Statistical Learning Theory
Fundamentals

Miguel A. Veganzones

Grupo Inteligencia Computacional
Universidad del País Vasco
Outline

1. Introduction
   - Learning problem
   - Statistical learning theory

2. Minimizing the risk functional on the basis of empirical data
   - The pattern recognition problem
   - The regression problem
   - The density estimation problem (Fisher-Wald setting)
   - Induction principles for minimizing the risk functional on the basis of empirical data
Outline

1. Introduction
   - Learning problem
     - Statistical learning theory

2. Minimizing the risk functional on the basis of empirical data
   - The pattern recognition problem
   - The regression problem
   - The density estimation problem (Fisher-Wald setting)
   - Induction principles for minimizing the risk functional on the basis of empirical data
Approaches to the learning problem

- Learning problem: the problem of choosing the desired dependence on the basis of empirical data.

- Two approaches:
  - To choose an approximating function from a given set of functions.
  - To estimate the desired stochastic dependences (densities, conditional densities, conditional probabilities).
First approach
To choose an approximating function from a given set of functions

- It’s a problem rather general.
- Three subproblems: pattern recognition, regression, 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, the regression and the 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
The Generator, \( G \), determines the environment in which the Supervisor and the Learning Machine act.

Simplest environment: \( G \) generates the vector \( \mathbf{x} \in \mathbf{X} \) independently and identically distributed (i.i.d.) according to some unknown but fixed probability distribution function, \( F(\mathbf{x}) \).
The 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 $(x_1, y_1), \ldots, (x_l, y_l)$, the training set, which is drawn randomly and independently according to a joint distribution function $F(x, y) = F(y|x)F(x)$.

- 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 $x_i$ generated by $G$. 

http://www.ehu.es/ccwintco (Grupo Inteligencia Computacional Universidad del País Vasco)
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 appropriate function from a given set of functions.
Outline

1. Introduction
   - Learning problem
   - Statistical learning theory

2. Minimizing the risk functional on the basis of empirical data
   - The pattern recognition problem
   - The regression problem
   - The density estimation problem (Fisher-Wald setting)
   - Induction principles for minimizing the risk functional on the basis of empirical data
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 $\mathbb{R}^n$, a set of admissible functions $\{g(z)\}$, $z \in Z$, is given and a functional $R = R(g(z))$ is defined.

- It’s required to find the function $g'(z)$ from the set $\{g(z)\}$ which minimizes the functional $R = R(g(z))$. 
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 p.d.f. $F(\mathbf{z})$ is defined on $Z$ and the functional is defined as the mathematical expectation

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

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 $\mathbf{z}_1, \ldots, \mathbf{z}_l$ 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 \( \{g(z)\} \) a function which will guarantee that the functional (1) is small?

The minimization of the functional (1) on the basis of empirical data \( z_1, \ldots, z_l \) is one of the main problems of mathematical statistics.
Parametrization

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

\[
R(\alpha) = \int Q(z, \alpha) \, dF(z), \quad \alpha \in \Lambda
\]

(2)

where \( Q(z, \alpha) = L(z, g(z, \alpha)) \) is called the loss function.
The expected loss

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

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

which is called the risk functional or the risk.
Introduction

Minimizing the risk functional on the basis of empirical data

Problem redefinition

Definition

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

1 Introduction
   - Learning problem
   - Statistical learning theory

2 Minimizing the risk functional on the basis of empirical data
   - The pattern recognition problem
   - The regression problem
   - The density estimation problem (Fisher-Wald setting)
   - Induction principles for minimizing the risk functional on the basis of empirical data
Informal definition

Definition

A supervisor observes occurring 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. \( F(x) \), situation \( x \) appears randomly and independently. the supervisor classifies each situation into one of \( k \) classes. We assume that the supervisor carries out this classification by \( F(\omega | x) \), where \( \omega \in \{0, 1, \ldots, k - 1\} \).

- Neither \( F(x) \) nor \( F(\omega, x) \) are known, but they exist.
- Thus, a joint distribution function \( F(\omega, x) = F(\omega | x)F(x) \) exists.
Loss function

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

\[
L(\omega, \phi) = \begin{cases} 
1 & \text{if } \omega = \phi \\
0 & \text{if } \omega \neq \phi 
\end{cases}
\]

- The problem of pattern recognition is to minimize the functional

\[
R(\alpha) = \int L(\omega, \phi(x, \alpha)) \, dF(\omega, x)
\]

on the set of functions \( \phi(x, \alpha), \alpha \in \Lambda \), where the p.d.f. \( F(\omega, x) \) is unknown but a random independent sample of pairs \( (x_1, \omega_1), \ldots, (x_l, \omega_l) \) 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 $Q(z, \alpha), \alpha \in \Lambda$, is not arbitrary as in the general case.

