Setting of the Learning Problem

Elvis Cui

PhD in Biostatistics, UCLA

January 7, 2020

Overview

Book chapters, papers and monographs

- V. N. Vapnik. The Nature of Statistical Learning Theory. Springer. 1996.

Function estimation model

A general model of learning from data:

- A generator G: generate $x \in \mathbb{R}^n$, $x \sim F(x)$ unknown.
- A supervisor S: Get y given x from $F(y|x)$ fixed but unknown.
- A learning machine LM: $\mathcal{F} = \{f(x, \alpha) : x \in \mathbb{R}^n, \alpha \in \Lambda\}$.
- Training set: $(x_1, y_1), \cdots, (x_l, y_l) \sim_{iid} F(x, y) = F(x)F(y|x)$.

The problem of risk minimization

Definition 1 (Risk Functional)

Given a loss function $L : \mathcal{Y} \times \mathcal{F} \to \mathbb{R}^*$, the risk functional (indexed by α) is defined as

$$R(\alpha) = \int_{\mathcal{X}\times\mathcal{Y}} L(y, f(x,\alpha))dF(x,y)$$

Our goal is to find

$$\alpha_0 = \arg\max_{\alpha\in\Lambda} R(\alpha)$$

Three main learning problems

- Pattern recognition: $y \in \{0,1\}, f(x,\alpha)$ is an indicator.

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

  Then $R(\alpha) = \mathbb{P}(y \neq f(x,\alpha))$.

- Regression estimation: Let $f(x,\alpha_0) = \int y dF(y|x)$ and

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

- Density estimation (Fisher-Wald setting): Suppose $f(x,\alpha)$'s are densities w.r.t. some measure. Take

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

  An empirical version gives us **maximum likelihood estimation**.

The general setting of the learning problem

Suppose $F(z)$, a probability measure, is defined on a space Z. Consider a family of functions $Q(z, \alpha), \alpha \in \Lambda$. Our goal is to minimize the risk functional

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

$F(Z)$ is unknown but we are given iid samples

$$z_1, \cdots, z_l.$$

The learning problems considered above are particular cases of minimizing risk functional.

The empirical risk minimization (ERM) inductive principle

Definition 2 (Empirical risk functional)

Since $F(Z)$ is unknown, we use the empirical version of risk functional:

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

Then minimizing it we get $Q(z, \alpha_l)$, an approximation of $Q(z, \alpha_0)$. This is called empirical risk minimization inductive principle.

The Classical Paradigm of Solving Learning Problems

Density estimation (maximum likelihood): Given

$$x_1, \cdots, x_l$$

In 1920s, R. A. Fisher suggested minimizing functional

$$L(\alpha) = \sum_{i=1}^{l} \ln p(x_i, \alpha)$$

to estimate α. This is called maximum likelihood. Under some conditions, MLE is consistent.

The Classical Paradigm of Solving Learning Problems

Pattern recognition (discriminant analysis): Bayes' decision rule is optimal.

$$f(x) = sign\{\ln p_1(x, \alpha^*) - \ln p_2(x, \beta^*) + \ln \frac{q_1}{1 - q_1}\}$$

We use ML method to estimate $p_1(x, \alpha^*)$ and $p_2(x, \beta^*)$.

The Classical Paradigm of Solving Learning Problems

Regression estimation model: $f_0(x) = f(x, \alpha_0), \alpha_0 \in \Lambda$ and also

$$y_i = f(x_i, \alpha_0) + \epsilon_i, \epsilon \perp x_i, \epsilon_i \sim p(\epsilon)$$

Then we use ML method to estimate α_0:

$$L(\alpha) = \sum_{i=1}^{l} \ln p(y_i - f(x_i, \alpha))$$

If we take $p(\cdot)$ to be Gaussian, then we get **least square estimation**.

A simple example where MLE fails

Suppose

$$p(x,\alpha,\sigma) = \frac{1}{2\sigma\sqrt{2\pi}}\exp\{-\frac{(x-\alpha)^2}{2\sigma^2}\} + \frac{1}{2\sqrt{2\pi}}\exp\{-\frac{x^2}{2}\}$$

Given $x_1, \cdots, x_l$, we have

$$\begin{aligned}
L(\alpha = x_1, \sigma_0) &= \sum_{i=1}^{l} \ln p(x_i; \alpha = x_1, \sigma_0) \\
&> \ln\left(\frac{1}{2\sigma_0\sqrt{2\pi}}\right) + \sum_{i=2}^{l} \ln\left(\frac{1}{2\sqrt{2\pi}}\exp\{-\frac{x_i^2}{2}\}\right) \\
&\to \infty \text{ as } \sigma_0 \to 0
\end{aligned}$$

Nonparametric methods of density estimation

(M. Rosenblatt, 1956; Parzen, 1962; Chentsov, 1963).

- Parzen 's Windows:

$$K(x, x_i, \gamma) = \frac{1}{\gamma^n} K(\frac{x - x_i}{\gamma}), x \in \mathbb{R}^n$$

$$p(x) = \frac{1}{l} \sum_{i=1}^{l} K(x, x_i, \gamma)$$

- Glivenko-Cantelli:

$$\sup_x |F(x) - F_l(x)| \to^{a.s.}_{l \to \infty} 0$$

where $F_l(x) = \frac{1}{l} \sum_{i=1}^{l} I(x \geq x_i)$.

Solving Problems Using a Restricted Amount of Information

- When solving a given problem, try to avoid solving a more general problem as an intermediate step.

Model Minimization of the Risk Based on Empirical Data

- Regression estimation: $y = f_0(x) + \epsilon$.

$$
\begin{aligned}
R(\alpha) &= \int (y - f(x,\alpha))^2 dF(x,y) \\
&= \int (f(x,\alpha) - f_0(x))^2 dF(x) + \int (y - f_0(x))^2 dF(x,y)
\end{aligned}
$$

The first term is $L_2(F)$ distance and the second term does not involve x. So we do not need to find the joint distribution $F(x, y)$.

Model Minimization of the Risk Based on Empirical Data

- Density estimation:

$$R(\alpha) = -\int \ln p(t,\alpha) dF(t) = -\int p_0(t) \ln p(t,\alpha) dt$$

If we add a constant

$$c = \int \ln p_0(t) dF(t)$$

We get Kullback-Leibler distance

$$R^*(\alpha) = -\int p_0(t) \ln \frac{p(t,\alpha)}{p_0(t)} dt$$

Stochastic Approximation Inference

(Robbins and Monroe, 1951) Given iid data $z_1, \cdots, z_l$, we minimize functional w.r.t. α:

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

One uses the following iterative procedure:

$$\alpha_{k+1} = \alpha_k - \gamma_k \text{grad}_\alpha Q(z_k, \alpha_k), k = 1, 2, \cdots, l$$

Such method is consistent under some general conditions.

Next time:
Support Vector Machine !