MACHINE LEARNING

ERNEST YEUNG ERNESTYALUMNI@GMAIL.COM

Contents

Abstract. Everything about Machine Learning.

Part 1. Introduction

0.0.1. *Terminology.*
inputs $\equiv$ independent variables $\equiv$ predictors (cf. statistics) $\equiv$ features (cf. pattern recognition)
outputs $\equiv$ dependent variables $\equiv$ responses

cf. Chapter 2 Overview of Supervised Learning, Section 2.1 Introduction of Hastie, Tibshirani, and Friedman (2009) [1]

cf. Chapter 2 Overview of Supervised Learning, Section 2.2 Variable Types and Terminology of Hastie, Tibshirani, and Friedman (2009) [1]

0.0.2. *FinSet.*
The category FinSet $\in$ Cat is the category of all finite sets (i.e. Obj(FinSet) $\equiv$ all finite sets) and all functions in between them; note that FinSet $\subset$ Set [1]

Recall that the FinSet *skeletal* is

0.1. Supervised Learning.

cf. http://cs229.stanford.edu/notes/cs229-notes1.pdf

Consider data to belong to the category of all possible data:

$$\text{Data} \equiv \text{Dat} = (\text{Obj}(\text{Dat}), \text{MorDat}, 1, \circ), \qquad \text{Dat} \in \text{Cat}$$

Consider the **training set**:

$$\text{training set} := \{(x^{(i)}, y^{(i)}) | i = 1 \dots m, x^{(i)} \in \mathcal{X}, y^{(i)} \in \mathcal{Y}\}$$

where $\mathcal{X}$ is a manifold (it can be topological or smooth, EY:20160502 I don't know exactly because I need to check the topological and/or differential structure); $\mathcal{Y} \in \text{Obj}(\text{FinSet})$, or ($\mathcal{Y} \in \text{Obj}(\text{Top})$(or $\mathcal{Y} \in \text{Obj}(\text{Man})$)).

So training set $\subset \mathcal{X} \times \mathcal{Y} \in \text{Obj}(\text{Dat})$.

I propose that there should be a functor H that represents the "learning algorithm":

$$\text{Dat} \xrightarrow{H} \text{ML}$$

s.t.

$$\begin{aligned} & H : \mathcal{X} \times \mathcal{Y} \to \text{Hom}(\mathcal{X}, \mathcal{Y}) \\ & H(\text{training set}) = H(\{(x^{(i)}, y^{(i)}) | i = 1 \dots m\}) = h \end{aligned}$$

When $\mathcal{Y} \in \text{Obj}(\text{FinSet})$, *classification.*
When $\mathcal{Y} \in \text{Obj}(\text{Top})$ (or Obj(Man)), *regression.*

0.1.1. *Linear Regression.* Keeping in mind

$$\text{Dat} \xrightarrow{H} \text{ML}$$

Consider

$$\begin{aligned} & h : \mathbb{R}^p \to \text{Hom}(\mathcal{X}, \mathcal{Y}) \\ & h : \theta \mapsto h_\theta \end{aligned}$$

s.t.

$$h_\theta : \mathcal{X} \to \mathcal{Y}$$

so (possibly) $h \in \text{Obj}ML$ (or is h part of the functor H?)

Consider the cost function J

$$\begin{aligned} & J : \mathbb{R}^p \to \text{Hom}(\mathfrak{X} \times \mathfrak{Y}, \mathbb{R}) = C^\infty(\mathcal{X} \times \mathcal{Y}) \\ & J(\theta) = \frac{1}{2} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)})^2 \end{aligned}$$

Date: 24 avril 2016.
Key words and phrases. Machine Learning, statistical inference, statistical inference learning.

[1] nlab FinSet https://ncatlab.org/nlab/show/FinSet

0.1.2. *LMS algorithm (least mean square (or Widrow-Hoff learning rule)).* Define **gradient descent** algorithm:

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

with := being assignment (I'll use := for "define", in mathematical terms, use context to distinguish the 2), where α is the *learning rate*.

