CS485/685
Lecture 19: March 13, 2012

Hypothesis Dependent Bounds

[BDSS] Chapters 6-7

Uniform Convergence Recap

Non-Uniform Convergence Idea

Weighted Hypotheses

- Idea: Assign weights to hypotheses such that the sum of all the weights is at most 1.
  - E.g. $w: H \rightarrow [0,1]$
    s.t. $\sum_{h \in H} w(h) \leq 1$
- For countable H, $\exists f$ such that $f: \mathbb{N} \rightarrow H$
- So we define $w(h) = \frac{1}{f^{-1}(h)^2}$

Non-uniform error bound

- **Theorem:** Let $w: H \to [0,1]$ such that $\sum_{h \in H} w(h) \leq 1$
  Then $\forall N, \delta > 0$ and D

$$\Pr_{S \sim D^N}[\exists h \in H\ s.t. |L_S(h) - L_D(h)| \geq \epsilon_h] \leq \delta$$

where $\epsilon_h = \sqrt{\frac{\ln\left(\frac{1}{w(h)}\right) + \ln\left(\frac{2}{\delta}\right)}{2N}}$

Proof

$P_{S \sim D^N}[\exists h \in H\ s.t. |L_S(h) - L_D(h)| \geq \epsilon_h]$

$\leq \sum_{h \in H} \Pr_{S \sim P^N}[|L_S(h) - L_D(h)| \geq \epsilon_h]$ by union bound

$\leq \sum_{h \in H} 2e^{-2N\epsilon_h^2}$ by Hoeffding's bound

$\forall D\ \Pr(|L_S(h) - L_D(h)| > \epsilon) \leq 2e^{-2N\epsilon^2}$

$= \sum_{h \in H} w(h)\delta$ since $\epsilon_h = \sqrt{\frac{\ln\left(\frac{1}{w(h)}\right) + \ln\left(\frac{2}{\delta}\right)}{2N}}$

$= \delta$

Hypothesis Dependent Bound

- Given a training set of size N, the loss $L_D(h)$ achieved by any algorithm that returns a hypothesis h is bounded by $L_S(h) + \epsilon_h$ with prob $\geq 1 - \delta$

  where $\epsilon_h = \sqrt{\frac{\ln\left(\frac{1}{w(h)}\right) + \ln\left(\frac{2}{\delta}\right)}{2N}}$

New Learning Paradigm

- Idea: instead of doing ML, MAP or Bayesian learning, choose the hypothesis with lowest error bound:

$$h^* = argmin_{h \in H} \underbrace{L_S(h)}_{\text{empirical error}} + \underbrace{\epsilon_h}_{\text{generalization error}}$$

- Hence $\epsilon_h = \sqrt{\frac{\ln\left(\frac{1}{w(h)}\right) + \ln\left(\frac{2}{\delta}\right)}{2N}}$ can be thought as a regularization term

Weighting Function

- Where does $w(h)$ come from?
- Idea: give more weight to hypotheses that are more likely.
- Minimize $L_S(h) + \epsilon_h$ where $\epsilon_h = \sqrt{\frac{\ln\left(\frac{1}{w(h)}\right)+\ln\left(\frac{2}{\delta}\right)}{2N}}$
  - As $L_D(h) \downarrow$ then $L_S(h) \downarrow$ and $\epsilon_h \downarrow$ since $w(h) \uparrow$
  - As $L_D(h) \uparrow$ then $L_S(h) \uparrow$ and $\epsilon_h \uparrow$ since $w(h) \downarrow$

Weighting Function

- How can we give more weight to more likely hypotheses?
- Two common ideas:
  - Description length
  - Prior probability

Description Length

- **Occam's razor**: simpler hypotheses are generally better
- Use description length as a proxy for the complexity of a hypothesis
- Set weight inversely proportional to description length

Description Length

- Let Σ be a set of symbols (alphabet)
  - E.g., $\Sigma = \{0,1\}$ or $\{a, b, c, d\}$
- Let $d: H \rightarrow \Sigma^*$ be a description language
- Let $|h|$ be the length of $d(h)$

Description Dependent Bound

- **Definition**: d is a prefix-free language if $\forall h\ d(h)$ is not the prefix of any $d(h')$

- **Theorem**: Let $d: H \rightarrow \{0,1\}^*$ be a prefix-free description language. Then $\forall N, \delta > 0, D$ with prob $\geq 1 - \delta$

$$\forall h \in H\ \ L_D(h) \leq L_S(h) + \sqrt{\left(|h| + \ln\left(\frac{2}{\delta}\right)\right)/2N}$$

Proof

- Let $w(h) = 1/2^{|h|}$
- **Kraft's inequality:** if $S \subseteq \{0,1\}^*$ is a prefix-free set of strings then

$$\sum_{\sigma \in S} \frac{1}{2^{|\sigma|}} \leq 1$$

- This can be verified by defining $w(h)$ to be the probability of generating $d(h)$ by repeatedly tossing a coin that outputs 0 for head and 1 for tail.
- Since probabilities sum up to 1, then Kraft's inequality holds.

Prior distribution

- Instead of using description length, we can use domain knowledge to set $w(h)$ based on some prior distribution $\Pr(h)$.

- **Theorem:** Let $\Pr(h)$ be some prior distribution over H. Then $\forall N, \delta > 0, D$ with prob $\geq 1 - \delta$

$$\forall h \in H \quad L_D(h) \leq L_S(h) + \sqrt{\left(\ln\left(\frac{1}{\Pr(h)}\right) + \ln\left(\frac{2}{\delta}\right)\right)/2N}$$

Example: Decision Trees

- Description Language: