Supervised Learning

Corso di AA, anno 2016/17, Padova

Fabio Aiolli

10 Ottobre 2016

Università degli Studi di Padova

Outline

- When and why do we need to learn?
    - Examples of applications
    - Common tasks in ML
- How can we learn?
    - Hypothesis space
    - Learning (searching) algorithm
    - Examples of hypothesis spaces
    - Examples of algorithms
- but above all... Can we actually learn?
    - VC-dimension
    - A learning bound

Università degli Studi di Padova

Supervised Learning

Formalization and terminology

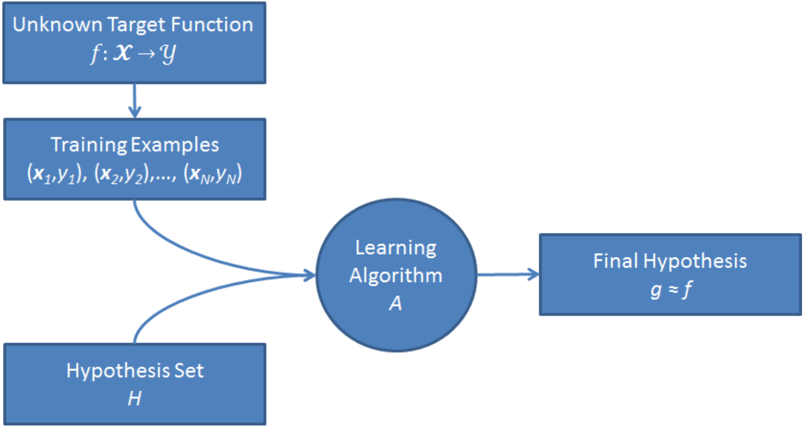
- Input: $\mathbf{x} \in \mathcal{X}$ (e.g. email representation)
- Output: $y \in \mathcal{Y}$ (e.g. spam/no_spam)
  - classification: $\mathcal{Y} \equiv \{-1, +1\}$
  - regression: $\mathcal{Y} \equiv \mathbb{R}$
- Oracle (two alternatives):
  - Target function $f : \mathcal{X} \rightarrow \mathcal{Y}$ deterministic (ideal and unknown!)
  - Probability distributions $P(\mathbf{x}), P(y|\mathbf{x})$ stochastic version (still unknown!)
- Data: $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ (e.g. historical records)

- Select an hypothesis $g : \mathcal{X} \rightarrow \mathcal{Y}$ from a set $\mathcal{H}$ using training data

Supervised Learning in action

- A series of pairs $(\mathbf{x}_i, y_i)$, called training set, is available. These pairs are supposed generated according to a probability function $P(\mathbf{x}, y) = P(\mathbf{x})P(y|\mathbf{x})$
- A 'plausible' hypothesis $g : \mathcal{X} \rightarrow \mathcal{Y}$ is selected from a set $\mathcal{H}$ using training data
- The selected hypothesis g should generalize well: correct predictions should be done for new unseen examples drawn according to $P(\mathbf{x}, y)$
- The error on training data is called empirical error/risk
- The expected error on any given pair $(\mathbf{x}, y)$ drawn according to $P(\mathbf{x}, y)$ is called ideal error/risk

Supervised Learning: the learning setting



Learning Puzzles

x	y
1	2
2	4
3	6
5	10
4	?

x	y
10101	+1
11011	+1
01100	-1
00110	-1
01110	?

Are these impossible tasks?

YES

Does this mean learning is an impossible task?

NO

The fundamental assumption in ML

ML Assumption

There is a stochastic process which explains observed data. We do not know the details about it but we know it is there!

e.g. the social behavior is not purely random!

The aim of Machine Learning is to build good (or useful) approximations of this process.

The Inductive Bias

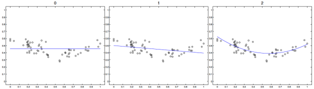
For learning to be feasible, further assumptions have to be made about the ‘complexity’ of the unknown target function and the hypothesis space.

- The hypothesis space cannot contain all possible formulas or functions
- The assumptions we make about the type of function to approximate is called inductive bias
- In particular, it consists of:
  - The hypothesis set: definition of the space $\mathcal{H}$
  - The learning algorithm, how the space $\mathcal{H}$ is explored

Example: Polynomial Regression

Let's take a simple example:

- Training data $\mathcal{S} = \{(x_1, y_1), \ldots, (x_n, y_n)\}, x \in \mathbb{R}, y \in \mathbb{R}$
- We want to find a polynomial curve that approximates the examples above, that is a function of type:
  $h_{\mathbf{w}}(x) = w_0 + w_1 x + w_2 x^2 + \cdots + w_p x^p,\ p \in \mathbb{N}$



- $\text{error}_S(h_{\mathbf{w}}) = \frac{1}{n} \sum_{i=1}^{n} (h_{\mathbf{w}}(x_i) - y_i)^2$
- $\text{TEST} = \{(x_{n+1}, y_{n+1}), \ldots, (x_N, y_N)\}$

Questions:

- How can we choose p? ($\mathcal{H}$ definition)
- How can we choose w's? ($\mathcal{H}$ search)

Example: Polynomial Regression

- Given a p, the problem becomes:
- $[\mathbf{X}]_i = [1, x_i, x_i^2, \ldots, x_i^p]$ (i-th row of the matrix $\mathbf{X}$)
- $[\mathbf{y}]_i = y_i$ (i-th row of the vector $\mathbf{y}$)
- TRAIN: Solve $\min_{\mathbf{w}} \frac{1}{n}||\mathbf{y} - \mathbf{X}\mathbf{w}||^2 + \beta||\mathbf{w}||^2$ by using the *ridge regression* method:
- $\mathbf{w} = (\mathbf{X}^\top \mathbf{X} + \beta \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$

Università degli Studi di Padova