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
- ...

©Hung Q. Ngo (SUNY at Buffalo) CSE 694 – A Fun Course
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 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, \((x, y)\) drawn from \(\Omega \times \{0, 1\}\) according to some unknown distribution \(D\)
- “Quality” of a hypothesis \(h\) is measured by

\[
\text{err}_D(h) := \text{Prob}_{(x, y) \leftarrow D} \left[ h(x) \neq y \right]
\]

(We will drop the subscript \(D\) when there’s no confusion.)
- \(\text{err}(h)\) is called the true error of \(h\)

The Problem in the Ideal Case

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

\[
h^* = \arg\min_{h \in \mathcal{H}} \text{err}(h).
\]

- But, we don’t know \(D\), and thus can’t even evaluate the objective function \(\text{err}(h)\)
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**

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

- **Question:** why is it optimal?

- \( \text{err}(h_{\text{OPT}}) \) is called the **Bayes error**, which is an absolute lowerbound on any \( \text{err}(h) \)

- Note that \( h_{\text{OPT}} \) may not belong to \( \mathcal{H} \), and thus \( h^* \) may be different from \( h_{\text{OPT}} \)
Since we don’t know $\mathcal{D}$: find another function approximating $\text{err}(h)$ well, and find $h$ minimizing that function instead!

Let $\hat{\text{err}}(h)$ be the fraction of examples wrongly labelled by $h$. Specifically, suppose $(x_1, y_1), \ldots, (x_m, y_m)$ are the examples, let

$$\hat{\text{err}}(h) = \frac{|\{i : h(x_i) \neq y_i\}|}{m}$$

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

**The Real Problem**

Find $h \in \mathcal{H}$ whose *empirical error* $\hat{\text{err}}(h)$ 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

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

examples, then

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

There’s also a VC-dimension version of this theorem.

Proof idea:

- $E_S[\hat{err}(h)] = err(h)$
- Apply Chernoff-Hoeffding and union bounds
Observations from the Uniform Convergence Theorem

- Note the dependence on $\epsilon^2$, instead of $\epsilon$ as in Valiant’s theorem
- Suppose
  \[ \hat{h}^* = \arg\min_{h \in \mathcal{H}} \hat{\text{err}}(h) \]
- Recall
  \[ h^* = \arg\min_{h \in \mathcal{H}} \text{err}(h) \]
- 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,
  \[ \text{err}(\hat{h}^*) \leq \hat{\text{err}}(\hat{h}^*) + \epsilon \leq \hat{\text{err}}(h^*) + \epsilon \leq \text{err}(h^*) + 2\epsilon. \]
- 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.)