Statistical Learning Theory

Fundamentals

Miguel A. Veganzones

Grupo Inteligencia Computacional
Universidad del País Vasco

Outline

Approaches to the learning problem

- Learning problem: the problem of choosing the desired dependence on the basis of empirical data.
- Two approaches:
  1. To choose an approximating function from a given set of functions.
  2. To estimate the desired stochastic dependences (densities, conditional densities, conditional probabilites).

First approach

To choose an aproximating function from a set of functions

- It's a problem rather general.
- Three subproblems: pattern recognition, regression estimation, density estimation.
- Based on the idea that the quality of the chosen function can be evaluated by a risk functional.
- Problem of minimizing the risk functional on the basis of empirical data.

Second approach

To estimate the desired stochastic dependences

- Using estimated stochastic dependence, the pattern recognition, regression and density estimation problems can be solved as well.
- Requires solution of integral equations for determining these dependences when some elements of the equation are unknown.
- It gives much more details but it's an ill-posed problem.

General problem of learning from examples

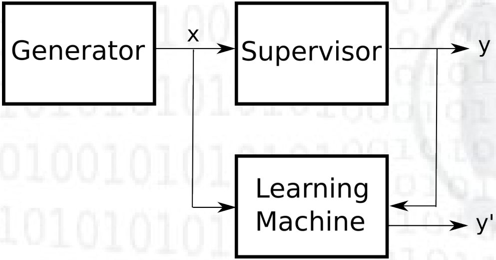

Figure: A model of learning by examples.

Generator

- The Generator G determines the environment in which the supervisor and the learning machine act.
- Simplest environment: G generates the vectors $\mathbf{x} \in \mathbf{X}$ independently and identically distributed (i.i.d.) according to some unknown but fixed probability distribution function $F(\mathbf{x})$.

Supervisor

- The target operator (supervisor) S transforms the input vectors $\mathbf{x}$ into the output values y.
- Supposition: S returns the output y on the vector $\mathbf{x}$ according to a conditional distribution function $F(y|\mathbf{x})$, which includes the case when the supervisor uses some function $y = f(\mathbf{x})$.

Learning Machine

- The Learning Machine LM observes the l pairs $(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_l}, y_l)$, the training set, which is drawn randonmly and independently according to a joint distribution function $F(\mathbf{x}, y) = F(y|\mathbf{x})F(\mathbf{x})$.
- Using the training set, LM constructs some operator which will be used for prediction of the supervisor's answer y_i on any specific vector $\mathbf{x_i}$ generated by G.

Two different Goals

1. To *imitate* the supervisor's operator: try to construct an operator which provides for a given G, the best predictions to the supervisor's outputs.
2. To *identify* the supervisor's operator: try to construct an operator which is close to the supervisor's operator.

- Both problems are based on the same general principles.
- The learning process is a process of choosing an appropiate function from a given set of functions.

Functional

- Among the totally of possible functions, one looks for the one that satisfies the given quality criterion in the best possible manner.
- Formally: on the subset Z of the vector space $\Re^n$, a set of admissible functions $\{g(\mathbf{z})\}$, $\mathbf{z} \in Z$, is given, and a functional $R = R(g(\mathbf{z}))$ is defined.
- It's required to find the function $g'(\mathbf{z})$ from the set $\{g(\mathbf{z})\}$ which minimizes the funcional $R(g(\mathbf{z}))$.

Two cases

1. When the set of functions $\{g(\mathbf{z})\}$ and the functional $R(g(\mathbf{z}))$ are explicitly given: calculus of variations.
2. When a probability distribution function $F(\mathbf{z})$ is defined on Z and the functional is defined as the mathematical expectation

$$R(g(\mathbf{z})) = \int L(\mathbf{z}, g(\mathbf{z}))dF(z) \quad (1)$$

- where function $L(\mathbf{z}, g(\mathbf{z}))$ is integrable for any $g(\mathbf{z}) \in \{g(\mathbf{z})\}$.
- The problem is then, to minimize (1) when $F(\mathbf{z})$ is unknown but the sample $\mathbf{z_1}, \ldots, \mathbf{z_n}$ of observations is available.

