CSE 250: Machine Learning Theory Spring 2017

Part 5 - Introduction to Convex Learning

Instructor: Raef Bassily Scribe: Andrew Leverentz

5.1 Convex Learning

Now we will discuss some generalizations to our learning framework. As before, $\mathcal{X}$ is the domain set (i.e., a set of feature vectors). Usually, we will consider the case where $\mathcal{X} \subseteq \mathbb{R}^m$. Instead of restricting the label set $\mathcal{Y}$ to a finite set of labels, $\mathcal{Y}$ can now be an arbitrary set. Usually, we will consider the case $\mathcal{Y} \subseteq \mathbb{R}$. In this context, $\mathcal{Y}$ is called the "target set."

In this case, $\mathcal{X}$ and $\mathcal{Y}$ describe the components of the data, and the goal is to "learn" a functional relationship between the features ($x \in \mathcal{X}$) and the target ($y \in \mathcal{Y}$).

For example, we could have

$$x = (\text{heart rate}, \text{blood pressure}, \ldots) \in \mathbb{R}^m,$$
$$y = \text{person's weight} \in \mathbb{R}.$$

As before, the hypothesis class $\mathcal{H}$ is a collection of functions from $\mathcal{X}$ to $\mathcal{Y}$. We will be looking at classes where each hypothesis $h \in \mathcal{H}$ is *uniquely* described by a real vector w of fixed length d; that is,

$$w \in C \subseteq \mathbb{R}^d,$$
$$\mathcal{H} = \{h_w : w \in C\}.$$

Here, C is called the parameter set. Each hypothesis is uniquely described by $w \in C$; moreover, there is a one-to-one correspondence between the hypothesis class $\mathcal{H}$ and the parameter set C. Sometimes, we will abuse notation and refer to $w \in C$ as the hypothesis and C as the hypothesis class.

Definition 5.1 (Loss function)**.** Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. We will consider loss functions of the form

$$\ell : C \times \mathcal{Z} \to \mathbb{R}_+.$$

Equivalently, ℓ maps $\mathcal{H} \times \mathcal{Z}$ to a non-negative number.

Example 5.2 (Ordinary Least Squares Regression)**.** Suppose we have

$$\mathcal{X} = \mathbb{R}^m,$$
$$\mathcal{Y} = \mathbb{R}.$$

Let $\tilde{x} = (x, 1) \in \mathbb{R}^{m+1}$ be a vector where the constant 1 has been appended to the end of $x \in \mathbb{R}^m$. Then, consider the hypothesis class

$$\mathcal{H} = \{h_w : \forall\, x \in \mathbb{R}^m, h_w(x) = \langle w, \tilde{x} \rangle, w \in \mathbb{R}^{m+1}\},$$
$$C = \mathbb{R}^{m+1}.$$

Here, $\langle \cdot, \cdot \rangle$ is the inner product.

Our loss function is squared loss:

$$\ell : \underbrace{\mathbb{R}^{m+1}}_{C} \times \underbrace{\mathbb{R}^{m+1}}_{\mathcal{Z}} \to \mathbb{R}_+,$$
$$\ell(w, (x, y)) = (\langle w, \tilde{x} \rangle - y)^2.$$

Remark 5.3 (Empirical Risk Minimization, revisited)**.** If we are given a training set $S = \{(x_i, y_i)\}_{i=1}^n$, the ERM notion still holds:

$$w^* = \operatorname*{argmin}_{w \in \mathbb{R}^{m+1}} \frac{1}{n} \sum_{i=1}^{n} \ell(w, (x_i, y_i)).$$

This optimization problem is "easy," in the sense that it can be solved efficiently, since C is a convex set, and for every $(x, y) \in \mathcal{Z}$, $\ell(\cdot, (x, y))$ is a convex function over C. Hence, it is a convex optimization problem. Under these conditions, there is a unique global minimum.

Next, we will define a generalization of agnostic PAC learnability. The notion of VC dimension will not apply, because the VC dimension requires binary labels.

Recall the setup for regression:

$$\mathcal{X} \subseteq \mathbb{R}^m,$$
$$\mathcal{Y} \subseteq \mathbb{R},$$
$$\mathcal{H} = \{h_w : \mathcal{X} \to \mathcal{Y}\}, \text{ where each } h_w \text{ is parameterized by } w \in C \subseteq \mathbb{R}^d.$$

The loss function is of the form $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$.

