Statistical Learning Theory: Generalization Error Bounds

CS6780 – Advanced Machine Learning
Spring 2015

Thorsten Joachims
Cornell University

Reading: Murphy 6.5.4
Schoelkopf/Smola Chapter 5 (beginning, rest later)

Outline

Questions in Statistical Learning Theory:
- How good is the learned rule after n examples?
- How many examples do I need before the learned rule is accurate?
- What can be learned and what cannot?
- Is there a universally best learning algorithm?

In particular, we will address:
What is the true error of h if we only know the training error of h?
- Finite hypothesis spaces and zero training error
- Finite hypothesis spaces and non-zero training error
- Infinite hypothesis spaces and VC dimension (later)

Can you Convince me of your Psychic Abilities?

- Game
  - I think of n bits
  - If somebody in the class guesses my bit sequence, that person clearly has telepathic abilities – right?
- Question:
  - If at least one of |H| players guesses the bit sequence correctly, is there any significant evidence that he/she has telepathic abilities?
  - How large would n and |H| have to be?

Discriminative Learning and Prediction Reminder

Real-world Process $P(X,Y)$

drawn i.i.d. drawn i.i.d.

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

- Goal: Find h with small prediction error $Err_P(h)$ over $P(X,Y)$.
- Discriminative Learning: Given H, find h with small error $Err_{S_{train}}(h)$ on training sample S_{train}.

- Training Error: Error $Err_{S_{train}}(h)$ on training sample.
- Test Error: Error $Err_{S_{test}}(h)$ on test sample is an estimate of $Err_P(h)$

Useful Formulas

- Binomial Distribution: The probability of observing *x* heads in a sample of *n* independent coin tosses, where in each toss the probability of heads is *p*, is

$$P(X = x|p,n) = \frac{n!}{r!\,(n-r)!} p^x (1-p)^{n-x}$$

- Union Bound:

$$P(X_1 = x_1 \vee X_2 = x_2 \vee \cdots \vee X_n = x_n) \leq \sum_{i=1}^{n} P(X_i = x_i)$$

- Unnamed:

$$(1-\epsilon) \leq e^{-\epsilon}$$

Generalization Error Bound: Finite H, Zero Error

- Setting
  - Sample of n labeled instances S_{train}
  - Learning Algorithm L with a finite hypothesis space H
  - At least one $h \in H$ has zero prediction error $Err_P(h)=0$ (→ $Err_{S_{train}}(h)=0$)
  - Learning Algorithm L returns zero training error hypothesis $\hat{h}$
- What is the probability that the prediction error of $\hat{h}$ is larger than ε?

$$P(Err_P(\hat{h}) \geq \epsilon) \leq |H| e^{-\epsilon n}$$

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

Sample Complexity: Finite H, Zero Error

- Setting
  - Sample of n labeled instances S_{train}
  - Learning Algorithm L with a finite hypothesis space H
  - At least one $h \in H$ has zero prediction error (→ $Err_{S_{train}}(h)=0$)
  - Learning Algorithm L returns zero training error hypothesis $\hat{h}$
- How many training examples does L need so that with probability at least (1-δ) it learns an $\hat{h}$ with prediction error less than ε?

$$n \geq \frac{1}{\epsilon}\left(\log(|H|) - \log(\delta)\right)$$

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

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

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

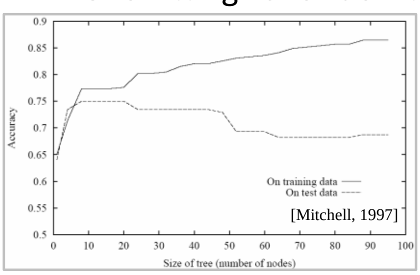
Generalization Error Bound: Finite H, Non-Zero Error

- 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 $(1-\delta)$:

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