Problem definition

1. Imitation problem: how can we obtain the minimum of the functional in the given set of functions?
2. Identification problem: what should be minimized in order to select from the set $\{g(\mathbf{z})\}$ a function which will guarantee that the functional (1) is small?

- The minimization of the functional (1) on the basis of empirical data $\mathbf{z_1}, \ldots, \mathbf{z_n}$ is one of the main problems of mathematical statistics.

Parametrization

- The set of functions $\{g(\mathbf{z})\}$ will be given in a parametric form $\{g(\mathbf{z},\alpha), \alpha \in \Lambda\}$.
- The study of only parametric functions is not a restriction on the problem, since the set Λ is arbitrary: a set of scalar quantities, a set of vectors, or a set of abstract elements.
- The functional (1) can be rewritten as

$$R(\alpha) = \int Q(\mathbf{z},\alpha)\partial dF(\mathbf{z}), \qquad \alpha \in \Lambda \tag{2}$$

  - where $Q(\mathbf{z},\alpha) = L(\mathbf{z}, g(\mathbf{z},\alpha))$ is called the *loss function*.

The expected loss

- It is assumed that each function $Q(\mathbf{z}, \alpha^*)$ determines the ammount of the loss resulting from the realization of the vector $\mathbf{z}$ for a fixed $\alpha = \alpha^*$.
- The expected loss with respect to $\mathbf{z}$ for the function $Q(\mathbf{z}, \alpha^*)$ is determined by the integral

$$R(\alpha^*) = \int Q(\mathbf{z}, \alpha^*) dF(\mathbf{z}) \quad (3)$$

  - which is called the *risk functional* or the *risk*.

Problem redefinition

The redefined Learning Problem

The problem is to choose in the set $Q(\mathbf{z}, \alpha)$, $\alpha \in \Lambda$, a function $Q(\mathbf{z}, \alpha_0)$ which minimizes the risk when the probability distribution function is unknown but random independent observations $\mathbf{z_1}, \ldots, \mathbf{z_n}$ are given.

Outline

Informal definition

Definition

A supervisor observes ocurring situations and determines to which of k classes each one of them belongs. It is required to construct a machine which, after observing the supervisor's classification, carries out the classification approximately in the same manner as the supervisor.

Formal definition

Definition

In a certain environment characterized by a p.d.f. $F(\mathbf{x})$, situation $\mathbf{x}$ appears randomly and independently. The supervisor classifies each situations into one of k classes. We assume that the supervisor carries out this classification by $F(\omega|\mathbf{x})$, where $\omega \in \{0, 1, \ldots, k-1\}$ ($w = p$ indicates that the supervisor assigns situation $\mathbf{x}$ to the class number p).

- Neither, $F(\mathbf{x})$ nor $F(\omega|\mathbf{x})$ are known, but they exist.
- Thus, a joint distribution $F(\omega, \mathbf{x}) = F(\omega|\mathbf{x})F(\mathbf{x})$ exists.

Loss Function

- Let a set of functions $\phi(\mathbf{x}, \alpha)$, $\alpha \in \Lambda$, which take only k values $\{0, 1, \ldots, k-1\}$ (a set of decision rules), be given.
- We shall consider the simplest loss function:

$$L(\omega, \phi) = \begin{cases} 0 & if\, \omega = \phi \\ 1 & if\, \omega \neq \phi \end{cases} \tag{4}$$

- The problem of pattern recognition is to minimize the functional

$$R(\alpha) = \int L(\omega, \phi(\mathbf{x}, \alpha)) dF(\omega, \mathbf{x})$$

on the set of functions $\phi(\mathbf{x}, \alpha)$, $\alpha \in \Lambda$, where the p.d.f. $F(\omega, \mathbf{x})$ is unknown but a random independent sample of pairs $(\omega_1, \mathbf{x_1}), \ldots, (\omega_l, \mathbf{x_l})$ is given.

