CS229M/STATS214: Machine Learning Theory

Lecturer: Tengyu Ma — Lecture # 1
Scribe: Anusri Pampari and Gabriel Poesia — Jan 11th, 2021

1.1 Overview

In this lecture, we will introduce formulations that will be used to study topics for the first five weeks of this class. We will set up the standard supervised learning formulation, introduce and justify the empirical risk minimization paradigm with an asymptotic result. Under the assumption that the number of data points in our training set goes to infinity, our *excess risk* (defined later) goes to zero. In practice it is often unrealistic to assume that our training set is "infinite"; thus, in later lectures we will prove results assuming a finite training set.

1.2 Supervised learning

Let's start with the standard formalization of the key problem in machine learning: supervised learning. Informally, we have a dataset where each data point is associated with a label, and we want to use the data to learn a function that maps a data point to its label. Once learnt, such a function can be used to infer the labels of data points with unknown labels. More formally, we'll say our data points come from some input space $\mathcal{X}$ (e.g. images of birds), and labels belong to the output space $\mathcal{Y}$ (e.g. bird species). Suppose we are interested in a specific joint probability distribution p over $\mathcal{X} \times \mathcal{Y}$ (e.g. images of birds in North America), from which we draw a *training set*, i.e we draw a a set of n independent and identically distributed (i.i.d.) data points $(x^{(i)}, y^{(i)})$ from p. The goal of supervised learning is to learn a mapping (i.e. a function) from $\mathcal{X}$ to $\mathcal{Y}$ using the training data. Any such function $h : \mathcal{X} \to \mathcal{Y}$ is called a *predictor* (also *hypothesis* or *model*).

Given two predictors, how do we decide which is better? For that, we define a *loss function* over the predictions. There are several ways to define loss functions: for now, define a loss function l as a function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Intuitively, the loss function takes two labels, the prediction made by a model $\hat{y}$ and the true label y, and gives a number that captures how different the two labels are. We assume ℓ is non-negative, i.e $\ell(\hat{y}, y) \geq 0$. Then, the loss of a model h on an example (x, y) is $\ell(h(x), y)$, i.e. the difference (as measured by ℓ) between the prediction made by h and the true label.

With these definitions, we are able to formalize the problem of supervised learning. Precisely, we seek to find a model h that minimizes what we call the expected loss (or population risk):

$$L(h) \triangleq \mathop{\mathbb{E}}_{(x,y)\sim p}[\ell(h(x), y)]. \tag{1.1}$$

In the best possible case, we would find an h with expected loss 0, since that's the best possible we can do.

1.2.1 Examples: Regression and classification

Here are two examples of standard supervised learning:

- In the problem of *regression*, predictions are real numbers ($\mathcal{Y} = \mathbb{R}$). We would like predictions to be as close as possible to the real labels. A classical loss function that captures this is the squared error, $\ell(\hat{y}, y) = (\hat{y} - y)^2$.
- In the problem of *classification*, predictions are in a discrete set of k unordered classes $\mathcal{Y} = [k] = \{1, \cdots, k\}$. One possible classification loss is the $0-1$ loss: $\ell(\hat{y}, y) = \mathbb{1}(\hat{y} \neq y)$, i.e. 0 if the prediction is equal to the true label, and 1 otherwise.

1.2.2 Hypothesis families

In the formulation of supervised learning Section 1.2, we said we would like to find *any function* that minimizes population risk. However, in practice, we do not have a way of optimizing over arbitrary functions. Instead, we work within a more constrained set of functions $\mathcal{H}$, which we call the *hypothesis family* (or *hypothesis class*). Each element of $\mathcal{H}$ is a function $h : \mathcal{X} \to \mathcal{Y}$. Usually, we choose a set $\mathcal{H}$ that we know how to optimize over (e.g. linear functions, or neural networks).

Given one particular function $h \in \mathcal{H}$, we define the *excess risk* of h with respect to $\mathcal{H}$ as the difference between the population risk of h and the best possible population risk inside $\mathcal{H}$:

$$E(h) \triangleq L(h) - \inf_{g \in \mathcal{H}} L(g).$$

Generally we need more assumptions about a specific problem and hypothesis class to bound absolute population risk, hence we focus on bounding the excess risk. Toward the end of this lecture we will show a general bound on the excess risk in the asymptotic case.

Usually, the family we choose to work with can be parameterized by a vector of parameters $\theta \in \Theta$. In that case, we can refer to an element of $\mathcal{H}$ by h_θ, making that explicit. An example of such a parametrization of the hypothesis class is $\mathcal{H} = \{h : h_\theta(x) = \theta^T x, \theta \in \mathbb{R}^d\}$.

