## Handout 1: Machine learning –A recap on: definitions, notation, and formalism

Lecturer & author: Georgios P. Karagiannis georgios.karagiannis@durham.ac.uk

**Aim.** To get some definitions and set-up about the learning procedure; essentially to formalize what introduced in term 1.

## Reading list & references:

- Bishop, C. M. (2006). Pattern recognition and machine learning. New York: Springer.
  - Ch. 1 Introduction
- Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge university press.
  - Ch. 1 Introduction

## 1. General Introductions and loose definitions

Pattern recognition is the automated discovery of patterns and regularities in data  $z \in \mathcal{Z}$ . Machine learning (ML) are statistical procedures for building and understanding probabilistic methods that 'learn'. ML algorithms  $\mathfrak{A}$  build a (probabilistic/deterministic) model able to make predictions or decisions with minimum human interference and can be used for pattern recognition. Learning (or training, estimation, fitting) is called the procedure where the ML model is tuned. Training data (or observations, sample data set, examples) is a set of observables  $\{z_i \in \mathcal{Z}\}$  used to tune the parameters of the ML model. By  $\mathcal{Z}$  we denote the examples (or observables) domain. Test set is a set of available examples/observables  $\{z_i'\}$  (different than the training data) used to verify the performance of the ML model for a given a measure of success. Measure of success (or performance) is a quantity that indicates how bad the corresponding ML model or Algorithm performs (eg quantifies the failure/error), and can also be used for comparisons among different ML models; eg, Risk function or Empirical Risk Function. Two main problems in ML are the supervised learning (we will focus on this here) and the unsupervised learning.

Supervised learning problems involve applications where the training data  $z \in \mathcal{Z}$  comprise examples of the input vectors  $x \in \mathcal{X}$  along with their corresponding target vectors  $y \in \mathcal{Y}$ ; i.e.  $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ . By  $\mathcal{X}$  we denote the inputs (or instances) domain, and by  $\mathcal{Y}$  we denote the target domain. Classification problems are those which aim to assign each input vector x to one of a finite number of discrete categories of y. Regression problems are those where the output y consists of one or more continuous variables. All in all, the learner wishes to discover an unknown pattern (i.e. functional relationship) between components  $x \in \mathcal{X}$  that serves as inputs and components  $y \in \mathcal{Y}$  that act as outputs; i.e.  $x \longmapsto y$ . Hence,  $\mathcal{X}$  is the input domain, and  $\mathcal{Y}$  is the output (or target) domain. The goal of learning is to discover a function which predicts (or help us make decisions about)  $y \in \mathcal{Y}$  from  $x \in \mathcal{X}$ .

1

Unsupervised learning problems involve applications where the training data  $z \in \mathcal{Z}$  consist of a set of input vectors  $x \in \mathcal{X}$  without any corresponding target values; i.e.  $\mathcal{Z} = \mathcal{X}$ . In clustering the goal is to discover groups of similar examples within the data of it is to discover groups of similar examples within the data.

2. (LOOSE) NOTATION & DEFINITIONS IN LEARNING

**Definition 1.** The learner's output is a function,  $h: \mathcal{X} \to \mathcal{Y}$  which predicts  $y \in \mathcal{Y}$  from  $x \in \mathcal{X}$ . It is also called Hypothesis, prediction rule, predictor, or classifier.

Notation 2. We often denote the set of hypothesis as  $\mathcal{H}$ ; i.e.  $h \in \mathcal{H}$ .

**Example 3.** (Linear Regression)<sup>1</sup> Consider the regression problem where the goal is to learn the mapping  $x \to y$  where  $x \in \mathcal{X} \subseteq \mathbb{R}^d$  and  $y \in \mathcal{Y} \subseteq \mathbb{R}$ . A hypothesis is a linear function  $h: \mathcal{X} \to \mathcal{Y}$ (that learner wishes to learn) with  $h(x) = \langle w, x \rangle$  approximating the mapping  $x \to y$ . The hypothesis set  $\mathcal{H} = \{x \to \langle w, x \rangle : w \in \mathbb{R}^d\}.$ 

**Example 4.** (Binary Classification) Consider the classification problem where the goal is to learn the mapping  $x \to y$  where  $x \in \mathcal{X} \subseteq \mathbb{R}^d$  and  $y \in \mathcal{Y}\{-1,+1\}$ . A hypothesis can be a function  $h: \mathcal{X} \to \mathcal{Y}$  with  $h(x) = \text{sign}(\langle w, x \rangle)$  approximating the mapping  $x \to y$ . The hypothesis set  $\mathcal{H} = \{ x \to \text{sign}(\langle w, x \rangle) : w \in \mathbb{R}^d \}.$ 

**Definition 5.** Training data set S of size m is any finite sequence of pairs  $(z_i = (x_i, y_i); i = 1, ..., m)$ in  $\mathcal{X} \times \mathcal{Y}$ ; i.e.  $\mathcal{S} = \{(x_i, y_i); i = 1, ..., m\}$ . This is the information that the learner has assess.