Restrictions

- The problem of pattern recognition has been reduced to the problem of minimizing the risk on the basis of empirical data, where the set of loss functions $Q(\mathbf{z},\alpha)$, $\alpha \in \Lambda$, is not arbitrary as in the general case. The following restrictions are imposed:
  1. The vector $\mathbf{z}$ consists of $n+1$ coordinates: coordinate ω (which takes a finite number of values) and n coordinates $x^1,\ldots,x^n$ which form the vector $\mathbf{x}$.
  2. The set of functions $Q(\mathbf{z},\alpha)$, $\alpha \in \Lambda$ is given by

$$Q(\mathbf{z},\alpha) = L(\omega,\phi(\mathbf{x},\alpha)), \qquad \alpha \in \Lambda$$

     and also takes on only a finite number of values.

Outline

Stochastic dependences

- There exist relationships (stochastic dependences) where to each vector $\mathbf{x}$ there corresponds a number y which we obtain as a result of random trials. $F(y|\mathbf{x})$ expresses the stochastic relationship between y and $\mathbf{x}$.
- Estimatimg the stochastic dependence based on the empirical data $(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_l}, y_l)$ is a quite difficult problem (ill-posed problem).
- However, the knowledge of $F(y|\mathbf{x})$ is often not required and it's sufficient to determine one of it characteristics.

Regression

- The function of conditional mathematical expectaction

$$r(\mathbf{x}) = \int y dF(y|\mathbf{x}) \tag{5}$$

  is called the *regression*.
- Estimate the regression in the set of functions $f(\mathbf{x}, \alpha)$, $\alpha \in \Lambda$, is referred to as the problem of regression estimation.

Conditions

- The problem of regression estimation is reduced to the model of minimizing risk based on empirical data under the following conditions:

$$\int y^2 dF(y,\mathbf{x}) < \infty \qquad \int r^2(\mathbf{x}) dF(y,\mathbf{x}) < \infty$$

- On the set $f(\mathbf{x},\alpha)$, $\alpha \in \Lambda$ ($f(\mathbf{x},\alpha) \in L_2(P)$), the minimum (if exists) of the functional

$$R(\alpha) = \int (y - f(\mathbf{x},\alpha))^2 dF(y,\mathbf{x})$$

is attained at:

  - The regression function if $r(\mathbf{x}) \in f(\mathbf{x},\alpha)$, $\alpha \in \Lambda$.
  - To the function $f(\mathbf{x},\alpha^*)$ which is the closest to $r(\mathbf{x})$ in the metric $L_2(P)$ if $r(\mathbf{x}) \notin f(\mathbf{x},\alpha)$, $\alpha \in \Lambda$.

Demostration

- Denote $\Delta f(\mathbf{x},\alpha) = f(\mathbf{x},\alpha) - r(\mathbf{x})$.
- The functional can be rewritten as:

$$R(\alpha) = \int (y - r(\mathbf{x}))^2 dF(y,\mathbf{x}) + \int (\Delta f(\mathbf{x},\alpha))^2 dF(y,\mathbf{x})$$

$$-2\int \Delta f(\mathbf{x},\alpha)(y - r(\mathbf{x}))dF(y,\mathbf{x})$$

  - $\int (y - r(\mathbf{x}))^2 dF(y,\mathbf{x})$ does not depend on α.
  - $\int \Delta f(\mathbf{x},\alpha)(y - r(\mathbf{x}))dF(y,\mathbf{x}) =$
    $\int \Delta f(\mathbf{x},\alpha)\left[(y - r(\mathbf{x}))dF(y|\mathbf{x})\right] dF(\mathbf{x}) = 0$.

