Machine Learning and Neural Networks (MATH3431) Epiphany term, 2023

Handout 0: Learning problem: Definitions, notation, and formulation –A recap

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:

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

Supervised learning problems involve applications where the training data $z \in \mathcal{Z}$ comprises 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 recover 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 $y \in \mathcal{Y}$ from $x \in \mathcal{X}$.

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)[1] 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 $y \in \mathcal{Y} \subseteq \mathbb{R}$. Hypothesis is a linear function $h : \mathcal{X} \to \mathcal{Y}$ (that learner wishes to learn) to approximate mapping $x \to y$. The hypothesis set $\mathcal{H} = \left\{x \to \langle w, x\rangle : w \in \mathbb{R}^d\right\}$. We can use the loss $\ell\left(h, (x, y)\right) = \left(h(x) - y\right)^2$.

Definition 4. Training data set $\mathcal{S}$ of size m is any finite sequence of pairs $\left((x_i, y_i)\,; i = 1, ..., m\right)$ in $\mathcal{X} \times \mathcal{Y}$. This is the information that the learner has assess.

Definition 5. Data generation model $g\left(\cdot\right)$ is the probability distribution over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, unknown to the learner that has generated the data.

Definition 6. We denote as $\mathfrak{A}\left(\mathcal{S}\right)$ the hypothesis (outcome) that a learning algorithm $\mathfrak{A}$ returns given training sample S.

Definition 7. (Loss function) Given any set of hypothesis $\mathcal{H}$ and some domain $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, a loss function $\ell\left(\cdot\right)$ is any function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_+$. The purpose of loss function $\ell\left(h, z\right)$ is to quantify the "error" for a given hypothesis h and example z –the greater the error the greater its value of the loss.

Example 8. (Cont. Example 3) In regression problems $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $\mathcal{Y} \subset \mathbb{R}$ is uncountable, a loss function can be

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

Example 9. In binary classification problems with $h : \mathcal{X} \to \mathcal{Y}$ a learner where $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $\mathcal{Y} = \{0, 1\}$ is discrete, a loss function can be

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

Definition 10. (Risk function) The risk function $R_g\left(h\right)$ of h is the expected loss of the hypothesis $h \in \mathcal{H}$, w.r.t. probability distribution g over domain Z; i.e.

$$R_g\left(h\right) = \mathrm{E}_{z\sim g}\left(\ell\left(h, z\right)\right) \tag{2.1}$$

[1] $\langle w, x\rangle = w^\top x$

Remark 11. In learning, an ideal way to obtain an optimal predictor h^* is to compute the risk minimizer

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

Example 12. (Cont. Ex. 8) The risk function is $R_g\left(h\right) = \mathrm{E}_{z\sim g}\left(h(x) - y\right)^2$, 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 y at every x.

Note 13. 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 resort to the Empirical risk function.

Definition 14. (Empirical risk function) The empirical risk function $\hat{R}_S\left(h\right)$ of h is the expectation of loss of h over a given sample $S = (z_1, ..., z_m) \in \mathcal{Z}^m$; i.e.

$$\hat{R}_S\left(h\right) = \frac{1}{m}\sum_{i=1}^{m} \ell\left(h, z_i\right).$$

Example 15. (Cont. Example 12) Given given sample $S = \{(x_i, y_i)\,; i = 1, ..., m\}$ the empirical risk function is $\hat{R}_S\left(h\right) = \frac{1}{m}\sum_{i=1}^{m}\left(h(x_i) - y_i\right)^2$.

Example 16. 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 $f_\theta\left(y\right) := f\left(y|\theta\right)$ labeled by an unknown parameter θ. The goal is to learn θ. If we assume loss function

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

then the risk is

$$R_g\left(\theta\right) = \mathrm{E}_{z\sim g}\left(\log\left(\frac{g\left(z\right)}{f_\theta\left(z\right)}\right)\right) = \mathrm{E}_{z\sim g}\left(\log\left(g\left(z\right)\right)\right) - \mathrm{E}_{z\sim g}\left(\log\left(f_\theta\left(z\right)\right)\right) \tag{2.2}$$

whose minimizer is

$$\theta^* = \arg\min_{\forall\theta}\left(R_g\left(\theta\right)\right) = \arg\min_{\forall\theta}\left(\mathrm{E}_{z\sim g}\left(-\log\left(f_\theta\left(z\right)\right)\right)\right)$$

as the first term in (2.2) is constant. Note that in the Maximum Likelihood Estimation technique the MLE θ_{MLE} is the minimizer of

$$\theta_{\mathrm{MLE}} = \arg\min_{\forall\theta}\left(\frac{1}{m}\sum_{i=1}^{m}\left(-\log\left(f_\theta\left(z_i\right)\right)\right)\right)$$

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

Definition 17. A learning problem with hypothesis class $\mathcal{H}$, examples domain $\mathcal{Z}$, and loss function ℓ may be denoted with a triplet $(\mathcal{H}, \mathcal{Z}, \ell)$.

Example 18. Consider the multiple linear regression problem $x \longmapsto y$ where $x \in \mathcal{X} \subseteq \mathbb{R}^d$ and $y \in \mathcal{Y} \subseteq \mathbb{R}$. Till now, we set the learning problem $(\mathcal{H}, \mathcal{Z}, \ell)$ in the linear regression with hypothesis class $\mathcal{H} = \left\{x \to \langle w, x\rangle : w \in \mathbb{R}^d\right\}$, loss $\ell\left(h, (x, y)\right) = (h(x) - y)^2$, and examples domain $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} \in \mathbb{R}^d$ and $\mathcal{Y} \in \mathbb{R}$. Because learning problem involves only linear functions as predictors $h\left(x\right) = \langle w, x\rangle$, this learning problem could be defined equivalently with a hypothesis class $\mathcal{H} = \left\{w \in \mathbb{R}^d\right\}$ and loss function loss $\ell\left(w, (x, y)\right) = \left(\langle w, x\rangle - y\right)^2$. The latter will be mainly used.

Appendix A. Useful bits

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

$$\begin{cases} f\left(\mathrm{E}\left(x\right)\right) \leq \mathrm{E}\left(f\left(x\right)\right) & \text{if } f \text{ is convex} \\ f\left(\mathrm{E}\left(x\right)\right) \geq \mathrm{E}\left(f\left(x\right)\right) & \text{if } f \text{ is concave} \end{cases}$$

- 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. $|\langle x, y\rangle| \leq \|x\| \, \|y\|$.