Rewriting the above,

$$\theta := \theta - \alpha \mathrm{grad} J(\theta)$$

where $\mathrm{grad} : C^\infty(M) \to \mathfrak{X}(M)$, with M being a smooth manifold.

This is *batch gradient descent*:

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) = \theta_j - \alpha \frac{\partial}{\partial \theta_j} \frac{1}{2} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)})^2 = \theta_j - \alpha \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) \left(\frac{\partial h_\theta(x^{(i)})}{\partial \theta} \right)$$

Simply notice how the entire training set of m rows is used.

I will expound on the so-called distinguished object $1 \xrightarrow{P} X$ on pp. 8, in Section 2 The Category of Conditional Probabilities of Culbertson and Sturtz (2013) [2] because it wasn't clear to me in the first place (the fault is mine; the authors wrote a very lucid and very fathomable, pedagogically-friendly exposition).

$\forall Y$ with indiscrete σ-algebra $\Sigma_Y = \{Y, \emptyset\}$

(remember, $((Y, \Sigma_Y), \mu_Y)$, $\mu_Y(\phi) = 0$, $\mu_Y(Y) = 1$),

$\exists !$ unique morphism in $\mathrm{Mor}\mathcal{P}$, $X \to Y$, since

$\forall P : X \to Y$, $P \in \mathrm{Mor}\mathcal{P}$, P_x must be a probability measure on Y, because

$$(X, \Sigma_X) \xrightarrow{P} (Y, \Sigma_Y)$$
$$P : \Sigma_Y \times X \to [0,1]$$
$$P(\cdot | x) : \Sigma_Y \to [0,1] \equiv \begin{matrix} P_x : \Sigma_Y \to [0,1] \text{ s.t.} \\ P_x(\emptyset) = 0, \, P_x(Y) = 1 \end{matrix}$$

i.e. EY: 20160503, Given $x \in X$ occurs, Y must occur.

By def. of terminal object ($\forall (X, \Sigma_X) \in \mathrm{Obj}\mathcal{P}$, $\exists !$ morphism P s.t. $(X, \Sigma_X) \xrightarrow{P} (Y, \Sigma_Y)$, Y *terminal* object, and denote unique morphism $!_X : X \to Y$, $!_X \in \mathrm{Mor}\mathcal{P}$.

Up to isomorphism, canonical terminal object is 1-element set denoted by $1 = \{*\}$, with the only possible σ-algebra ($\mu(*) = 1$, $\mu(\emptyset) = 0$),

$$\forall P : 1 \to X, \, P \in \mathrm{Mor}\mathcal{P}, \, P \in \mathrm{Hom}_{\mathcal{P}}(1, X), \, \forall X \in \mathrm{Mor}\mathcal{P}$$

P is an "absolute" probability measure on X because "there's no variability (conditioning) possible within singleton set $1 = \{*\}$." [2]

Now

$$P : \Sigma_X \times 1 \to [0,1]$$
$$P(\cdot | *) : \Sigma_X \to [0,1]$$

where $P(\cdot|*) : \Sigma_X \to [0,1]$ perfect probability measure on X, $P(\cdot|*) : \Sigma_X \to [0,1] \equiv P_*$, i.e. $P(\cdot|*) = p(\cdot)$ (usual probability on X).

$\forall A \in \Sigma_X$, $P(A|\cdot) : 1 \to [0,1]$, but $P(A|*) = P(A)$, $P(A|\emptyset) = 0$.

Refer to

$$1 \xrightarrow{P} X$$

morphism $P : 1 \to X \in \mathrm{Mor}\mathcal{P}$ as probability measure or distribution on X.

References

[1] Trevor Hastie, Robert Tibshirani, Jerome Friedman. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**, Second Edition (Springer Series in Statistics) 2nd ed. 2009. Corr. 7th printing 2013 Edition. ISBN-13: 978-0387848570. https://web.stanford.edu/~hastie/local.ftp/Springer/OLD/ESLII_print4.pdf

[2] Jared Culbertson, Kirk Sturtz. *Bayesian machine learning via category theory.* arXiv:1312.1445 [math.CT] http://arxiv.org/abs/1312.1445

[3] CS229 Stanford University. http://cs229.stanford.edu/materials.html