Machine Learning Talk X

Learning Frameworks

Axel G. R. Turnquist

NJIT Department of Mathematical Sciences

December 4, 2020

Machine learning is fundamentally about: **generalization**. Task: choose

1. A hypothesis set (approximation error)
2. A specific function in that set (estimation error)

$$R(h) - R^* = \left(R(h) - \inf_{h \in \mathcal{H}} R(h)\right) + \left(\inf_{h \in \mathcal{H}} R(h) - R^*\right) \qquad (1)$$

- How does one minimize the first one?
- How does one minimize the second one?

Main Question: Is the error ever small?

Mathematical Formalism

Definitions:

- $\mathcal{X}$ set of examples or instances
- $\mathcal{Y}$ labels or target values $\mathcal{Y} = \{0, 1\}$
- Concept class $\mathcal{C}$ what you desire to learn
- Hypothesis set $\mathcal{H}$

Assume examples are i.i.d. with law $\mathcal{D}$.

Learning Problem: Learner considers a fixed set $\mathcal{H}$, which may or may not coincide with $\mathcal{C}$. Receives sample $S = (x_1, \ldots, x_m)$, which is drawn i.i.d. according to $\mathcal{D}$ as well labels $(c(x_1), \ldots, c(x_m))$, where $c \in \mathcal{C}$. Task is to use S to learn $h_S \in \mathcal{H}$, that has a small generalization error with respect to c

Generalization Error

Given $h \in \mathcal{H}$, a target concept $c \in \mathcal{C}$, an underlying distribution D, the generalization error of h is defined by:

$$R(h) = \mathbb{P}_{x \sim \mathcal{D}}[h(x) \neq c(x)] = \mathbb{E}_{x \sim \mathcal{D}}[1_{h(x) \neq c(x)}] \tag{2}$$

But, $\mathcal{D}$ and c are unknown. One can measure the empirical error:

$$\hat{R}_S(h) = \frac{1}{m} \sum_{i=1}^{m} 1_{h(x_i) \neq c(x_i)} \tag{3}$$

There are a number of guarantees that relate these two quantities with high probability.

PAC Learning

A concept class $\mathcal{C}$ is said to be **PAC-learnable** if there exists an algorithm $\mathcal{A}$ and a polynomial function $\text{poly}(\cdot,\cdot,\cdot,\cdot)$ such that for any $\epsilon > 0$ and $\delta > 0$, for all distributions $\mathcal{D}$ on $\mathcal{X}$ and for any target concept $c \in \mathcal{C}$, the following holds for any sample size $m \geq \text{poly}(1/\epsilon, 1/\delta, n, \text{size}(c))$:

$$\mathbb{P}_{S \sim \mathcal{D}^m}[R(h_S) \leq \epsilon] \geq 1 - \delta \tag{4}$$

Note that training and test samples are drawn from the same distribution. This learnability is related to $\mathcal{C}$, which is known, but $c \in \mathcal{C}$ which is unknown.

Hypothesis complexity

If we have an algorithm that returns a consistent hypothesis, i.e. $\hat{R}_S(h_S) = 0$ for any concept $c \in \mathcal{H}$, then if the hypothesis set $\mathcal{H}$ has finite cardinality, the concept class is PAC-learnable provided that the sample size satisfies:

$$m \geq \frac{1}{\epsilon}\left(\log|\mathcal{H}| + \log\frac{1}{\delta}\right) \tag{5}$$

Q: What if the cardinality of the hypothesis set is infinite?
A: We have to examine some exotic concepts...

Radamacher Complexity & VC Dimension

Rademacher complexity: ability of family of functions to correlate with noise. This concept seems related somehow to amount of information.

VC-Dimension: Largest size of set that can be shattered.

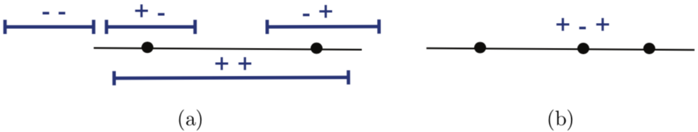

Figure 3.1
VC-dimension of intervals on the real line. (a) Any two points can be shattered. (b) No sample of three points can be shattered as the $(+, -, +)$ labeling cannot be realized.

Bounds

Using these concepts, we can derive nice asymptotic bounds using concentration inequalities like Hoeffding. Let $\mathcal{H}$ have VC-dimension d. Then, for any $\delta > 0$, with probability at least $1-\delta$, the following holds for all $h \in \mathcal{H}$:

$$R(h) \leq \hat{R}_S(h) + \mathcal{O}\left(\sqrt{\frac{\log(m/d)}{(m/d)}}\right) \tag{6}$$

- Too many samples with a simple hypothesis set? Not very generalizable.
- Not many samples and a complex hypothesis set? Not very generalizable.

Generalizations

In reality, the distribution $\mathcal{D}$ is over $\mathcal{X} \times \mathcal{Y}$, meaning that even the labeling is unreliable to some extent. For example, input height, output gender. Then, we instead define:

$$R(h) = \mathbb{P}_{(x,y) \in \mathcal{D}}[h(x) \neq y] \tag{7}$$

PAC learning:

$$\mathbb{P}_{S \sim \mathcal{D}^m}[R(h_S) - \min_{h \in \mathcal{H}} R(h) \leq \epsilon] \geq 1 - \delta \tag{8}$$

Note, the deterministic case guarantees $\exists h$ s.t. $R(h) = 0$

Bayes Hypothesis

- We define the infimum over all measurable functions h_{Bayes}. This is called the **Bayes hypothesis**.
- The error of the Bayes hypothesis at a point $x \in \mathcal{X}$ is called the **noise**. This is pretty unavoidable.

For example, perhaps given an age $x = 40$ years old, can we predict if it's a man or a woman? No, too noisy. Given the age $x = 110$, can we? Most likely a woman.

Reducing the Empirical Error

$$R(h) - R^* = \left(R(h) - \inf_{h \in \mathcal{H}} R(h) \right) + \left(\inf_{h \in \mathcal{H}} R(h) - R^* \right) \tag{9}$$

- What if we pick a very rich hypothesis set $\mathcal{H}$? Then, second term, the approximation error, is small, but the first term, the estimation error, is large for a fixed h.
- If we pick a simple $\mathcal{H}$? Then, the first term is easy to make small, but the second term is usually not small.

Estimation Error & Approximation Error

- Since $R(h)$ can be bounded by the empirical error $\hat{R}(h)$, bounding the first term, the estimation error is akin to reducing the empirical error (**empirical risk management**). Theoretically can be bounded well by having a large sample and a small complexity (Rademacher or VC-dimension). In practice, the bound is usually poor.
- Another way is to pick a hypothesis that balances the estimation and approximation errors (**structural risk management**).
- In practice? Use **cross-validation**, setting aside part of the training sample as a validation set. This gives nice bounds.

Questions?

Highlighted Resources

- "**Foundations of Machine Learning**" Mohri, Mehryar; Rostamizadeh, Afshin; Talwalkar, Ameet

Future Talks

Next Talk:

Dec. 11: Yuexin Liu
Reinforcement Learning