**Definition 6.** Data generation model  $q(\cdot)$  is the probability distribution over  $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ , unknown to the learner that has generated the data. E.g.  $z \sim g$ .

**Definition 7.** We denote as  $\mathfrak{A}(\mathcal{S})$  the hypothesis (outcome) that a learning algorithm  $\mathfrak{A}$  returns given training sample S.

**Definition 8.** (Loss function) Given any set of hypothesis  $\mathcal{H}$  and some domain  $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ , a loss function  $\ell(\cdot)$  is any function  $\ell: \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_+$ . Loss function  $\ell(h, z)$  for  $h \in \mathcal{H}$  and  $z \in \mathcal{Z}$  is specified according to the purpose the machine learning algorithm. It reflects how the "error" is quantified for a given hypothesis h and a given example z. The rule is "the greater the error the greater the value of the loss".

**Example 9.** (Cont. Example 3) In regression problems  $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$  and  $\mathcal{Y} \subset \mathbb{R}$  is uncountable, a potential loss function is

$$\ell_{\text{sq}}(h,(x,y)) = (h(x) - y)^2$$

**Example 10.** (Cont. Example 4) In binary classification problems with hypothesis  $h: \mathcal{X} \to \mathcal{Y}$ where  $\mathcal{Y} = \{0, 1\}$  is discrete, a loss function can be

$$\ell_{0-1}(h,(x,y)) = 1(h(x) \neq y),$$

 $<sup>1\</sup>langle w, x \rangle = w^{\top} x$ 

**Definition 11.** (Risk function) The risk function  $R_g(h)$  of h is the expected loss of the hypothesis  $h \in \mathcal{H}$ , w.r.t. the data generation model (which is a probability distribution) g over domain Z; i.e.

(2.1) 
$$R_{g}(h) = \mathcal{E}_{z \sim g}(\ell(h, z))$$

Remark 12. In learning, an ideal way to obtain an optimal predictor  $h^*$  is to compute the minimizer of the risk; i.e.

$$(2.2) h^* = \arg\min_{\forall h} \left( R_g \left( h \right) \right)$$

**Example 13.** (Cont. Ex. 9) The risk function is  $R_g(h) = \mathbb{E}_{z \sim g} (h(x) - y)^2$ , and it measures the quality of the hypothesis function  $h: \mathcal{X} \to \mathcal{Y}$ , (or equiv. the validity of the class of hypotheses  $\mathcal{H}$ ) against the data generating model g, as the expected square difference between the predicted values form h and the true target values g at every g.

Note 14. Computing the risk minimizer may be practically challenging due to the integration w.r.t. the unknown data generation model g involved in the expectation (2.1). Sub-optimally, one may use the Empirical risk function instead of the Risk function in (2.2).

**Definition 15.** (Empirical risk function) The Empirical Risk Function (ERF)  $\hat{R}_S(h)$  of h is the expectation of loss of h over a given sample  $S = (z_1, ..., z_m) \in \mathbb{Z}^m$ ; i.e.

$$\hat{R}_{S}(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_{i}).$$

Remark 16. Given Empirical Risk Function (ERF)  $R_S(h)$  of h the optimal predictor  $h^*$  is the minimizer of the ERF; i.e.

$$(2.3) h^* = \arg\min_{\forall h} \left( \hat{R}_S(h) \right)$$

**Example 17.** (Cont. Example 13) Given given sample  $S = \{(x_i, y_i); i = 1, ..., m\}$  the empirical risk function is  $\hat{R}_S(h) = \frac{1}{m} \sum_{i=1}^m (h(x_i) - y_i)^2$ .

**Example 18.** (Cont. Example 10) Given given sample  $S = \{(x_i, y_i); i = 1, ..., m\}$  the empirical risk function is  $\hat{R}_S(h) = \frac{1}{m} \sum_{i=1}^m \mathbb{1}(h(x_i) \neq y_i)$ .

Remark 19. If the Hypothesis set  $\mathcal{H}$  is a known parametric family of functions; i.e.  $\mathcal{H} = \{h_w(\cdot); w \in \mathcal{W}\}$  parameterized by unknown  $w \in \mathcal{W}$ , then we can equivalently consider  $\mathcal{H} = \{w \in \mathcal{W}\} = \mathcal{W}$  keeping in mind that the learner's output is restricted to  $h_w(\cdot)$ .

