R(f) = \|Af - F_\delta\|^2$ cannot produce reasonable solution
- Regularization
  – Change the objective function to regularized functional

$$R^*(f) = \|Af - F_\delta\| + \lambda(\delta)\Omega(f)$$

where $\Omega(f)$ is a functional measuring certain property of solution f

Recall: Multicategory SVM

- Yoonkyung Lee, Yi Lin & Grace Wahba, Multicategory Support Vector Machines, *Technical Report No. 1043, Dept of Statistics, Univ of Wisconsin, Madison*, Sept. 29, 2001
  – slides adopted from Lee's presentations

SVM in Regularization Framework

Wahba (1998)

- $f(\mathbf{x}) = h(\mathbf{x}) + b$ with $h \in \mathcal{H}_K$ a Reproducing Kernel Hilbert space (RKHS) with reproducing kernel K.
- Classification rule : $\phi(\mathbf{x}) = sign(f(\mathbf{x}))$
- SVM scheme : to find f minimizing

Regularization term

$$\frac{1}{n}\sum_{i=1}^{n}(1 - y_i f(\mathbf{x}_i))_+ + \lambda\|h\|^2_{\mathcal{H}_K}$$

λ is a tuning parameter which balances the data fit and the complexity of $f(\mathbf{x})$.

Another view of SRM: ERM + Regularization

$$\min \quad \psi(\mathbf{w}, \xi) = \frac{1}{2}(\mathbf{w} \cdot \mathbf{w}) + C\left(\sum_{i=1}^{n} \xi_i\right)$$

$$\min_{f} \sum_{i=1}^{n} V(f(x_i), y_i) + \lambda R(f)$$

Regularization Methods

- L_0 Regularization

$$\min_{\beta} \quad \frac{1}{l}\sum_{i=1}^{l} V(y_j, \boldsymbol{\beta}^T \boldsymbol{x}_j) + \lambda \|\boldsymbol{\beta}\|_0$$

- L_1 Regularization, Lasso or Basis Pursuit

$$\min_{\beta} \quad \frac{1}{l}\sum_{i=1}^{l} \left(y_j - \boldsymbol{\beta}^T \boldsymbol{x}_j\right)^2 + \lambda \|\boldsymbol{\beta}\|_1$$

- L_2 Regularization: Tikhonov Regularization

$$\min_{\beta} \quad \frac{1}{l}\sum_{i=1}^{l} V(y_j, \boldsymbol{\beta}^T \boldsymbol{x}_j) + \lambda \|\boldsymbol{\beta}\|^2$$

Regularization Methods

- L_q Regularization

$$\min_{\beta} \frac{1}{l}\sum_{i=1}^{l} V(y_j, \boldsymbol{\beta}^T \boldsymbol{x}_j) + \lambda \sum_j \left|\beta_j^q\right|^{\frac{1}{q}}$$

- Elastic Net

$$\min_{\beta} \frac{1}{l}\sum_{i=1}^{l} \left(y_j - \boldsymbol{\beta}^T \boldsymbol{x}_j\right)^2 + \lambda(\alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\|\boldsymbol{\beta}\|^2)$$

R

W

Readings

R

- Vapnik, *The Nature of Statistical Learning Theory*
  - Chapters 1-4, pp.17-122
  - Chapter 8, pp.291-299
- Shai Shalev-Shwartz & Shai Ben-David, *Understanding Machine Learning: From Theory to Algorithms* (http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning)
  - Chapter 6. The VC-Dimension, pp.67-82 (43-57 on 5th print 2016)
  - Chapter 13. Regularization and Stability, pp. 171-183 (137-149)
- Abu-Mostafa, Magdon-Ismail, Lin, *Learning from Data* (http://amlbook.com/; http://work.caltech.edu/telecourse)
  - Chapter 1, The Learning Problem, pp. 15-33
  - Chapter 2, Training versus Testing, pp. 39-69

No homework for this chapter.


See you next week
for
Feature Selection and Extraction
Xuegong Zhang
91