1.3 Empirical risk minimization

Our ultimate goal is to minimize population risk. However, in practice we do not have access to the entire population: we only have a *training set* of n data points, drawn from the same distribution as the entire population. While we cannot compute population risk, we can compute *empirical risk*, the loss over the training set, and try to minimize that. This is, in short, the paradigm known as *empirical risk minimization* (ERM): we optimize the training set loss, with the hope that this leads us to a model that has low population loss. From now on, with some abuse of notation, we often write $\ell(h_\theta(x), y)$ as $\ell((x, y), \theta)$ and use the two notations interchangeably. Formally, we define the empirical risk of a model h as:

$$\widehat{L}(h_\theta) \triangleq \frac{1}{n} \sum_{i=1}^{n} \ell(h_\theta(x^{(i)}), y^{(i)}) = \frac{1}{n} \sum_{i=1}^{n} \ell((x^{(i)}, y^{(i)}), \theta). \tag{1.2}$$

Empirical risk minimization means of finding the minimizer of $\widehat{L}$, which we call $\hat{\theta}$:

$$\hat{\theta} \triangleq \operatorname*{argmin}_{\theta \in \Theta} \widehat{L}(h_\theta). \tag{1.3}$$

Fortunately, since for now we're assuming that our training examples are drawn from the same distribution as the whole population, we know that empirical risk and population risk are equal *in expectation*:

$$\mathop{\mathbb{E}}_{(x^{(i)},y^{(i)})\sim p} \widehat{L}(h_\theta) = \mathop{\mathbb{E}}_{(x^{(i)},y^{(i)})\sim p} \frac{1}{n}\sum_{i=1}^{n} \ell(h_\theta(x^{(i)}), y^{(i)}) \tag{1.4}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathop{\mathbb{E}}_{(x^{(i)},y^{(i)})\sim p} \ell(h_\theta(x^{(i)}), y^{(i)}) \tag{1.5}$$

$$= \frac{1}{n} \cdot n \cdot \mathop{\mathbb{E}}_{(x^{(i)},y^{(i)})\sim p} \ell(h_\theta(x^{(i)}), y^{(i)}) \tag{1.6}$$

$$= L(h_\theta). \tag{1.7}$$

This is one reason why it makes sense to use empirical risk: it is an unbiased estimator of the population risk.

The key question that we seek to answer in the first part of this course is: **what guarantees do we have on the excess risk for the parameters learned by ERM?** The hope with ERM is that minimizing the training error will lead to small testing error. One way to make this rigorous is by showing that the ERM minimizer's excess risk is bounded.

1.4 Asymptotics of empirical risk minimization

In this section, we use an asymptotic approach (i.e assuming number of training samples $n \to \infty$) to achieve a bound on the ERM. In future lectures we will assume finite n and provide a non-asymptotic analysis.

For the asymptotic analysis of ERM, we would like to prove that excess risk is bounded as shown below:

$$L(\hat{\theta}) - \operatorname*{argmin}_{\theta\in\Theta} L(\theta) \leq \frac{c}{n} + o\left(\frac{1}{n}\right). \tag{1.8}$$

Here c is a problem dependent constant. The equation above shows that as we have more training data (i.e as n increases) the excess risk of ERM decrease at the rate of $\frac{1}{n}$.

Let $\{(x^{(1)}, y^{(1)}), \cdots, (x^{(n)}, y^{(n)})\}$ be the training data and let $\mathcal{H} = \{h_\theta : \theta = \mathbb{R}^p\}$ be the parameterized family of hypothesis functions. Let the ERM minimizer be $\hat{\theta}$ as defined in Equation (1.3). Let θ^* be the minimizer of the population risk L, i.e. $\theta^* = \operatorname{argmin}_\theta L(\theta)$. The theorem below quantifies the excess risk $L(\hat{\theta}) - L(\theta^*)$:

Theorem 1.1. *Suppose that (a) $\hat{\theta} \xrightarrow{p} \theta^*$ as $n \to \infty$ (i.e consistency of $\hat{\theta}$), (b) $\nabla^2 L(\theta^*)$ is full rank, and (c) other appropriate regularity conditions hold. Then,*

1. *$\sqrt{n}(\hat{\theta} - \theta^*) = O_P(1)$, i.e. for every $\epsilon > 0$, there is an M such that $\sup_n \mathbb{P}(\|\sqrt{n}(\hat{\theta} - \theta^*)\|_2 > M) < \epsilon$. (This means that the sequence $\{\sqrt{n}(\hat{\theta} - \theta^*)\}$ is "bounded in probability".)*
2. *$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}\left(0, (\nabla^2 L(\theta^*))^{-1} Cov(\ell((x, y), \theta^*))(\nabla^2 L(\theta^*))^{-1}\right)$.*
3. *$n(L(\hat{\theta}) - L(\theta^*)) = O_P(1)$.*

4. $n(L(\hat{\theta}) - L(\theta^*)) \xrightarrow{d} \frac{1}{2}||S||_2^2$ *where* $S \sim \mathcal{N}\left(0, (\nabla^2 L(\theta^*))^{-1/2} Cov(\ell((x,y),\theta^*))(\nabla^2 L(\theta^*))^{-1/2}\right)$.

Remark: In the theorem above, Parts 1 and 3 only show the rate or order of convergence, while Parts 2 and 4 define the limiting distribution for the random variables.

