An Overview of Statistical Learning Theory

German Hernandez

Ing. Sistemas UN-Bogota.
Vast amounts of data are being generated in many fields, and the statistician’s job is to make sense of it all: to extract important patterns and trends, and understand “what the data says”. We call this learning from data.

Hastie, Tibshirani and Friedman [9]
Abstract

This lecture presents an overview of statistical learning. We review some automated (machine) learning algorithms from a statistical viewpoint. Also we present a unifying formal setting for this learning algorithms.
1. **Goal**

The focus of this lecture is to give an introductory overview of some automated (machine) learning algorithms, from the statistical point of view. These include disparate algorithms such as regression, neural networks and decision trees. Some other learning algorithms such as principal components analysis (PAC), Bayesian networks, hidden Markov models are not included in this overview.
2. Introduction

In recent years we have witnessed the application of a variety of automated (machine) learning algorithms, particularly processing data sets whose sizes and complexities are beyond the ability of humans to handle.

Such algorithms have appeared in diverse fields such as: computer vision (face and character recognition), text and information retrieval (web mining e.g. google, spam detection), data mining (discovering patterns in large data bases), bioinformatics (DNA and protein sequence analysis, genomics), speech recognition, time series prediction (weather, finance), game playing (reinforcement, web agents), etc.

This collection of data sets is growing (e.g. the genome project) and the data sets are generally available online.
3. Learning

Here learning will be considered as the process of gaining experience, knowledge or skill through training or schooling, in order to perform a task.

Some examples of tasks where learning algorithms have been successful are autonomous driving, speech recognition, handwritten character recognition and systems which learn user’s interests (e.g. newsreaders).
In general we have two types of learning: **supervised learning** when we have a teacher or critic that provides feedback to the learner about responses to the training examples, and **unsupervised** when we have only training examples but no feedback about the answers of the learner.
Learning algorithms try to emulate the learning capabilities of humans; a computer program or algorithm is said to learn if it improves its performance at some task through experience [7].

3.1. Elements of a model of learning

There are some elements of any model of learning needs to specify, these are:

3.1.1. Learner:

“Who” does the learning? In our case, we will consider a computer program, and it is often required that an algorithm performs a task efficiently (in polynomial time), and using a finite amount of memory.
3.1.2. Domain:

“What” is being learned? In automated learning we have:

- **Concept learning (Negative-positive classification):** when the learner is trying to learn an unknown concept (such as what is a chair and what is not a chair). In concept learning, the learner is trying to come up with a “rule” to separate positive examples from negative examples, i.e., the learner’s goal is to infer how an unknown target function classifies examples as positive or negative from a given domain. As a concrete example consider the task of handwritten character recognition.

- **Function learning:** when the learner is trying to learn an unknown function (a classical example is learning a physical law by curve fitting to data[3]).

- **Technique/behavior learning:** when the learner is trying to learn an unknown technique (such as a how to play a game in order to win).
3.1.3. Information Source:

“From what” is the learner learning? How is the learner informed about the domain? This can happen in many ways:

a. *Examples*: the learner is given examples. These examples can be chosen in a variety of ways:
   - these can be chosen at random from some known or unknown distribution,
   - these can be chosen arbitrarily,
   - these can be chosen maliciously by some adversary who wants to know the worst these behavior of a learning algorithm,
   - these can be carefully chosen by a helpful teacher who wants to facilitate the learning process,
   - these can be generated for some random process (e.g., nature).

b. *Queries*: the learner may get information about the domain by asking questions to a teacher.

c. *Experimentation*: the learner may get information about the domain by actively experimenting with it.
3.1.4. Information Representation:

How are features of the examples described?

3.1.5. Prior Knowledge:

What does the learner know about the domain initially?

3.1.6. Performance Criteria:

We have several possible criteria for evaluating a learning algorithm:

a. Accuracy
b. Speed of learning
c. Speed of answering
d. Space requirements
In this lecture we will consider two tasks that are solved by supervised learning algorithms: function-fitting and classification.

