Statistical Learning Theory: PAC Learning

CS4780/5780 – Machine Learning
Fall 2014

Thorsten Joachims
Cornell University

Reading: Mitchell Chapter 7 (not 7.4.4 and 7.5)
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
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

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

- Training Error: Error $\text{Err}_{S_{\text{train}}}(h)$ on training sample.
- Test Error: Error $\text{Err}_{S_{\text{test}}}(h)$ on test sample is an estimate of $\text{Err}_P(h)$.
Definition: A particular instance of a learning problem is described by a probability distribution \( P(X, Y) \).

Definition: A sample \( S = ((x_1, y_1), ..., (x_n, y_n)) \) is independently identically distributed (i.i.d.) according to \( P(X, Y) \).

Definition: The error on sample \( S \) \( Err_S(h) \) of a hypothesis \( h \) is \( Err_S(h) = \frac{1}{n} \sum_{i=1}^{n} \Delta(h(x_i), y_i) \).

Definition: The prediction/generalization/true error \( Err_P(h) \) of a hypothesis \( h \) for a learning task \( P(X, Y) \) is

\[
Err_P(h) = \sum_{\bar{x} \in X, y \in Y} \Delta(h(\bar{x}), y) P(X = \bar{x}, Y = y).
\]

Definition: The hypothesis space \( H \) is the set of all possible classification rules available to the learner.
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:

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

• Unnamed:

$$(1 - \epsilon) \leq e^{-\epsilon}$$
Generalization Error Bound: Finite H, Zero Error Error

• Setting
  – Sample of n labeled instances $S_{\text{train}}$
  – Learning Algorithm $L$ with a finite hypothesis space $H$
  – At least one $h \in H$ has zero prediction error $\text{Err}_P(h) = 0$ ($\Rightarrow \text{Err}_{S_{\text{train}}}(h) = 0$)
  – Learning Algorithm $L$ returns zero training error hypothesis $\hat{h}$

• What is the probability that the prediction error of $\hat{h}$ is larger than $\varepsilon$?

\[ P(\text{Err}_P(\hat{h}) \geq \varepsilon) \leq \frac{H}{n} e^{-\varepsilon n} \]

Training Sample $S_{\text{train}}$

$S_{\text{train}} (x_1, y_1), \ldots, (x_n, y_n)$

Learner

\[ \hat{h} \]

Test Sample $S_{\text{test}}$

$(x_{n+1}, y_{n+1}), \ldots$
Sample Complexity: Finite H, Zero Error

- **Setting**
  - Sample of $n$ labeled instances $S_{\text{train}}$
  - Learning Algorithm $L$ with a finite hypothesis space $H$
  - At least one $h \in H$ has zero prediction error ($\Rightarrow \text{Err}_{S_{\text{train}}}(h) = 0$)
  - Learning Algorithm $L$ returns zero training error hypothesis $\hat{h}$

- **How many training examples does $L$ need so that with probability at least (1-$\delta$) it learns an $\hat{h}$ with prediction error less than $\epsilon$?**

$$n > \frac{1}{\epsilon}(\log(|H|) - \log(\delta))$$

---

**Training Sample** $S_{\text{train}}$

$(x_1,y_1), \ldots, (x_n,y_n)$

**Test Sample** $S_{\text{test}}$

$(x_{n+1},y_{n+1}), \ldots$