$A_1, \ldots A_k$ events

$$P(\cup_i A_i) \leq P(A_1) + \ldots + P(A_k)$$

Hoeffding inequality

$Z_1, \ldots Z_m$ independent, identically distrib^{d} (iid)

drawn from Bernoulli dist.

$$P(z_i = 1) = \phi \,, \quad P(z_i = 0) = 1 - \phi$$

Let $\hat{\phi} = \frac{1}{m} \sum_i z_i$

For $\gamma > 0$

$$\boxed{P(|\phi - \hat{\phi}| > \gamma) \leq 2e^{-2\gamma^2 m}}$$

Training set $S = \{(x^i, y^i), i = 1, \ldots, m\}$

Samples drawn iid from prob dist D

Training error for hypothesis h

$$\hat{\varepsilon}(h) = \frac{1}{m} \sum 1\{h(x^i) \neq y^i\}$$

Generalization error

$$\epsilon(h) = P_{(x,y) \sim D}(h(x) \neq y)$$

Ex linear classifier

$$h_\theta(x) = 1\{\theta^T x \geq 0\}$$

Pick $\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\ \hat{\varepsilon}(h_\theta)$

In general hypothesis class H

Finite $\mathcal{H}$

$$\mathcal{H} = \{h_1, \ldots, h_k\}$$

Draw a sample, and let Z denote whether $h_i(x)$ misclassifies it, i.e.

$$Z = 1\{h_i(x) \neq y\}$$

$$\hat{\epsilon}(h_i) = \frac{1}{m} \sum_j Z_j$$

But this is just the est. mean of a Bernoulli distribution!

So $$P(|\epsilon(h_i) - \hat{\epsilon}(h_i)| > \gamma) \leq 2e^{-\gamma^2 m}$$

Want this to be true for any $h_i \in H$

$$P(\exists h \in H \; |\epsilon(h) - \hat{\epsilon}(h_i)| > \gamma) = P(A_1 \cup \dots \cup A_k)$$

$$\leq \sum_{i=1}^{k} P(A_k)$$

$$\leq \sum_{i=1}^{k} 2e^{-2\gamma^2 m}$$

$$= 2ke^{-2\gamma^2 m}$$

Subtract 1 from both sides...

$$P(\neg \exists h \in H \; |\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma)$$

$$= P(\forall h \in H \; |\varepsilon(h_i) - \hat{\varepsilon}(h_i)| \leq \gamma) \geq 1 - 2ke^{-2\gamma^2 m}$$

Q: Given γ and some $\delta > 0$
how large must m be to guarantee that w. Prob $1-\delta$ the training error will be within γ of generalization error?

A: $m > \frac{1}{2\gamma^2} \log \frac{2k}{\delta}$ sample complexity

---

Q: $|\hat{\epsilon}(h) - \epsilon(h) \leq ???$

$$h^* = \arg\min_{h \in H} \epsilon(h)$$

$$\begin{aligned} \epsilon(\hat{h}) &\leq \hat{\epsilon}(\hat{h}) + \gamma \\ &\leq \hat{\epsilon}(h^*) + \gamma \\ &\leq \epsilon(h^*) + 2\gamma \end{aligned}$$

Thm

$$\epsilon(\hat{h}) \leq \min_{h \in H} \epsilon(h) + 2\sqrt{\frac{1}{2m} \log \frac{2k}{\delta}}$$

k hypotheses result

Case of ∞ h

Not the right argument but has the right intuitions

- H parameterized by d real numbers
- Assume 64 bits per real #.

Then 2^{64d} possibilities

So using sample complexity

$$m \geq O\left(\frac{1}{\gamma^2} \log \frac{2^{64d}}{\delta}\right) = O\left(\frac{d}{\gamma^2} \log \frac{1}{\delta}\right)$$

$$= O_{\gamma,\delta}(d)$$

These notes follow Andrew Ng's notes

VC Dimension

$$\left|\epsilon(h) - \hat{\epsilon}(h)\right| \leq O\left(\sqrt{\left(\frac{d}{m}\log\frac{m}{d} + \frac{1}{m}\log\frac{1}{\delta}\right)}\right)$$

$$() = \frac{1}{m}\log\left(\frac{m}{d}\right)^{d} + \frac{1}{m}\log\frac{1}{\delta}$$

$$= \frac{1}{m}\log\left(\frac{1}{\delta}\left(\frac{m}{d}\right)^{d}\right)$$