15-859(B) Machine Learning Theory

Lecture 02/12/02, Avrim Blum

Uniform convergence: the proofs

Recap

Let $err(h)$ = true error of h, $\widehat{err}(h)$ = empirical error of h. m = sample size.

- From first principles: if $err(h) \geq \varepsilon$, then

$$\Pr\left[\widehat{err}(h) = 0\right] \leq (1 - \varepsilon)^m.$$

- Set rhs to $\delta/|C|$ and solve for m. After

$$m = \frac{1}{\varepsilon}\left[\ln|C| + \ln\frac{1}{\delta}\right]$$

examples, whp *all* $h \in C$ of true error $\geq \varepsilon$ have empirical error > 0.

- From Hoeffding:

$$\Pr\left[|\widehat{err}(h) - err(h)| \geq \varepsilon\right] \leq 2e^{-2m\varepsilon^2}.$$

- Set rhs to $\delta/|C|$ and solve for m. After

$$m = \frac{2}{\varepsilon^2}\left[\ln|C| + \ln\frac{2}{\delta}\right]$$

examples, whp *all* $h \in C$ have $|\widehat{err}(h) - err(h)| \leq \varepsilon$.

Drawbacks

- Use of the union bound makes this loose when many hypothesis in C are very similar.
- This is especially bad for continuous hypothesis spaces.

What we'll prove today

Let $C[S]$ be the set of splittings of dataset S using concepts in C, and let $C[m] = \max_{|S|=m} |C[S]|$.

- **Theorem 2:** For any class C, distrib. D, if
$$m > \frac{2}{\varepsilon}[\log_2(2C[2m]) + \log_2(1/\delta)]$$
then with prob. $(1-\delta)$, all $h \in C$ with $err(h) > \varepsilon$ have $\widehat{err}(h) > 0$.

- **Theorem 2':** For any class C, distrib. D, if
$$m > \frac{2}{\varepsilon^2}[\ln(2C[2m]) + \ln(1/\delta)]$$
then with prob. $(1-\delta)$, all $h \in C$ have
$$|err(h) - \widehat{err}| \leq \varepsilon.$$

- **Theorem 3:** Can replace bound in Theorem 2 with: $O(\frac{1}{\varepsilon}[VCdim(C)\log(1/\varepsilon) + \log(1/\delta)])$.

For the proofs, let's go to the board....