Restrictions

- The problem of estimating the regression may also be reduced to the scheme of minimizing the risk. The following restrictions are imposed:
  1. The vector $\mathbf{z}$ consists of $n+1$ coordinates: the coordinate y and n coordinates $x^1, \ldots, x^n$ which form the vector $\mathbf{x}$. However, the coordinate y as well as the function $f(\mathbf{x}, \alpha)$ may take any value in the interval $(-\infty, \infty)$.
  2. The set of functions $Q(\mathbf{z}, \alpha)$, $\alpha \in \Lambda$ is on the form

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

and can take on arbitrary non-negative values.

Outline

1. Introduction
2. Minimizing the risk functional on the basis of empirical data
   - The problem of Pattern Recognition
   - The problem of Regression Estimation
   - The problem of Density Estimation (The Fisher-Wald setting)
   - Induction principles for minimizing the risk functional on the basis of empirical data

Problem definition

- Let $p(\mathbf{x},\alpha)$, $\alpha \in \Lambda$, be a set of probability densities containing the required density:

$$p(\mathbf{x},\alpha_0) = \frac{dF(\mathbf{x})}{d\mathbf{x}}$$

- Consider the functional:

$$R(\alpha) = \int \ln p(\mathbf{x},\alpha)dF(\mathbf{x}) \qquad (6)$$

- The problem of estmating the density in the L_1 metric is reduced to the minimization of the functional (6) on the basis of empirical data (Fisher-Wald formulation).

Assertions (I)

Functional's minimum

- The minimum of the functional (6) (if it exists) is attained at the functions $p(\mathbf{x}, \alpha^*)$ which may differ from $p(\mathbf{x}, \alpha_0)$ only on a set of zero measure.
- Demostration (based on Jensen's inequality):
  - Jensen's inequality implies:

$$\int \ln \frac{p(\mathbf{x}, \alpha)}{p(\mathbf{x}, \alpha_0)} dF(x) \leq \int \ln \frac{p(\mathbf{x}, \alpha)}{p(\mathbf{x}, \alpha_0)} p(\mathbf{x}, \alpha_0) d\mathbf{x} = \ln 1 = 0$$

  - So, the first assertion is proved by:

$$\int \ln \frac{p(\mathbf{x}, \alpha)}{p(\mathbf{x}, \alpha_0)} dF(x) = \int \ln p(\mathbf{x}, \alpha) dF(x) - \int \ln p(\mathbf{x}, \alpha_0) dF(x) \leq 0$$

Assertions (II)

The Bretagnolle-Huber inequality

- The Bretagnolle-Huber inequality

$$\int |p(\mathbf{x},\alpha) - p(\mathbf{x},\alpha_0)|\, d\mathbf{x} \leq 2\sqrt{1 - \exp\{R(\alpha_0) - R(\alpha)\}}$$

is valid.

Conclussion

- The functions $p(x, \alpha^*)$ which are ε-close to the minimum

$$R(\alpha^*) - \inf_{\alpha \in \Lambda} R(\alpha) < \varepsilon$$

will be $2\sqrt{1 - \exp\{-\varepsilon\}}$-close to the required density in the metric L_1.

Restrictions

- The density estimation problem in the Fisher-Wald setting is that the set of functions $Q(\mathbf{z}, \alpha)$ is subject to the following restrictions:
  - The vector $\mathbf{z}$ coincides with the vector $\mathbf{x}$.
  - The set of functions $Q(\mathbf{z}, \alpha)$, $\alpha \in \Lambda$, is on the form

    $$Q(\mathbf{z}, \alpha) = -\log p(\mathbf{x}, \alpha)$$

    where $p(\mathbf{x}, \alpha)$ is a set of density functions.
  - The loss function takes on arbitrary values on the interval $(-\infty, \infty)$.

Outline

Introduction

- We have seen that pattern recognition, regression estimation and density estimation problems can be reduced to this scheme by specifying a loss function in the risk functional.
- Now, how can we minimize the risk functional?
    - Classical: empirical risk minimization.
    - New one: structural risk minimization.