**Example 20.** Consider the multiple linear regression problem with regressors  $x \in \mathcal{X} \subseteq \mathbb{R}^d$  and response  $y \in \mathcal{Y} \subseteq \mathbb{R}$ . Because it involves only linear functions as predictors  $h_w(x) = \langle w, x \rangle$ , we could consider a hypothesis class  $\mathcal{H} = \{w \in \mathbb{R}^d\} = \mathbb{R}^d$  and loss function loss  $\ell(w, (x, y)) = (\langle w, x \rangle - y)^2$  for computational simplicity. The latter will be mainly used.

**Example 21.** Consider a learning problem where the true data generation distribution (unknown to the learner) is g(z), the statistical model (known to the learner) is given by a sampling distribution

 $f_{\theta}(y) := f(y|\theta)$  labeled by an unknown parameter  $\theta$ . The goal is to learn  $\theta$ . If we assume loss function

$$\ell(\theta, z) = \log\left(\frac{g(z)}{f_{\theta}(z)}\right)$$

then the risk is

$$(2.4) R_g(\theta) = E_{z \sim g}\left(\log\left(\frac{g(z)}{f_{\theta}(z)}\right)\right) = E_{z \sim g}\left(\log\left(g(z)\right)\right) - E_{z \sim g}\left(\log\left(f_{\theta}(z)\right)\right)$$

whose minimizer is

$$\theta^* = \arg\min_{\forall \theta} (R_g(\theta)) = \arg\min_{\forall \theta} (E_{z \sim g}(-\log(f_{\theta}(z))))$$

as the first term in (2.4) is constant. Note that in the Maximum Likelihood Estimation technique the MLE  $\theta_{\text{MLE}}$  is the minimizer

$$\theta_{\text{MLE}} = \underset{\theta}{\operatorname{arg \, min}} \left( \frac{1}{m} \sum_{i=1}^{m} \left( -\log \left( f_{\theta} \left( z_{i} \right) \right) \right) \right)$$

where  $S = \{z_1, ..., z_m\}$  is an IID sample from g. Hence, MLE  $\theta_{\text{MLE}}$  can be considered as the minimizer of the empirical risk  $R_S(\theta) = \frac{1}{m} \sum_{i=1}^m \left(-\log\left(f_\theta\left(z_i\right)\right)\right)$ .

**Definition 22.** A learning problem with hypothesis class  $\mathcal{H}$ , examples domain  $\mathcal{Z}$ , and loss function  $\ell$  may be denoted with a triplet  $(\mathcal{H}, \mathcal{Z}, \ell)$ .

**Example 23.** The standard multiple linear regression problem with regressors  $x \in \mathcal{X} \subseteq \mathbb{R}^d$  and response  $y \in \mathcal{Y} \subseteq \mathbb{R}$ , is a learning problem with examples domain  $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} = \mathbb{R}^{d+1}$ , hypothesis class  $\mathcal{H} = \{x \to \langle w, x \rangle : w \in \mathbb{R}^d\}$ , and loss function  $\ell_{\text{sq}}(h, (x, y)) = (h(x) - y)^2$ .

## APPENDIX A. USEFUL THINGS

Below are some standard notation used as default in the notes except in cases that is defined otherwise.

- q-norm: When  $x \in \mathbb{R}^d \|x\|_q := \left(\sum_{j=1}^d x_j^q\right)^{1/q}$
- Manhattan norm: When  $x \in \mathbb{R}^d \|x\|_1 := \sum_{j=1}^d |x_j|$
- Euclidean norm: When  $x \in \mathbb{R}^d \|x\|_2 := \sqrt{\sum_{j=1}^d x_j^2}$ . When  $\|\cdot\|$  we will assume the Euclidean norm.
- Infinity norm or maximum norm:  $\|x\|_{\infty} := \max_{\forall j} |x_j|$
- Inner product of x , y: If  $x,y \in \mathbb{R}^d$  then  $\langle x,y \rangle = x^\top y$ . So  $\langle x,x \rangle = \|x\|^2$

Also some standard formulas.

• Jensens' inequality: If If  $x \in \mathbb{R}^d$  and  $f : \mathbb{R}^d \to \mathbb{R}$  then

$$\begin{cases} f(\mathbf{E}(x)) \le \mathbf{E}(f(x)) & \text{if } f \text{ is convex} \\ f(\mathbf{E}(x)) \ge \mathbf{E}(f(x)) & \text{if } f \text{ is concave} \end{cases}$$

• Cauchy–Schwarz inequality: If  $x, y \in \mathbb{R}^d$  then  $|\langle x, y \rangle|^2 \leq \langle x, x \rangle \langle y, y \rangle$  equiv.  $|\langle x, y \rangle| \leq ||x|| \, ||y||$ .