Statistical Learning Theory: Error Bounds and VC-Dimension

CS4780/5780 – Machine Learning
Fall 2014
Thorsten Joachims
Cornell University
Reading: Mitchell Chapter 7 (not 7.4.4 and 7.5)

Example: Smart Investing

• Task: Pick stock analyst based on past performance.
• Experiment:
  – Review analyst prediction “next day up/down” for past 10 days. Pick analyst that makes the fewest errors.
  – Situation 1:
    • 2 stock analyst (A1, A2), A1 makes 5 errors
  – Situation 2:
    • 5 stock analysts (A1, A2, B1, B2, B3), B2 best with 1 error
  – Situation 3:
    • 1005 stock analysts (A1, A2, B1, B2, B3, C1, ..., C1000), C543 best with 0 errors
• Question: Which analysts are you most confident in, A1, B2, or C543?

Useful Formula

Hoeffding/Chernoff Bound:
For any distribution P(X) where X can take the values 0 and 1, the probability that an average of an i.i.d. sample deviates from its mean p by more than ε is bounded as

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} x_i - p\right| > \varepsilon\right) < 2e^{-2\varepsilon^2 n}$$

Overfitting vs. Underfitting

With probability at least (1-δ):

$$\text{error}\left(h_{\text{train}}\right) - \text{error}\left(h_{\text{test}}\right) \leq \sqrt{\frac{\ln(2IH/\delta)}{2n}}$$

[Mitchell, 1997]
Generalization Error Bound: Infinite H, Non-Zero Error

- Setting
  - Sample of $n$ labeled instances $S$
  - Learning Algorithm $L$ using a hypothesis space $H$ with $\text{VCDim}(H) = d$
  - $L$ returns hypothesis $\hat{h} = L(S)$ with lowest training error

- Definition: The VC-Dimension of $H$ is equal to the maximum number $d$ of examples that can be split into two sets in all $2^d$ ways using functions from $H$ (shattering).

- Given hypothesis space $H$ with $\text{VCDim}(H) = d$ and an i.i.d. sample $S$ of size $n$, with probability $(1 - \delta)$ it holds that

$$\forall \epsilon \in (0, 1): \text{Err}_S(\hat{h}) \leq \text{Err}_S(h^*) + \frac{d \cdot \ln(\frac{2n}{d}) + 1}{n} - \ln(\frac{1}{\delta})$$