Empirical Risk Minimization
Definition

- Instead of minimizing the risk functional

$$R(\alpha) = \int Q(\mathbf{z}, \alpha) dF(\mathbf{z}), \qquad \alpha \in \Lambda$$

minimize the empirical risk functional

$$R_{emp}(\alpha) = \frac{1}{l} \sum_{i=1}^{l} Q(\mathbf{z}_l, \alpha), \qquad \alpha \in \Lambda$$

on the basis of empirical data $\mathbf{z_1}, \ldots, \mathbf{z_l}$ obtained according to a distribution function $F(\mathbf{z})$.

Empirical Risk Minimization

Considerations

- This functional is explicitely defined and it is subject to minimization.
- The problem is to establish conditions under which the minimum of the empirical risk functional, $Q(\mathbf{z}, \alpha_f)$, is close to the desired one, $Q(\mathbf{z}, \alpha_0)$.

Empirical Risk Minimization
Pattern Recognition problem (I)

- The Pattern Recognition problem is considered as the minimization of the functional

$$R(\alpha) = \int L(\omega, \phi(\mathbf{x}, \alpha))dF(\omega, \mathbf{x}), \qquad \alpha \in \Lambda$$

on a set of functions $\phi(\mathbf{x}, \alpha)$, $\alpha \in \Lambda$, that take on only a finite number of values, on the basis of empirical data

$$(\omega_1, \mathbf{x_1}), \ldots, (\omega_l, \mathbf{x_l})$$

Empirical Risk Minimization

Pattern Recognition problem (II)

- Considering the empirical risk functional

$$R_{emp}(\alpha) = \frac{1}{l}\sum_{i=1}^{l} L(\omega_i, \phi(\mathbf{x_i}, \alpha)), \qquad \alpha \in \Lambda$$

- When $L(\omega_i, \phi) \in \{0, 1\}$ (0 if $\omega = \phi$ and 1 if $\omega \neq \phi$), minimization of the empirical risk functional is equivalent to minimize the number of training errors.

Empirical Risk Minimization
Regression Estimation problem (I)

- The Regression Estimation problem is considered as the minimization of the functional

$$R(\alpha) = \int (y - f(\mathbf{x}, \alpha))^2 dF(y, \mathbf{x}), \qquad \alpha \in \Lambda$$

on a set of functions $f(\mathbf{x}, \alpha)$, $\alpha \in \Lambda$, on the basis of empirical data

$$(y_1, \mathbf{x_1}), \ldots, (y_l, \mathbf{x_l})$$

Empirical Risk Minimization
Regression Estimation problem (II)

- Considering the empirical risk functional

$$R_{emp}(\alpha) = \frac{1}{l}\sum_{i=1}^{l}(y_i - f(\mathbf{x_i}, \alpha))^2, \qquad \alpha \in \Lambda$$

- The method of minimizing this empirical risk functional is known as the *Least-Squares Method*.

Empirical Risk Minimization
Density Estimation problem (I)

- The Density Estimation problem is considered as the minimization of the functional

$$R(\alpha) = \int \ln p(\mathbf{x}, \alpha) dF(\mathbf{x}), \qquad \alpha \in \Lambda$$

on a set of densities $p(\mathbf{x}, \alpha)$, $\alpha \in \Lambda$, using i.i.d. empirical data

$$\mathbf{x_1}, \ldots, \mathbf{x_l}$$

Empirical Risk Minimization

Density Estimation problem (II)

- Considering the empirical risk functional

$$R_{emp}(\alpha) = -\sum_{i=1}^{l} \ln p(\mathbf{x}, \alpha), \qquad \alpha \in \Lambda$$

- It is the same solution which comes from the *Maximum Likelihood Method* (in the maximum likelihood method a plus sign is used in from of the sum instead of the minus sign).

Questions?

Thank you very much for your attention.

- Contact:
  - Miguel Angel Veganzones
  - Grupo Inteligencia Computacional
  - Universidad del País Vasco - UPV/EHU (Spain)
  - E-mail: miguelangel.veganzones@ehu.es
  - Web page: http://www.ehu.es/computationalintelligence