The Agnostic Setting

The labeling function may be noisy so we don't have a function $f: X \to Y$

Instead we have a joint distribution on $X \times Y$.

This is called the agnostic setting or the unrealizable setting.

We have to redefine L and L_S

We define $err(h) = L(h) = \mathbb{P}_{(x,y)\sim D}(h(x) \neq y)$

Note this subsumes the case where the labelling is done by a function.

Previously we tried to make $L(h)$ as small as possible by choosing large enough samples. This may no longer be possible, for example if the labelling "function" is noisy.

We will try to get close to the best function in $\mathcal{H}$.

Previously we looked for a function consistent with our sample. This may not be possible. We can only try to minimize the empirical error L_S defined as before.

As before $\mathbb{E}_{S\sim D^m}[L_S(h)] = L(h)$ so the empirical error is an unbiased estimator of the true error.

We are now going to derive general learning bounds in this setting. Assume we have an algorithm that can produce $\hat{h}$ with minimum empirical error from S.

What we want: a lower bound on m such that we can <u>guarantee</u> for any $\delta, \epsilon > 0$

$$\mathbb{P}_{S \sim \mathcal{D}^m}\left(\left[L(h_S) - \min_{h \in \mathcal{H}} L(h)\right] > \epsilon\right) < \delta .$$

Strategy: use Chernoff-Hoeffding bounds to ensure

$$\forall h \in \mathcal{H} \quad |L(h) - L_S(h)| \leq \epsilon/2$$

Now we can calculate:

$$\begin{aligned} L(h_S) &\leq L_S(h_S) + \epsilon/2 \\ &\leq L_S(h) + \epsilon/2 \quad [h_S \text{ gives the minimal empirical error}] \\ &\leq L(h) + \epsilon \end{aligned}$$

Since this is true for any h it holds for the h that minimizes the error

$$L(h_S) \leq \min_{h \in \mathcal{H}} L(h) + \epsilon$$

<u>Lemma 1</u> Let $\{(x_i, y_i)\}_{i=1}^m$ be iid samples drawn from $\mathcal{D}$.

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

then with probability at least $1-\delta$, $\forall h \in \mathcal{H}\ |L(h) - L_S(h)| \leq \frac{\epsilon}{2}$

<u>Proof</u> Fix any $h \in \mathcal{H}$. Define $Z_i = \begin{cases} 1 \text{ if } h(x_i) \neq y_i \\ 0 \text{ otherwise} \end{cases}$

$L_S(h) = \frac{1}{m}\sum_{i=1}^m Z_i$ and $L(h) = \mathbb{E}[L_S(h)]$

Now we can use C-H bound $\forall i, (b_i - a_i) = 1$ & use $\epsilon/2$ for ϵ

$$\mathbb{P}\left(|L(h) - L_S(h)| \geq \epsilon/2\right) \leq 2e^{-\epsilon^2 m/2}$$

Applying the union bound we get

$$P(\exists h \in \mathcal{H} \text{ s.t. } |L(h) - L_S(h)| \geq \epsilon/2) \leq 2|\mathcal{H}| e^{-\epsilon^2 m/2}$$

Now we want this to be $\leq \delta$ so we have

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

Now we can combine this with the argument just before Lemma 1 to obtain

Thm For any finite concept class $\mathcal{H}$, if we have an algorithm A that takes iid samples $S = \{(x_1, y_1), \ldots (x_m, y_m)\}$ drawn from $\mathcal{D}$ over $X \times Y$ & outputs a function $h_S \in \mathcal{H}$ with minimal empirical error, then $\forall \mathcal{D}$ over $X \times Y$, $\forall \epsilon, \delta > 0$ if

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

we have with probability <u>at least</u> $1-\delta$

$$L(h_S) \leq \min_{h \in \mathcal{H}} L(h) + \epsilon$$

<u>Remarks</u> (1) We pay $1/\epsilon$ penalty in the sample size in the agnostic case.

(2) The smaller we have $\mathcal{H}$ the better for reducing the error. Small $\mathcal{H}$ reduces the danger of overfitting.