Machine Learning for Intelligent Systems

Lecture 18: Statistical Learning Theory 2

Reading: UML 6

Instructors: Nika Haghtalab (this time) and Thorsten Joachims

Sample Complexity – 0 Empirical Error

**Theorem:** Sample Complexity (0 empirical error)

Let \( m \geq \frac{1}{\epsilon^2} \left( \ln(|H|) + \ln \left( \frac{2}{\delta} \right) \right) \). For any instance space \( X \), labels \( Y = \{-1, 1\} \), distribution \( P \) on \( X \times Y \), with probability \( 1 - \delta \) over i.i.d draws of set \( S \) of \( m \) samples, we have Any \( h \in H \) that has 0 empirical error, has true error of \( \epsilon \).

Learning Algorithm: Given a sample set \( S \) and hypothesis class \( h \in H \), if there is a \( h_\delta \in H \) that is consistent with \( S \), return \( h_\delta \). (Eq: Return \( h_\delta \) in version space \( V_S(H, S) \))

Sample Complexity – General

**Theorem:** Sample Complexity (non-zero empirical error)

Let \( m \geq \frac{1}{\epsilon^2} \left( \ln(|H|) + \ln \left( \frac{2}{\delta} \right) \right) \). For any instance space \( X \), labels \( Y = \{-1, 1\} \), distribution \( P \) on \( X \times Y \), with probability \( 1 - \delta \) over i.i.d draws of set \( S \) of \( m \) samples, \( h_\delta \in H \), with least empirical error, has true error \( \epsilon \).

Fundamental Questions

Questions in Statistical Learning Theory:

- Trying to learn a classifier from \( H \)?
- How good is the learned rule after \( m \) examples?
- How many examples is needed for the learned rule to be accurate?
- What can be learned and what cannot?
- Is there a universally best learning algorithm?

In particular, we will address:

- What kind of a guarantee on the true error of a classifier can I get if I know its training error?
- Is there a universally best learning algorithm?
- What can be learned and what cannot?
- How many examples is needed for the learned rule to be accurate?
- Trying to learn a classifier from \( H \)?

No Consistent Hypothesis

**A reasonable learning Algorithm:** Given a sample set \( S \) and hypothesis class \( h \in H \), return \( h_\delta = \arg\min_{h \in H} \text{err}_S(h) \).

What can go wrong?

Best hypothesis on distribution \( h^* = \arg\min_{h \in H} \text{err}_S(h) \).

The true error of \( h_\delta \) is within \( \epsilon \) of the optimal true error, \( \text{err}_S(h^*) \), if For all \( h \in H \), we have \( |\text{err}_S(h) - \text{err}_S(h^*)| \leq \epsilon \).

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, C543?
Infinite Hypothesis Classes

Linear thresholds in

\[ w \]

Thresholds on the line

Neural Networks

\[ W_1, W_2, \ldots \]

Intervals on the real line

Sample Complexity bounds for finite hypothesis spaces become meaningless:

\[ \frac{1}{\varepsilon} \left( \ln(|\mathcal{H}|) + \ln \left( \frac{1}{\delta} \right) \right) \]

\[ \frac{2}{\varepsilon^2} \left( \ln(|\mathcal{H}|) + \ln \left( \frac{1}{\delta} \right) \right) \]

Example 1: Growth Function

What is \( H[m] \) for thresholds on a line:

- \( h\_x(x) = 1 \) if \( x \geq w \) and \(-1\) otherwise.
- \( H \) is infinitely large
- \( H[m] \)

\[ w \in \mathbb{R} \]

\[ \_+ \_+ \_+ \_+ \_+ \_+ \_+ \_+ \_+ \_+ \}

- For any \( m \) points, \( H[m] \) is the number of intervals they divide the line to, which is at most \( m + 1 \ll 2^m \).

Example 2: Growth Functions

What is \( H[m] \) for intervals on the line:

- \( h\_x(x) = 1 \) if \( w' \geq x \geq w \) and \(-1\) otherwise
- \( H \) is infinitely large

\[ H[m] = m \choose 0 + m \choose 1 + m \choose 2 + \cdots + m \choose k = 1 + m + \frac{m(m-1)}{2} = O(m^2) \ll 2^m \]

- Where \( m \choose k \) is the number of ways we can choose a subset of size \( k \) from a set of \( m \) items.

Sample Complexity – growth Function

Let \( m \geq \frac{2}{\varepsilon} \left( \ln(|\mathcal{H}|) + \ln \left( \frac{1}{\delta} \right) \right) \) for some constant \( c_0 \).

For any instance space \( X \), labels \( Y = \{-1, 1\} \), distribution \( P \) on \( X \times Y \), with probability \( 1 - \delta \) over i.i.d draws of set \( S \) of \( m \) samples, we have any \( h \in H \) that has \( 0 \) empirical error, has true error of \( \text{err}_P(h) \leq \varepsilon \).

- Difficult to interpret:
  \[ m \geq \Omega \left( \frac{\ln(|\mathcal{H}|) + \ln \left( \frac{1}{\delta} \right)}{\varepsilon} \right) \]

- If \( H[m] = 2^m \), the sample complexity is Impossible to learn from samples.

VC Dimension

Shattering and VC Dimension

\( H \) shatters a sample set \( S \) if \( \|H[S]\| = 2^{|S|} \).

VC Dimension of \( H \) is the size of the largest set \( S \) that can be shattered by \( H \).

- \( \text{VCDim}(H) \): Largest \( m \) for which \( H[m] = 2^m \).

VC Dimension is roughly the point where the growth function stops being exponential and becomes polynomial.

- When is learning from samples possible?
  - \( \text{VCDim}(H) = \infty \) then \( H[m] = 2^m \) for all \( m \)
  - \( \text{VCDim}(H) = \delta \) then \( H[m] < O(m^\delta) \) for all \( m \)
  - \( \text{VCDim}(H) = \delta \) then \( H[m] < O(m^\delta) \) for all \( m \)

  We can learn!