- The following restrictions are imposed:
  - The vector $z$ consist of $n+1$ coordinates: coordinate $\omega$ (which takes a finite number of values) and $n$ coordinates $x_1, x_2, \ldots, x_n$ which form the vector $x$.
  - The set of functions $Q(z, \alpha), \alpha \in \Lambda$ is given by
    \[
    Q(z, \alpha) = L(\omega, \phi(x, \alpha)), \quad \alpha \in \Lambda
    \]
    and also takes on only a finite number of values.
Outline

1 Introduction
   - Learning problem
   - Statistical learning theory

2 Minimizing the risk functional on the basis of empirical data
   - The pattern recognition problem
   - The regression problem
   - The density estimation problem (Fisher-Wald setting)
   - Induction principles for minimizing the risk functional on the basis of empirical data
Stochastic dependences

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

- The function of conditional mathematical expectation
  
  \[ r(x) = \int y dF(y|x) \]

  is called the *regression*.

- Estimate the regression in the set of functions \( f(x, \alpha), \alpha \in \Lambda \), 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:

\[
\int y^2 dF(y,x) < \infty \quad \int r^2(x) dF(y,x) < \infty
\]

- On the set \( f(x, \alpha), \alpha \in \Lambda, f(x, \alpha) \in L_2 \), the minimum (if exists) of the functional

\[
R(\alpha) = \int (y - f(x, \alpha))^2 dF(y,x)
\]

is attained at:

- The regression function if \( r(x) \in f(x, \alpha), \alpha \in \Lambda \).
- The function \( f(x, \alpha^*) \) which is the closest to \( r(x) \) in the \( L_2 \) metric if \( r(x) \notin f(x, \alpha), \alpha \in \Lambda \).
Demonstration

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

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

$$-2 \int \Delta f (x, \alpha) (y - r (x))^2 \, dF (y, x)$$

- $\int (y - r (x))^2 \, dF (y, x)$ does not depend of $\alpha$.
- $\int \Delta f (x, \alpha) (y - r (x))^2 \, dF (y, x) = 0$. 
Restrictions

- The problem of estimating the regression may be also reduced to the scheme of minimizing the risk. The following restrictions are imposed:
  - The vector $z$ consist of $n + 1$ coordinates: coordinate $y$ and $n$ coordinates $x^1, x^2, \ldots, x^n$ which form the vector $x$. However, the coordinate $y$ as well as the function $f(x, \alpha)$ may take any value on the interval $(-\infty, \infty)$.
  - The set of functions $Q(z, \alpha), \alpha \in \Lambda$ is on the form
    \[
    Q(z, \alpha) = (y - f(x, \alpha))^2, \quad \alpha \in \Lambda
    \]
    and can take on arbitrary non-negative values.
Outline

1. Introduction
   - Learning problem
   - Statistical learning theory

2. Minimizing the risk functional on the basis of empirical data
   - The pattern recognition problem
   - The regression problem
   - The density estimation problem (Fisher-Wald setting)
   - Induction principles for minimizing the risk functional on the basis of empirical data
Problem definition

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

\[
p(x, \alpha_0) = \frac{dF(x)}{dx}
\]

- Considering the functional:

\[
R(\alpha) = \int \ln p(x, \alpha) dF(x)
\] (4)

the problem of estimating the density in the \( L_1 \) metric is reduced to the minimization of the functional (4) on the basis of empirical data (Fisher-Wald formulation).

http://www.ehu.es/ccwintco (Grupo Inteligencia Computacional Universidad del País Vasco)
Assertions (I)
Functional’s minimum

- The minimum of the functional (4) (if it exists) is attained at the functions \( p(x, \alpha^*) \) which may differ from \( p(x, \alpha_0) \) only on a set of zero measure.