4. **Function-fitting and classification learning tasks**

4.1. **Function-fitting**

Assume that we want to “learn” a function \[ f : \mathbb{R}^p \rightarrow \mathbb{R}^k \]

that underlies the relationship between some inputs and outputs and we have at hand a *training set* of observations or examples

\[
\mathcal{T} = \{(x_i, y_i) \mid i = 1, 2, \ldots, n\}.
\]

If the measurements generating this set were exact, then \( f(x_i) \) should be equal to \( y_i \), but in general, we expect this values to be affected by noise.
In some contexts the $x_i$ rather than be chosen to do the training, are being generated by a probability distribution on $\mathbb{R}^p$. Thus, one might take as starting point, instead of the unknown $f$, a probability measure on $\mathbb{R}^k$ for each $x \in \mathbb{R}^p$, i.e., we have the conditional probability measures $\mu(y|x)$ on $\mathbb{R}^k$, for each $x \in \mathbb{R}^p$ and this case the $y_i$ observed is a sample of the probability measure on $\mathbb{R}^k$ for the given $x_i$, (see figure 2). In many cases the starting point is only a single probability measure $\mu$ on $\mathbb{R}^p \times \mathbb{R}^k$ from which the pairs $(x_i, y_i)$ are randomly drawn. In section 4 we will relate this probability with the unknown function $f$. We have many learning algorithms to approximate the unknown function next we review two of them: regression and neural networks.
Observe obtained for $\mu(y|x_i)$

Conditional probability measures $\mu(y|x)$
4.1.1. Regression

In regression we assume that the unknown function $f$, has a specific form and that the space of all this functions can be parameterized by $N$ real numbers, i.e., we have one candidate function $f_\omega$ for every vector $\omega \in \mathbb{R}^N$. Usually we look for the best fit estimate by the least squares method

$$f^*(x) = \arg \min_{f_\omega} \sum_{i=1}^{n} (f_\omega(x) - y_i)^2$$

In linear regression we consider that the functions of the form

$$f(x) = x^t \beta$$

with $\beta$ the matrix of $N = p \times k$ of parameters. In a more general context we can consider functions of the form

$$f_\omega(x) = \sum_{r=1}^{R} \phi_r(x) \omega_k;$$

where the functions $\phi_r$ form a linear basis, by example polynomial or trigonometric expansions.
4.1.2. Multilayer Neural Networks

The use multilayer neural networks to approximate functions is equivalent to do a regression in which the function is seen as superposition of sigmoid functions [2]

\[ \sigma(x) = \frac{1}{1 + e^{-x}}, \]

this function has the nice property of \( \frac{d\sigma(x)}{dx} = \sigma(x)(1 - \sigma(x)) \) that allows to derive gradient descent rules to train.
**Goal**

Introduction

Learning

Function-fitting and...

Formal Setting

---

**Neuron (7)**

\[
\text{net} = \sum_{i=0}^{n} w_i x_i
\]

\[
o = \sigma(\text{net}) = \frac{1}{1 + e^{-\text{net}}}
\]
An element that applies the sigmoid function $\sigma$ to a linear combinations of its inputs is a called neuron, it is called this way because it has some analogy with the neurons of the nervous system. An array of at least three layers of this neuron is called a multilayer neural network.
**Multilayer (3-layer) Neural Networks Topology [7]**
This neural network defines a function of the form

$$f_\omega(x) = (\sigma(\omega_{o1} h(x)), \sigma(\omega_{o2} h(x)), \ldots, \sigma(\omega_{ok} h(x)))$$

and

$$h(x) = (\sigma(\omega_{h1} x), \sigma(\omega_{h2} x), \ldots, \sigma(\omega_{hl} x))$$

where $\omega_{oj}$ are the weights of the $j$th output neuron, $\omega_{hj}$ are the weights of the $j$th hidden neuron and $h$ are the outputs of the hidden neurons. In this case we have $N = (p \times l) + (l \times k)$ parameters. There is a gradient decent method of training for this type of networks [4] and a proof [1][2] that with only three layers this networks can approximate, to nay degree of precision, any measurable function.
4.2. Classification

