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)

Probably Approximately Correct Learning

Definition: C is **PAC-learnable** *by learning algorithm* $\mathcal{L}$ *using* H *and a sample* S *of* n *examples drawn i.i.d. from some fixed distribution* $P(X)$ *and labeled by a concept* $c \in C$, *if for sufficiently large* n

$$P(Err_P(h_{\mathcal{L}(S)}) \leq \epsilon) \geq (1-\delta)$$

for all $c \in C, \epsilon > 0, \delta > 0$, *and* $P(X)$. $\mathcal{L}$ *is required to run in polynomial time dependent on* $1/\epsilon, 1/\delta, n$, *the size of the training examples, and the size of* c.

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

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

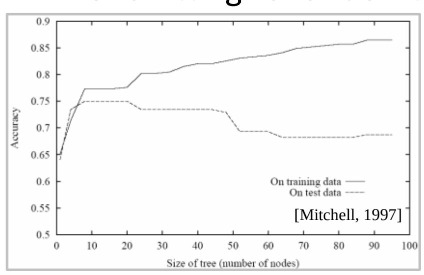
Generalization Error Bound: Finite H, Non-Zero Error

- Setting
  - Sample of n labeled instances *S*
  - Learning Algorithm *L* with a finite hypothesis space *H*
  - *L* returns hypothesis $\hat{h}=L(S)$ with lowest training error
- What is the probability that the prediction error of $\hat{h}$ exceeds the fraction of training errors by more than ε?

$$P\left(\left|Err_S(h_{\mathcal{L}(S)}) - Err_P(h_{\mathcal{L}(S)})\right| \geq \epsilon\right) \leq 2|H|e^{-2\epsilon^2 n}$$

Training Sample S_{train} $(x_1,y_1), \ldots, (x_n,y_n)$ $\xrightarrow{S_{train}}$ Learner $\xrightarrow{\hat{h}}$ Test Sample S_{test} $(x_{n+1},y_{n+1}), \ldots$

Overfitting vs. Underfitting



With probability at least *(1-δ)*:

$$Err_P(h_{\mathcal{L}(S_{train})}) \leq Err_{S_{train}}(h_{\mathcal{L}(S_{train})}) + \sqrt{\frac{(\ln(2|H|) - \ln(\delta))}{2n}}$$

Generalization Error Bound: Infinite H, Non-Zero Error

- Setting
  - Sample of n labeled instances *S*
  - Learning Algorithm *L* using a hypothesis space *H* with *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 *VCDim(H)* equal to *d* and an i.i.d. sample *S* of size *n*, with probability ($1-\delta$) it holds that

$$Err_P(h_{\mathcal{L}(S)}) \leq Err_S(h_{\mathcal{L}(S)}) + \sqrt{\frac{d\left(\ln\left(\frac{2n}{d}\right)+1\right)-\ln\left(\frac{\delta}{4}\right)}{n}}$$