For example, regression can use squared loss:

$$\ell(w, (x, y)) = (w \cdot \tilde{x} - y)^2.$$

Definition 5.4 (Generalized Agnostic PAC Learnability)**.** A class $\mathcal{H}$ is agnostic PAC-learnable with respect to $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and the loss function ℓ in the generalized model if there is a function

$$n_{\mathcal{H}} : (0, 1)^2 \to \mathbb{N}$$

and a learning algorithm A such that:

- for every $0 < \varepsilon, \delta < 1$, and
- for every distribution $\mathcal{D}$ over $\mathcal{Z}$,

when running A on a set S of $n \geq n_{\mathcal{H}}(\varepsilon, \delta)$ i.i.d training examples from $\mathcal{D}$, then w.p. $\geq 1 - \delta$, A returns a hypothesis $h_S \in \mathcal{H}$ with

$$\mathbb{E}_{z \sim \mathcal{D}}[\ell(h_S, z)] \leq \inf_{h \in \mathcal{H}} \mathbb{E}_{z \sim \mathcal{D}}[\ell(h, z)] + \varepsilon.$$

Remark 5.5. Note that the quantity $\mathbb{E}_{z \sim \mathcal{D}}[\ell(h_S, z)]$ is a random variable, since it depends on S, and the expectation is not taken over the choice of S.

Remark 5.6. A few observations:

- Learnability is hard to reason about in this general model.
- The notion of VC dimension is not necessarily valid for regression problems, since it deals with set-shattering and binary labelings.
- However, under certain restrictions on ℓ and $\mathcal{H}$, learnability can be achieved.
- There exists an analogue of VC dimension called *fat-shattering dimension* for regression, but there is no full characterization of learnability using this notion.

Remark 5.7. We will study this model under the following assumptions (hence the name "convex learning"):

- For all $z \in \mathcal{Z}$, $\ell(\cdot, z)$ is a convex loss function over the parameter set C.
- The parameter set C is a convex set.

5.2 Preliminaries on Convex Analysis

Definition 5.8 (Convex set)**.** A set C in a vector space is convex if for any $u, v \in C$, and any $\lambda \in [0, 1]$,

$$\lambda u + (1 - \lambda)v \in C.$$

In other words, convex combinations of points from C also lie in C. For example: line segments joining points from C also lie entirely in C.

Definition 5.9 (Convex function)**.** Let C be a convex set. A function $f : C \to \mathbb{R}$ is convex if for all $u, v \in C$ and for all $\lambda \in [0, 1]$,

$$f(\lambda u + (1 - \lambda)v) \leq \lambda f(u) + (1 - \lambda)f(v).$$

Furthermore, f is called "strictly convex" if this inequality is strict when $u \neq v$ and $\lambda \in (0, 1)$. (For example: x^2 is strictly convex but $|x|$ is not.)

Remark 5.10. An equivalent definition: $f : C \to \mathbb{R}$ is convex if for every $u \in C$ there is a vector in C denoted by $\partial f(u)$ s.t. for all $v \in C$, we have

$$f(v) \geq f(u) + \langle \partial f(u), v - u \rangle.$$

Then f is strictly convex if the inequality is strict for all $v \neq u$.

- This is analogous to the fact that the tangent line of a convex function always lies at or below the function itself.
- Also, when f is convex and differentiable, the first order Taylor expansion (the RHS of the inequality above) is always $\leq$ the function value.
- The vector $\partial f(u)$ is called a *subgradient* of f at u.
- There can be more than one subgradient at any given point. However, if f is differentiable, then there is a unique subgradient at any given point x, called the *gradient* of the function at this point $\nabla f(x)$.

The gradient of a differentiable function at point $x \in \mathbb{R}^d$ is given by

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_d} \end{bmatrix}.$$

Fact 5.11. Let C be a convex set. If $f : C \to \mathbb{R}$ is convex, then f has a unique minimum over C, but not necessarily a unique minimizer. (For example, the function may have a flat "trough" at the minimum.)

If the function is strictly convex, then there is a unique minimum and a unique minimizer.

Fact 5.12. Let $f_i : \mathbb{R}^d \to \mathbb{R}$ for $i = 1, \ldots, r$ be convex functions, then

- $g(x) = \max_{i \in [r]} f_i(x)$ is convex, and
- $g(x) = \sum_{i \in [r]} \gamma_i f_i(x)$, where $\gamma_i \geq 0$ for all i, is convex.

Given $f_1 : \mathbb{R}^d \to \mathbb{R}$ and $f_i : \mathbb{R} \to \mathbb{R}$ for $i = 2, \ldots, r$, the composition

$$f_r \circ f_{r-1} \circ \cdots \circ f_2 \circ f_1$$

is convex if each individual function is convex and *one* of the following conditions holds:

- $f_{r-1}, \ldots, f_1$ are affine,
- $f_r, f_{r-1}, \ldots, f_2$ are non-decreasing,
- there is a $j \in \{2, \ldots, r-1\}$ s.t. $f_{j-1}, \ldots f_1$ are affine, and $f_r, f_{r-1}, \ldots, f_{j+1}$ are non-decreasing.