Given some objects with $k$ characteristics that can be codified as real numbers. Let's assume that there are $p$ different classes of this objects $\{C_1, C_2, ..., C_p\}$. We “learn” to classify this objects with a training set of classified examples

$$\mathcal{T} = \{(x_k, c_k) | i = 1, 2, ..., n\}.$$ 

This problem reduces to “learn” a function

$$f: \mathbb{R}^p \rightarrow [0, 1]^k$$

that assigns to an object in class $C_i$, more exactly to its vector of characteristics, the vector $e_i$. We have many learning algorithms to approximate this function, next we discuss two of them: decision trees and multilayer neural networks
4.2.1. Decision Trees

Decision trees are considered when we have a classification problem that have:

- Instances describable by attribute–value
- Target function is discrete valued
- Disjunctive hypothesis may be required
- Possibly noisy training data

Some examples are equipment or medical diagnosis and credit risk analysis. In a decision tree we have that:

- Each internal node tests an attribute
- Each branch corresponds to attribute value
- Each leaf node assigns a classification
DECISION TREE FOR PLAYING TENNIS[7]
The top-down induction algorithm for decision trees is

Main loop:

1. $A \leftarrow$ the “best decision attribute using information gain for next node

2. Assign $A$ as decision attribute for node

3. For each value of $A$, create new descendant of node

4. Sort training examples to leaf nodes

5. If training examples perfectly classified, Then STOP, Else iterate over new leaf nodes

The information gain (entropy) measures how well a given attribute separates the training examples according to the target classification.
Here are the classification function is of the form

\[ f(x) = \begin{cases} 
(1, 0, 0, 0, \ldots, 0) & \text{if } x^1 \in A^{11}, x^2 \in A^{12}, \ldots, x^p \in A^{1p} \\
(0, 1, 0, 0, \ldots, 0) & \text{if } x^1 \in A^{21}, x^2 \in A^{22}, \ldots, x^p \in A^{2p} \\
\vdots & \\
(0, 0, 0, \ldots, 0, 1) & \text{if } x^1 \in A^{k1}, x^2 \in A^{k2}, \ldots, x^p \in A^{kp} 
\end{cases} \]

where \( x = (x^1, x^2, x^3, \ldots, x^p) \) and \( A^{ij} \) is the set defined by the union of all the conditions on attribute \( j \) over branches of the tree that lead to class \( i \).
4.2.2. **Multilayer Neural Networks Classifier**

Here the application to approximate the classification function is the application of the multilayer neural network to approximate particular case of function. To illustrate this use let an example
Multilayer neural networks classifier of phonemes using the two first formants [7]
5. **Formal Setting**

Now we present a unifying formal setting for the algorithms presented based on [3]. Let us consider the following probability space

$$(\Omega = X \times Y, \beta_\Omega, \mu)$$

where

i) $X$ is the space of inputs of our learning task,

ii) $Y$ is the space of outputs of our learning task,

iii) $\beta_\Omega$ is the Borel $\sigma$-algebra on $\Omega$,

vi) $\mu$ is a probability measure on $\Omega$.

Then let us assume that the training set is produced by sampling $\Omega$ independently with probability $\mu$ and obtaining
\[(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).\]

For every integrable function

\[g : X \times Y \to \mathbb{R}^k\]

a version of the Fubini’s theorem states that

\[\int_{\Omega} g(x, y) \, d\mu = \int_X \left( \int_Y g(x, y) \, d\mu(y|x) \right) \, d\mu_X\]

where \(\mu_X(B) = \mu(\pi^{-1}(B))\). We define the regression function \(f_\mu\) of \(\mu\) by

\[f_\mu = \int_Y y \, d\mu(y|x),\]

this is \(f_\mu(x) = \mathbb{E}[y|x]\).
In each the learning algorithms presented before we define a particular family of integrable functions $\mathcal{F}$ and our goal is to “learn” (i.e., to find a good approximation of) $f_\mu$ on $\mathcal{F}$. Usually the metric is

$$
\mathcal{E}(f, f_\mu) = \int_X [f - f_\mu]^2 \, d\mu_X.
$$

Given that we have only a random sample of training examples from $\Omega$ we look for the best approximation in least squares error.
References


