3.1-3.2: Overview and Formal Setup

1. Overview

- Central question: Why training error $\downarrow \Rightarrow$ test error $\downarrow$?

Two deficiencies of asymptotics analysis:

(i) No direct instruction about how large n should be.

(ii) Only apply to unregularized estimators on smooth loss functions.

This unit develop a new suite of tools to answer "How can we generalize to other problems and estimators.

2. Formal setup (supervised learning)

- Problem: to predict an output $y \in \mathcal{Y}$ given $x \in \mathcal{X}$.
- Hypothesis: $\mathcal{H} = \{ h \mid h: \mathcal{X} \to \mathcal{Y} \}$
- Loss function: $\ell: (\mathcal{X} \times \mathcal{Y}) \times \mathcal{H} \to \mathbb{R}$
- Let p^* denote the true underlying data-generating distribution over $\mathcal{X} \times \mathcal{Y}$.

[Definition 4] expected risk $L(h)$: (test error)

$$L(h) := \mathbb{E}_{(x,y)\sim p^*}[\ell((x,y),h)]$$

$h^* \in \arg\min_{h \in \mathcal{H}} L(h)$ is called the expected risk minimizer.

- training examples: $(x^{(1)}, y^{(1)}), \cdots, (x^{(n)}, y^{(n)})$ are a set of input-output pairs, each of which is drawn i.i.d. from p^*.

Note: (i) training and test distributions are same.

(ii) the independence assumption ensures more training data gives us more information.

(while (i) & (ii) seem impossibly to be practical.)

[Definition 5] empirical risk $\hat{L}(h)$: (training error)

$$\hat{L}(h) := \frac{1}{n} \sum_{i=1}^{n} \ell((x^{(i)}, y^{(i)}), h)$$

$\hat{h} \in \arg\min_{h \in \mathcal{H}} \hat{L}(h)$ is called the empirical risk minimizer. (ERM)

Note: (i) $\hat{h}$ is r.v. but h^* is non-random.

(ii) we are interested in $L(\hat{h})$ - two questions:

① quantify $L(\hat{h}) - \hat{L}(\hat{h})$

② quantify $L(\hat{h}) - L(h^*)$ (excess risk)

How can we analyse the excess risk?

Constraints: (i) $\hat{h}$ is r.v., excess risk could be high.

(ii) n is small finite, CLT can't work.

Idea: use concentration to show that bad outcomes are not too likely.

Probably Approximately Correct (PAC) framework:

An algorithm A returns $\hat{h} \in \mathcal{H}$ s.t. $L(\hat{h}) - L(h^*) < \varepsilon$ with prob. $\geq 1-\delta$ and runs in $\mathrm{Poly}(n, \mathrm{size}(x), 1/\varepsilon, 1/\delta)$ time.