- Demonstration:
  - Jensen’s inequality implies:
    \[
    \int \ln \frac{p(x, \alpha)}{p(x, \alpha_0)} \, dF(x) \leq \int \ln \frac{p(x, \alpha)}{p(x, \alpha_0)} p(x, \alpha_0) \, dx = \ln 1 = 0
    \]
  - So, the first assertion is proved by:
    \[
    \int \ln \frac{p(x, \alpha)}{p(x, \alpha_0)} \, dF(x) = \int \ln p(x, \alpha) \, dF(x) - \int \ln p(x, \alpha_0) \, dF(x) \leq 0
    \]
The Bregtanolle-Huber inequality:

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

is valid.
Conclusion

- The functions \(p(x, \alpha^*)\) which are \(\varepsilon\)-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 \(L_1\) metric.
Restrictions

- The density estimation problem in the Fisher-Wald setting is that the set of functions $Q(z, \alpha)$ is subject to the following restrictions:
  - The vector $z$ coincides with the vector $x$.
  - The set of functions $Q(z, \alpha), \alpha \in \Lambda$, is on the form
    \[
    Q(z, \alpha) = -\ln p(x, \alpha)
    \]
    where $p(x, \alpha)$ is a set of density functions.
  - The loss function takes on arbitrary values on the interval $(-\infty, \infty)$. 

http://www.ehu.es/ccwintco

Vapnik

UPV/EHU
Outline

1 Introduction
   - Learning problem
   - Statistical learning theory

2 Minimizing the risk functional on the basis of empirical data
   - The pattern recognition problem
   - The regression problem
   - The density estimation problem (Fisher-Wald setting)
   - Induction principles for minimizing the risk functional on the basis of empirical data

http://www.ehu.es/ccwintco (Grupo Inteligencia Computacional Universidad del País Vasco)
We have seen that the pattern recognition, the regression and the 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 when the density function is unknown?

- Classical: empirical risk minimization (ERM).
- New one: structural risk minimization (SRM).
Empirical Risk Minimization

Definition

Instead of minimizing the risk functional:

\[ R(\alpha) = \int Q(z, \alpha) dF(z), \quad \alpha \in \Lambda \]

minimize the empirical risk functional

\[ R_{\text{emp}}(\alpha) = \frac{1}{l} \sum_{i=1}^{l} Q(z_i, \alpha), \quad \alpha \in \Lambda \]

on the basis of empirical data \( z_1, \ldots, z_l \) obtained according to a distribution function \( F(z) \).
Empirical Risk Minimization
Considerations

- The functional is explicitly defined and it is subject to minimization.
- The problem is to establish conditions under which the minimum of the empirical risk functional, $Q(z, \alpha_f)$, is closed to the desired one $Q(z, \alpha_0)$. 
Empirical Risk Minimization
Pattern recognition problem (I)

The pattern recognition problem is considered as the minimization of the functional

$$R(\alpha) = \int L(\omega, \phi(x, \alpha)) dF(\omega, x), \quad \alpha \in \Lambda$$

on a set of functions $\phi(x, \alpha), \alpha \in \Lambda$, that take on only a finite number of values, on the basis of empirical data $(x_1, \omega_1), \ldots, (x_l, \omega_l)$. 
Empirical Risk Minimization
Pattern recognition problem (II)

- Considering the empirical risk functional

\[ R_{\text{emp}}(\alpha) = \frac{1}{l} \sum_{i=1}^{l} L(\omega_i, \phi(x_i, \alpha)), \quad \alpha \in \Lambda \]

- When \( L(\omega_i, \alpha) \in \{0, 1\} \) (0 if \( \omega = \phi \) and 01 if \( \omega \neq \phi \)), minimization of the empirical risk functional is equivalent to minimizing the number of training errors.
Empirical Risk Minimization
Regression problem (I)

- The regression problem is considered as the minimization of the functional

\[ R(\alpha) = \int (y - f(x, \alpha))^2 dF(y, x), \quad \alpha \in \Lambda \]

on a set of functions \( f(x, \alpha), \alpha \in \Lambda \), on the basis of empirical data \( (x_1, y_1), \ldots, (x_l, y_l) \).
Empirical Risk Minimization
Regression problem (II)

- Considering the empirical risk functional

\[ R_{\text{emp}}(\alpha) = \frac{1}{l} \sum_{i=1}^{l} (y_i - f(x_i, \alpha))^2, \quad \alpha \in \Lambda \]

- The method of minimizing the 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

\[ R(\alpha) = \int \ln p(x, \alpha) \, dF(x), \quad \alpha \in \Lambda \]

on a set of densities \( p(x, \alpha), \alpha \in \Lambda \), using i.i.d. empirical data \( x_1, \ldots, x_l \).
Empirical Risk Minimization
Density estimation problem (II)

Considering the empirical risk functional

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

it is the same solution which comes from the Maximum Likelihood method (in the Maximum Likelihood method a plus sign is used in front of the sum instead of the minus sign).
For Further Reading

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