An Overview of Statistical Learning Theory

Diego Garrido

1 Function Estimation Model

The autor describe that a model of learning from examples can be conducted in the general statistical framework of minimizing expected loss given data.

The model of learning from examples considers:

1. A unknown distribution $P(x)$ called the generator, which draws independent random vectors x.
2. A unknown conditional distribution $P(y|x)$ called the supervisor, which draw an output vector y given input vector x.
3. A learning machine capable of implementing a set of functions $f(x, \alpha), \alpha \in \Lambda$.

The problem is choose the function from the set of function $f(x, \alpha), \alpha \in \Lambda$ with best performance, i.e, the one which predicts the supervisor's responese in the best possible way. This selection is based on a training set of l random idepedent identically distributed (i.i.d.) observations drawn according to the join probability $P(x, y) = P(x)P(y|x)$.

$$(x_1, y_1), \dots, (x_l, y_l) \tag{1}$$

2 Risk Minimization

To choose the best function from the set of function $f(x, \alpha), \alpha \in \Lambda$, one measures the loss $L(y, f(x, \alpha))$ between the output y given the input x and $f(x, \alpha)$. Then one can consider the following risk functional:

$$R(\alpha) = \int L(y, f(x, \alpha)) dP(x, y),\ \alpha \in \Lambda \tag{2}$$

The functional $R(\alpha)$ measures the expected loss between the response y given input x an the response $f(x, \alpha)$ given α. The goal is to find the functi贸n $f(x, \alpha_0)$ which minimizes the risk functional $R(\alpha), \alpha \in \Lambda$, i.e, over the set of functions $f(x, \alpha), \alpha \in \Lambda$, in the situation where the join probability distributi贸n $P(x, y)$ is unknown, but a training set is given (1).

3 The Problem of Pattern Recognition

Let the supervisor's output take on only two values $y \in \{0, 1\}$ and let $f(x, \alpha), \alpha \in \Lambda$, be a set of indicator functions (functions which take on only two values zero and one). Consider the following loss function:

$$L(y, f(x, \alpha)) = \begin{cases} 0 \text{ if } y = f(x, \alpha) \\ 1 \text{ if } y \neq f(x, \alpha) \end{cases} \tag{3}$$

So the expected loss can be written as

$$\mathbb{E}(L(x, f(x, \alpha))) = \mathbb{E}(\mathbb{I}_{\{y \neq f(x,\alpha)\}}) = P(y \neq f(x, \alpha)) \tag{4}$$

then $R(\alpha) = P(y \neq f(x, \alpha))$, that means the risk functional (2) provides the probability of classification error, i.e, when the answer y given by supervisor an the answers given by indicator function $f(x, \alpha)$ differ.

4 Empirical risk

The general setting of the learning problem can be described as follows. Let the probability measure $P(z)$ 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

$$R(\alpha) = \int Q(z, \alpha) dP(z),\ \alpha \in \Lambda \tag{5}$$

if probability measure $P(z)$ is unknown but an i.i.d. sample is given

$$z_1, \dots, z_l \tag{6}$$

The empirical risk functional replace the expected risk functinoal $R(\alpha)$ by

$$R_{emp}(\alpha) = \frac{1}{l}\sum_{i=1}^{l} Q(z,\alpha) \tag{7}$$

based in the training set (6).

According to law of large numbers a collection of $z_1, z_2, \ldots, z_l$ i.i.d. samples, the arithmetic mean converges to the expected value when $l \to \infty$ and the approximation is better the larger the data set. The empirical risk functional correspond to the Montel Carlo approximation of the expected value of $Q(z,\alpha)$ using the empirical distribution of $\{Q(z_i,\alpha)\}_{i=1}^{l}$. It is a good approximation because it is based on the idea of law of large numbers, where the sequence $Q(z_1,\alpha), \ldots, Q(z_l,\alpha)$ is a sequence of i.i.d. random variables, such that the sample mean of this sequence converges in probability to $\mathbb{E}(Q(z,\alpha))$.

5 Density estimation

To estimate a density function from a given set of functions $p(x,\alpha), \alpha \in \Lambda$ one can use the loss function $L(p(x,\alpha)) = -ln\ p(x,\alpha)$. Putting this loss into (7) one obtains

$$R_{emp}(\alpha) = -\frac{1}{l}\sum_{i=1}^{l} ln\ p(x_i,\alpha) \tag{8}$$

minimizing this functional is equivalent to minimizing the negative log-likelihood, which is equivalent to the maximum likelihood method.

6 The four parts of learning theory

The four parts of learning theory are:

1. The theory of consistency of learning processes. This is related to the necessary and sufficient conditions for convergence in probability of the values of risk $R(\alpha_l)$ and the empirical risks $R_{emp}(\alpha_l)$ to the minimal possible value of the risk $R(\alpha_0)$.
2. The nonasymptotic theory of the rate of convergence of learning processes. This refer to how fast the sequence of smallest empirical risk values converge to the smallest actual risk.
3. The theory of controlling the generalization of learning processes. This is related to control the rate of generalization of the learning machine.
4. The theory of constructing learning algorithms. This topic attempt to build algorithms that can control the rate of generalzation.