

Supervised Learning

Machine Learning, A.Y. 2022/23, Padova



Fabio Aioli

October 10th, 2022



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

Supervised Learning

Formalization and terminology



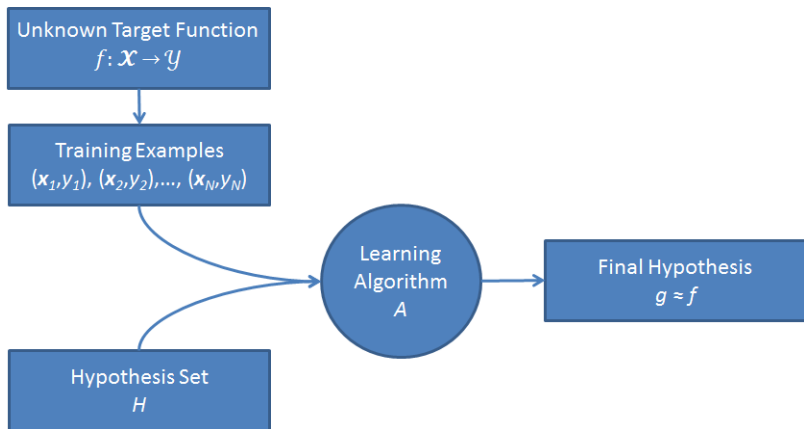
- Input: $\mathbf{x} \in \mathcal{X}$ (e.g. email representation)
- Output: $y \in \mathcal{Y}$ (e.g. spam/no_spam)
 - classification: $\mathcal{Y} \equiv \{-1, +1\}$
 - multi-class classification: $\mathcal{Y} \equiv \{1, \dots, m\}$
 - regression: $\mathcal{Y} \equiv \mathbb{R}$
- Oracle/Nature (two 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: typically *independent and identically distributed (i.i.d.)* data $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$. No time-series (independent). No concept drift (identically distributed).
- A predefined set of hypotheses/functions $\mathcal{H} \equiv \{h|h : \mathcal{X} \rightarrow \mathcal{Y}\}$ (a.k.a. hypothesis space)



Supervised Learning in action

- A set of pairs (\mathbf{x}_i, y_i) , called **training set**, is available. These pairs are supposed to be generated according to an unknown 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 the hypothesis space \mathcal{H} using training data
- The error on training data is called **empirical error/risk** (this is easily computable)
- The selected hypothesis g should **generalize well**: correct predictions should be done for **new unseen examples** drawn according to $P(\mathbf{x}, y)$, i.e. minimizing the ideal risk.
- The expected error on given pairs (\mathbf{x}, y) drawn according to $P(\mathbf{x}, y)$ is called **ideal error/risk** (this can be only estimated)

Supervised Learning: the learning setting





x	y
10101	+1
11011	+1
01101	-1
00110	-1
01110	?

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

Are these impossible tasks?

YES

Does this mean **learning** is an impossible task?

NOT NECESSARILY



ML Assumption

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

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

e.g. images of the same digit share common patterns!

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

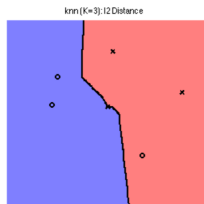
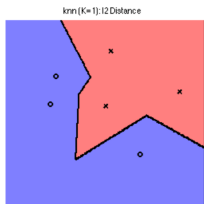
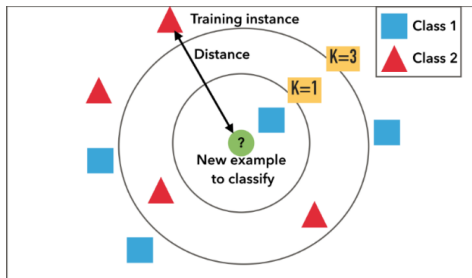


Machine Learning is not magic!!

For learning to be feasible, further assumptions must 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**.
- Specifically, the inductive bias consists of:
 - The **hypothesis space**: how the space \mathcal{H} is defined
 - The **learning algorithm**, how the space \mathcal{H} is explored

Example: K-Nearest-Neighbor (kNN) inductive bias



- Nearby instances are expected to have the same label
- Complexity of \mathcal{H} is tuned by the K parameter

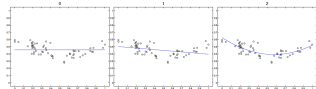


Example: Polynomial Regression

Let's take another simple example:

- Training data $\mathcal{S} = \{(x_1, y_1), \dots, (x_n, y_n)\}, x \in \mathbb{R}, y \in \mathbb{R}$
- We want to find a polynomial curve that approximates the examples above, that is, a function of type:

$$h_{\mathbf{w}}(x) = w_0 + w_1x + w_2x^2 + \dots + w_px^p, p \in \mathbb{N}$$



- $\text{error}_{\mathcal{S}}(h_{\mathbf{w}}) = \frac{1}{n} \sum_{i=1}^n (h_{\mathbf{w}}(x_i) - y_i)^2$ (empirical error)
- TEST = $\{(x_{n+1}, y_{n+1}), \dots, (x_N, y_N)\}$ (for estimating the ideal error)

Questions:

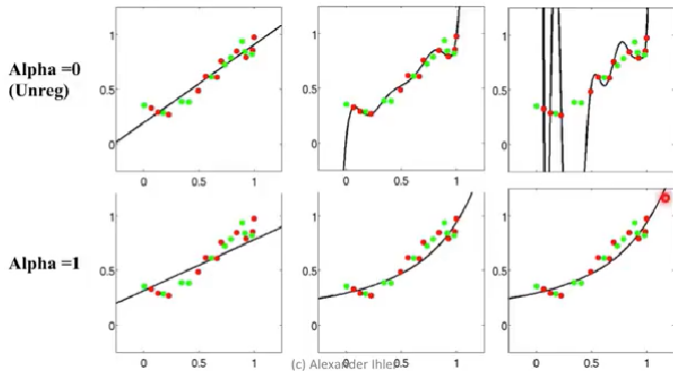
- How can we choose p ? (\mathcal{H} definition)
- How can we choose w 's? (\mathcal{H} search)



Example: Polynomial Regression

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

The effect of α

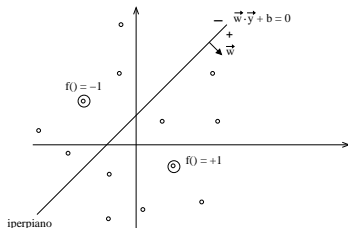


Hypothesis Space: Example 1



Hyperplanes in \mathbb{R}^2

- Instance space: points in the plane $\mathcal{X} = \{y | y \in \mathbb{R}^2\}$
- Hypothesis space: dichotomies induced by hyperplanes in \mathbb{R}^2 , that is $\mathcal{H} = \{f_{\mathbf{w},b}(y) = \text{sign}(\mathbf{w} \cdot y + b), \mathbf{w} \in \mathbb{R}^2, b \in \mathbb{R}\}$



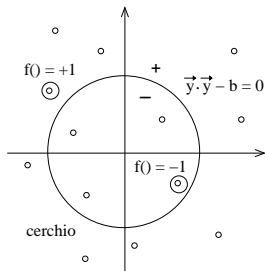
What changes in \mathbb{R}^n ?



Hypothesis Space: Example 2

Circles in \mathbb{R}^2

- Instance space: points in the plane $\mathcal{X} = \{y | y \in \mathbb{R}^2\}$
- Hypothesis space: dichotomies induced by circles centered at the origin in \mathbb{R}^2 , that is $\mathcal{H} = \{f_b(y) = \text{sign}(\|y\|^2 - b), b \in \mathbb{R}\}$



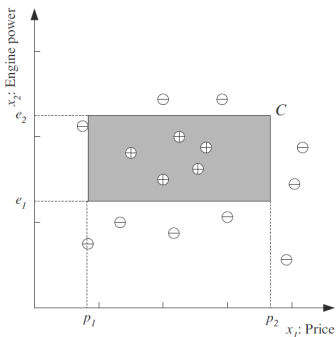
What changes in \mathbb{R}^n ?



Hypothesis Space: Example 3

Rectangles in \mathbb{R}^2

- Instance space: points in the plane $\mathcal{X} = \{(p, e) | (p, e) \in \mathbb{R}^2\}$
- Hypothesis space: dichotomies induced by rectangles in \mathbb{R}^2 , that is $\mathcal{H} = \{f_\theta(y) = [p_1 \leq p \leq p_2 \cap e_1 \leq e \leq e_2], \theta = \{p_1, p_2, e_1, e_2\}\}$ where $[z] = +1$ if $z = \text{True}$, -1 otherwise.



What changes in \mathbb{R}^n



Hypothesis Space: Example 4

Conjunction of m positive literals

- Instance space: strings of m bits, $\mathcal{X} = \{s | s \in \{0, 1\}^m\}$
- Hypothesis space: all the logic sentences involving positive literals l_1, \dots, l_m (l_1 is true if the first bit is 1, l_2 is true if the second bit is 1, etc.) and just containing the operator \wedge (**and**)

$$\mathcal{H} = \{f_{\{i_1, \dots, i_j\}}(s) | f_{\{i_1, \dots, i_j\}}(s) \equiv l_{i_1} \wedge l_{i_2} \wedge \dots \wedge l_{i_j}, \{i_1, \dots, i_j\} \subseteq \{1, \dots, m\}\}$$

E.g. $m = 3$, $\mathcal{X} = \{0, 1\}^3$

Examples of instances: $s_1 = 101$, $s_2 = 001$, $s_3 = 100$, $s_4 = 111$

Examples of hypotheses: $h_1 \equiv l_2$, $h_2 \equiv l_1 \wedge l_2$, $h_3 \equiv \text{true}$, $h_4 \equiv l_1 \wedge l_3$,
 $h_5 \equiv l_1 \wedge l_2 \wedge l_3$

h_1 , h_2 , and h_5 are false for s_1 , s_2 and s_3 and true for s_4 ; h_3 is true for any instance; h_4 is true for s_1 and s_4 but false for s_2 and s_3



Hypothesis Space: Example 4

Conjunction of m positive literals

- Question 1: how many and which are the distinct hypotheses for $m = 3$?
 - Ans.(which): *true, $l_1, l_2, l_3, l_1 \wedge l_2, l_1 \wedge l_3, l_2 \wedge l_3, l_1 \wedge l_2 \wedge l_3$*
 - Ans.(how many): *8*
- Question 2: how many distinct hypotheses there are as a function of m ?
 - Ans.: *2^m , in fact for each possible bit of the input string the corresponding literal may occur or not in the logic formula, so:*

$$\underbrace{2 \cdot 2 \cdot 2 \cdots 2}_{m \text{ times}} = 2^m$$



Notions

- Target function: deterministic vs. stochastic
- Training/empirical error
- Test error vs. ideal error
- Inductive Bias implementation

Exercises

- Think about some possible binary classification, multi-class classification and regression tasks. What hypothesis spaces can be considered?
- Implement polynomial regression using the Ridge Regression method available in scikit-learn, see `sklearn.linear_model.Ridge()` and look at the behavior of the solution when changing the parameter α