Computational Learning Theory

- Brief Overview of Machine Learning
- Consistency Model
- Probably Approximately Correct Learning
- Sample Complexity and Occam's Razor
- **Dealing with Noises and Inconsistent Hypotheses**
- ...

Problems with PAC

What we have seen so far isn't realistic:

- There may not be any $h \in \mathcal{H}$ such that $h = c$, thus, there will be examples which we can't find a consistent h
- There may be some $h \in \mathcal{H}$ such that $h = c$, but the problem of finding a consistent h (with examples) is $\mathbf{NP}$-hard
- In practices, examples are noisy. There might be some x labelled with both 0 and 1. Some "true" label might be flipped due to noise.
- There may not be any c at all!

Conclusions

Have to relax the model:

- Allow outputting h inconsistent with examples
- Measure h's performance somehow, even when c does not exist!

A New Model: Inconsistent Hypothesis Model

- In this model, $(\mathbf{x}, y)$ drawn from $\Omega \times \{0, 1\}$ according to some unknown distribution $\mathcal{D}$
- "Quality" of a hypothesis h is measured by

$$\mathsf{err}_{\mathcal{D}}(h) := \underset{(\mathbf{x},y)\leftarrow\mathcal{D}}{\mathsf{Prob}}[h(\mathbf{x}) \neq y]$$

(We will drop the subscript $\mathcal{D}$ when there's no confusion.)

- $\mathsf{err}(h)$ is called the *true error* of h

The Problem in the Ideal Case

Find $h^* \in \mathcal{H}$ whose $\mathsf{err}(h*)$ is minimized, i.e.

$$h^* = \underset{h\in\mathcal{H}}{\operatorname{argmin}}\, \mathsf{err}(h).$$

- But, we don't know $\mathcal{D}$, and thus can't even evaluate the objective function $\mathsf{err}(h)$

Bayes Optimal Classifier

- But suppose we do know $\mathcal{D}$, what is the best possible classifier? (There might be more than one.)
- The following is called the *Bayes optimal classifier*

$$h_{\text{OPT}}(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathsf{Prob}[y = 1 \mid \mathbf{x}] \geq 1/2 \\ 0 & \text{if } \mathsf{Prob}[y = 0 \mid \mathbf{x}] < 1/2 \end{cases}$$

  Question: why is it optimal?
- err(h_{OPT}) is called the *Bayes error*, which is an absolute lowerbound on any err(h)
- Note that h_{OPT} may not belong to $\mathcal{H}$, and thus h^* may be different from h_{OPT}

Empirical Error

- Since we don't know $\mathcal{D}$: find another function approximating $\text{err}(h)$ well, and find h minimizing that function instead!
- Let $\widehat{\text{err}}(h)$ be the fraction of examples wrongly labelled by h. Specifically, suppose $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)$ are the examples, let

$$\widehat{\text{err}}(h) = \frac{|\{i : h(\mathbf{x}_i) \neq y_i\}|}{m}$$

- We will prove that, with enough examples, $\widehat{\text{err}}(h) \approx \text{err}(h)$ with high probability. This is called the *uniform convergence* theorem.

The Real Problem

Find $h \in \mathcal{H}$ whose *empirical error* $\widehat{\text{err}}(h)$ is minimized.

Chernoff-Hoeffding Bound

(We've seen the "multiplicative" version of Chernoff, here's the "additive" version.)

Suppose X_i, $i \in [m]$ are i.i.d. Bernoulli variables with $\text{Prob}[X_i = 1] = p$. Let

$$\hat{p} = \frac{X_1 + \cdots + X_m}{m}$$

Then, for any $\epsilon > 0$,

$$\text{Prob}[\hat{p} \geq p + \epsilon] \leq e^{-2\epsilon^2 m}$$

and

$$\text{Prob}[\hat{p} \leq p - \epsilon] \leq e^{-2\epsilon^2 m}$$

Thus,

$$\text{Prob}[|\hat{p} - p| \geq \epsilon] \leq 2e^{-2\epsilon^2 m}$$

Uniform Convergence Theorem

Theorem

Suppose the hypothesis class $\mathcal{H}$ is finite. If we take

$$m \geq \frac{\log\left(\frac{2|\mathcal{H}|}{\delta}\right)}{2\epsilon^2}$$

examples, then

$$\text{Prob}\left[|\text{err}(h) - \widehat{\text{err}}(h)| \leq \epsilon,\ \text{for all } h \in \mathcal{H}\right] \geq 1 - \delta.$$

There's also a VC-dimension version of this theorem.
Proof idea:

- $\text{E}_S[\widehat{\text{err}}(h)] = \text{err}(h)$
- Apply Chernoff-Hoeffding and union bounds

Observations from the Uniform Convergence Theorem

- Note the dependence on ϵ^2, instead of ϵ as in Valiant's theorem
- Suppose
$$\hat{h}^* = \operatorname*{argmin}_{h \in \mathcal{H}} \widehat{\mathsf{err}}(h)$$
- Recall
$$h^* = \operatorname*{argmin}_{h \in \mathcal{H}} \mathsf{err}(h)$$
- We really want h^*, but don't know $\mathcal{D}$, and thus settled for $\hat{h}^*$ instead
- How good is $\hat{h}^*$ compared to h^*? By uniform convergence theorem,
$$\mathsf{err}(\hat{h}^*) \leq \widehat{\mathsf{err}}(\hat{h}^*) + \epsilon \leq \widehat{\mathsf{err}}(h^*) + \epsilon \leq \mathsf{err}(h^*) + 2\epsilon.$$
- The true error of $\hat{h}^*$ is not too far from the true error of the best hypothesis! (Even though we only minimize the empirical error.)