01/27/2017 CS group meeting

The nature of Statistical Learning Theory V. N. Vapnik 1st edition

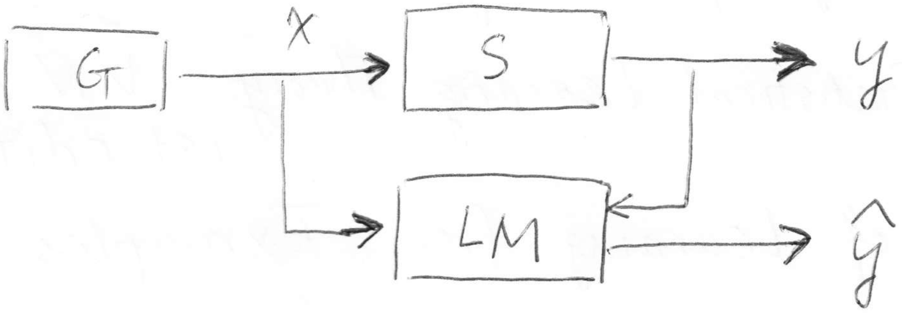
General Model of Learning from Examples

(i). Generator (G) of random vectors $x \in \mathbb{R}^n$ drawn independently from a fixed but unknown probability distribution function [redacted] $F(x)$.

(ii). 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

(iii) A learning machine (LM) capable of implementing a set of functions $f(x, \alpha)$ $\alpha \in \Lambda$ where Λ is a set of parameters.

Problem of learning Choosing from the given set of functions $f(x, \alpha)$, $\alpha \in \Lambda$, the one which best approximates the supervisor's response.



Selection of the desired function is based on a training set of l independent and identically distributed observations drawn according to $F(x, y) = F(x)\, F(y|x)$: $(x_1, y_1), \ldots, (x_l, y_l)$

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

Goal: Find the function $f(x, \alpha_0)$ which minimize the risk functional $R(\alpha)$ over $\alpha \in \Lambda$.

where the joint probability distribution function $F(x,y)$ is unknown and the only available information is contained in the training set $(x_1, y_1), \ldots, (x_l, y_l)$

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

"pattern recognition"

$R(\alpha)$ — classification error

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

$f(x,\alpha)$ $\alpha \in \Lambda$ contains the regression function

$f(x,\alpha_0) = \int y \, dF(y|x)$

"Regression Estimation": Minimizing $R(\alpha)$ in the situation where $F(x,y)$ is unknown but training set is given

(3). Density Estimation $p(x,\alpha)$ $\alpha \in \Lambda$

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

<u>General Setting of the Learning Problem</u>

We defined $F(z)$ as a probability measure on a set of functions $Q(z,\alpha)$ $\alpha \in \Lambda$

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

where $F(z)$ is unknown but an i.i.d. sample $z_1, \dots, z_\ell$ is given.

Empirical Risk Minimization (ERM) inductive principle

Empirical Risk Functional $R_{emp}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha)$

One approximates the function $Q(z, \alpha_0)$ which minimizes the risk $R(\alpha)$ by the function $Q(z, \alpha_\ell)$ which minimizes $R_{emp}(\alpha)$

We say that an inductive principle defines a <u>learning process</u> if for any given set of observations the learning machine chooses the approximation using this inductive principle. , "consistency"

▷ Least Squares Regression $R_{emp}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - f(x_i, \alpha))^2$

▷ ML method $R_{emp}(\alpha) = -\frac{1}{\ell} \sum_{i=1}^{\ell} \ln p(x_i, \alpha)$

Learning Theory addresses:

(i). What are (necessary and sufficient) conditions for consistency of a learning process based on the ERM principle?

(ii). How fast is the rate of convergence of the learning process?

(iii). How can one control the rate of convergence (the generalization ability) of the learning process In other words, this problem is devoted to constructing an inductive principle for minimizing the risk functional using a small sample of training instances.

(iv). How can one construct algorithms that can control the generalization ability?

$Q(z, \alpha_\ell)$ minimizes $R_{emp} = \frac{1}{\ell}\sum_{i=1}^{\ell} Q(z_i, \alpha)$

where $z_1 \ldots z_\ell$ is a given i.i.d. sequence of observations

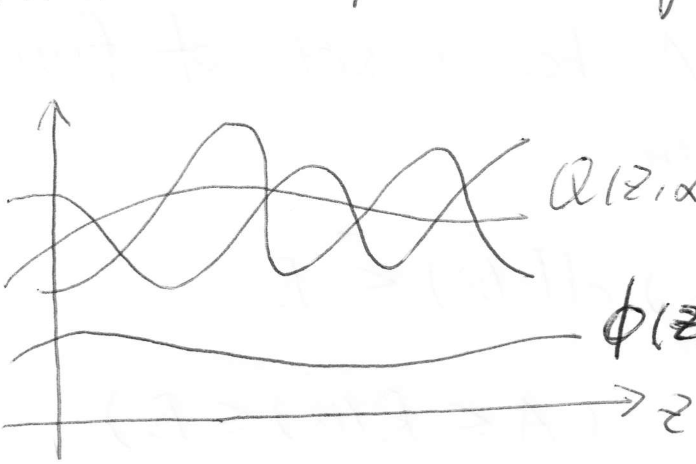
<u>Definition</u> We say that the principle (method) of ERM is consistent for the set of functions $Q(z, \alpha)$ $\alpha \in \Lambda$ and for the probability distribution function $F(z)$ if the following two sequences converge in probability to the same limit

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

$$R_{emp}(\alpha_\ell) \xrightarrow[\ell \to \infty]{P} \inf_{\alpha \in \Lambda} R(\alpha).$$

What are the conditions of consistency for the ERM method? These conditions are obtained in terms of general characteristics of the set of functions and the probability measure.

"Trivial Case of Consistency"



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

<u>Definition</u> We say that the ERM method is non-trivially consistent for the set of functions $Q(z,\alpha)$, $\alpha\in\Lambda$ and the probability distribution function $F(z)$ if for any non-empty 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

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

is valid.

Key Theorem of Learning Theory

Let $Q(z,\alpha)$, $\alpha \in \Lambda$ be a set of functions that satisfy the condition

$$A \leq \int Q(z,\alpha)\, dF(z) \leq B$$

$$(A \leq R(\alpha) \leq B)$$

Then for the ERM principle to be consistent it is necessary and sufficient that the empirical risk $R_{emp}(\alpha)$ converge uniformly to the actual risk $R(\alpha)$ over the set $Q(z,\alpha)$ $\alpha \in \Lambda$ in the following sense

$$\lim_{l \to \infty} \mathbb{P}\left\{ \sup_{\alpha \in \Lambda}\left(R(\alpha) - R_{emp}(\alpha)\right) > \varepsilon \right\} = 0$$

$$\forall \varepsilon > 0$$

Consistency of ERM principle $\Longleftrightarrow$ Existence of uniform one-sided convergence.