1.4.1 Key ideas of proof

We will prove the theorem above by applying the following main ideas:

1. Obtain an expression for the excess risk by Taylor expansion of the derivative of the empirical risk $\nabla\widehat{L}(\theta)$ around θ^*.
2. By the law of large numbers, we have that $\widehat{L}(\theta) \xrightarrow{p} L(\theta)$, $\nabla\widehat{L}(\theta) \xrightarrow{p} \nabla L(\theta)$ and $\nabla^2\hat{L}(\theta) \xrightarrow{p} \nabla^2 L(\theta)$ as $n \to \infty$.
3. Central limit theorem (CLT).

First, we state the CLT for i.i.d. means and a lemma that we will use in the proof.

Theorem 1.2 (Central Limit Theorem)**.** *Let* $X_1, \cdots, X_n$*, be i.i.d. random variables, where* $\widehat{X} = \frac{1}{n}\sum_{i=1}^n X_i$ *and the covariance matrix* Σ *is finite. Then, as* $n \to \infty$ *we have*

1. $\widehat{X} \xrightarrow{p} \mathbb{E}[X]$*, and*
2. $\sqrt{n}(\widehat{X} - \mathbb{E}[X]) \xrightarrow{d} \mathcal{N}(0, \Sigma)$*. In particular,* $\sqrt{n}(\widehat{X} - \mathbb{E}[X]) = O_P(1)$.

Lemma 1.3.

1. *If* $Z \sim N(0, \Sigma)$ *and* A *is a deterministic matrix, then* $AZ \sim N(0, A\Sigma A^T)$.
2. *If* $Z \sim N(0, \Sigma^{-1})$ *and* $Z \in \mathbb{R}^p$*, then* $Z^T\Sigma Z \sim \chi^2(p)$*, where* $\sim \chi^2(p)$ *is the chi-squared distribution with* p *degrees of freedom.*

1.4.2 Main Proof

We give heuristic arguments for Parts 1 and 2 here, and defer the proofs for Parts 3 and 4 to the next lecture. First, note that by definition, the gradient of the empirical risk at the empirical risk minimizer, $\nabla\widehat{L}(\hat{\theta})$, is equal to 0. From the Taylor expansion of $\nabla\widehat{L}$ around θ^*, we have that

$$0 = \nabla\widehat{L}(\hat{\theta}) = \nabla\widehat{L}(\theta^*) + \nabla^2\widehat{L}(\theta^*)(\hat{\theta} - \theta^*) + O(\|\hat{\theta} - \theta^*\|_2^2). \tag{1.9}$$

Rearranging, we have

$$\hat{\theta} - \theta^* = -(\nabla^2\widehat{L}(\theta^*))^{-1}\nabla\widehat{L}(\theta^*) + O(||\hat{\theta} - \theta^*||_2^2). \tag{1.10}$$

Multiplying by $\sqrt{n}$ on both sides,

$$\sqrt{n}(\hat{\theta} - \theta^*) = -(\nabla^2\widehat{L}(\theta^*))^{-1}\sqrt{n}(\nabla\widehat{L}(\theta^*)) + O(||\hat{\theta} - \theta^*||_2^2)) \tag{1.11}$$

$$\approx -(\nabla^2\widehat{L}(\theta^*))^{-1}\sqrt{n}(\nabla\widehat{L}(\theta^*)). \tag{1.12}$$

Applying the Central Limit Theorem (Theorem 1.2) using $X_i = \nabla \ell((x^{(i)}, y^{(i)}), \theta^*)$ and $\widehat{X} = \nabla \widehat{L}(\theta^*)$, and noticing that $\mathbb{E}[\nabla \widehat{L}(\theta^*)] = \nabla L(\theta^*)$, we have

$$\sqrt{n}(\nabla \widehat{L}(\theta^*) - \nabla L(\theta^*)) \xrightarrow{d} \mathcal{N}(0, Cov(\ell((x, y), \theta^*))). \tag{1.13}$$

Note that $\nabla L(\theta^*) = 0$ because θ^* is the minimizer of L, so $\sqrt{n}(\nabla \widehat{L}(\theta^*)) \xrightarrow{d} \mathcal{N}(0, Cov(\ell((x, y), \theta^*)))$. By the law of large numbers, $\nabla^2 \widehat{L}(\theta^*) \xrightarrow{p} \nabla^2 L(\theta^*)$. Applying these results to (1.12) (together with an application of Slutsky's theorem),

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} \nabla^2 L(\theta^*)^{-1} \mathcal{N}(0, Cov(\ell((x, y), \theta^*)) \tag{1.14}$$

$$\stackrel{d}{=} \mathcal{N}\left(0, \nabla^2 L(\theta^*)^{-1} Cov(\ell((x, y), \theta^*)) \nabla^2 L(\theta^*)^{-1}\right), \tag{1.15}$$

where the second step is due to Lemma 1.3. This proves Part 2 of Theorem 1.1.

Part 1 follows directly from Part 2 by the following fact: If $X_n \xrightarrow{d} P$ for some probability distribution P, then $X_n = O_P(1)$.