Machine Learning for Intelligent Systems

Lecture 18: Statistical Learning Theory 2

Reading: UML 6

Instructors: Nika Haghtalab (this time) and Thorsten Joachims

1

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?

2

Sample Complexity – 0 Empirical Error

Theorem: Sample Complexity (zero empirical error)

Let $m \geq \frac{1}{\epsilon}\left(\ln(|H|) + \ln\left(\frac{1}{\delta}\right)\right)$. 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 $err_P(h) \leq \epsilon$.

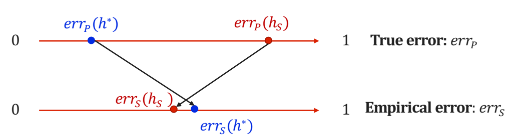
Learning Algorithm: Given a sample set S and hypothesis class $h \in H$, if there is a $h_S \in H$ that is *consistent* with S, return h_S. (Eqv. Return h_S in version space $VS(H, S)$)

4

No Consistent Hypothesis

A reasonable learning Algorithm: Given a sample set S and hypothesis class $h \in H$, return $h_S = argmin_{h \in H}\, err_S(h)$.

What can go wrong?
Best hypothesis on distribution $h^* = argmin_{h \in H}\, err_P(h)$.



The **true error** of h_S is within ϵ of the **optimal true error,** $err_P(h^*)$, if

For all $h \in H$, we have $|err_S(h) - err_P(h)| \leq \frac{\epsilon}{2}$.

5

Sample Complexity – General

Theorem

For any instance space X, labels $Y = \{-1, 1\}$, and distribution P on $X \times Y$, consider a set S of m i.i.d. samples from P. We have

$$\Pr_{S \sim P^m}\left[\exists\, h \in H, \quad |err_S(h) - err_P(h)| > \frac{\epsilon}{2}\right] \leq 2|H|e^{-\epsilon^2 m/2}.$$

Theorem: Sample Complexity (non-zero empirical error)

Let $m \geq \frac{2}{\epsilon^2}\left(\ln(|H|) + \ln\left(\frac{2}{\delta}\right)\right)$. 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, $h_S \in H$, with **least empirical error,** has **true error**

$$err_P(h_S) \leq err_P(h^*) + \epsilon.$$

6

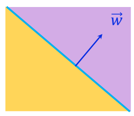
Example: Smart Investing

- **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?

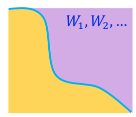
7

Infinite Hypothesis Classes

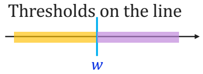
Linear thresholds in



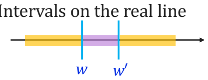
Neural Networks



Thresholds on the line



Intervals on the real line



Sample Complexity bounds for finite hypothesis spaces become meaningless:

$$\frac{1}{\epsilon}\left(\ln(|H|) + \ln\left(\frac{1}{\delta}\right)\right) \qquad \frac{2}{\epsilon^2}\left(\ln(|H|) + \ln\left(\frac{2}{\delta}\right)\right)$$

8

Effective Number of Hypotheses

How many different ways hypotheses in H label the sample set S?

Most complex: Many unique rows
2^m unique rows

$H \backslash S$	x_1	x_2	x_3	…	x_m
h_1	-1	-1	1		-1
h_2	1	-1	-1		1
h_3	-1	1	-1		1
h_4	1	1	1		-1
⋮					

$h_i(x_j)$

Least complex: Just one unique row
1 unique row

$H \backslash S$	x_1	x_2	x_3	…	x_m
h_1	1	1	-1		1
h_2	1	1	-1		1
h_3	1	1	-1		1
h_4	1	1	-1		1
⋮	1	1	-1		1

Growth function

The set all m-tuples produced by hypotheses in H on the sample set S

$$H[S] = \{\, (h(x_1), h(x_2), h(x_3), \ldots, h(x_m)) \,\}_{h \in H}$$

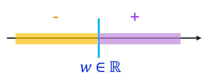
Growth function: $H[m] = \max_{|S|=m} |H[S]|$ is the largest number of unique rows that H can produce on any set of m elements.

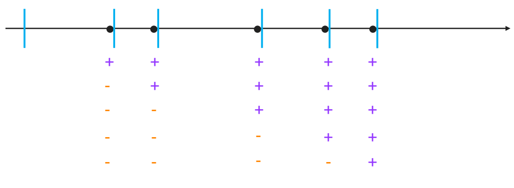
9

Example 1: Growth Function

What is $H[m]$ for *thresholds on a line*:

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





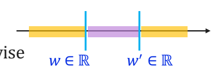
- For any m points, $H[m]$ is the number of intervals they divide the line to, which is at most $m + 1 \ll 2^m$.

10

Example 2: Growth Functions

What is $H[m]$ for *intervals on the line*:

- $h_{w,w'}(x) = 1$ if $w' \geq x \geq w$ and -1 otherwise
- H is infinitely large



$$H[m] = \binom{m}{0} + \binom{m}{1} + \binom{m}{2} = 1 + m + \frac{m(m-1)}{2} = O(m^2) \ll 2^m$$

- Where $\binom{m}{k}$ is the number of ways we can choose a subset of size k from a set of m items.

$$\binom{m}{k} = \frac{m!}{(m-k)!\,k!}$$

11

Sample Complexity – growth Function

Let $m \geq \frac{c_0}{\epsilon}\left(\ln(H[2m]) + \ln\left(\frac{1}{\delta}\right)\right)$ 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 $err_P(h) \leq \epsilon$.

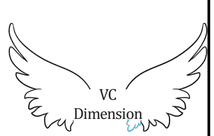
- Difficult to interpret:

$$m \geq \Omega\left(\frac{\ln(H[2m])}{\epsilon}\right)$$

- If, $H[m] = 2^m$, the sample complexity is

Impossible to learn from samples.

$$m \geq \Omega\left(\frac{m}{\epsilon}\right)$$



12

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. ← $\mathrm{VCDim}(H)$: 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?

- If $\mathrm{VCDim(H)} = \infty$ then $H[m] = 2^m$ for all m
→ It would be impossible to learn!
- If $\mathrm{VCDim(H)} = d$ then $H[m] < O(m^d)$ for all m
→ We can learn!

13