Learning Theory

CISC 5800
Professor Daniel Leeds

The classifier

Function C that provides correct label (Y) based on features (X)

\[ C(x) = y \]

Goal: identifier classifier that maximizes correct labels for most inputs

Sample complexity

How many training examples needed to learn concept?

• \( X \) – set of data points
• \( P(X) \) – Probability of drawing data point \( x \)
• \( H \) – space of hypotheses \( H = \{ h : X \rightarrow \text{classes} \} \)
• \( C \) – correct assignment \( C = \{ c : c(x) = y \ \forall x \in X \} \)

Probability of error

\( H = \{ h : X \rightarrow \{0,1\} \} \)

True error of \( h \): probability randomly selected data point from \( P(X) \) misclassified

\[ \text{error}_{\text{true}}(h) = \Pr_{x \sim P(X)}[h(x) \neq c(x)] \]

• Hard to compute, but can prove properties of \( \text{error}_{\text{true}} \)
Example: Learner picks one of fixed number of classifiers $h \in H$
Correct classifier $c$ is some assignment of each $x$ to a label

How many training points $m$ needed for $\text{error}_{\text{true}}(h) < \varepsilon$?

$\text{Prob}[\text{error}_{\text{true}}(h) \leq \varepsilon] > 1 - \delta$

"Probability learned classifier $h$ has worse than $\varepsilon$ error is $< \delta"$

"Probably Approximately Correct Learning" – PAC Learning

Binary example: sample complexity
Note for $\varepsilon = [0,1], (1 - \varepsilon) \leq e^{-\varepsilon}$

What is the chance learned $h$ is bad but classifies training data correctly?

If $\text{error}_{\text{true}}(h) > \varepsilon$:
- $\text{Prob}[h \text{ correctly labels } x^1] < (1 - \varepsilon) \leq e^{-\varepsilon}$
- $\text{Prob}[h \text{ correctly labels } x^1 \text{ and } x^2 \ldots \text{ and } x^m] < (1 - \varepsilon)^m \leq e^{-m\varepsilon}$

If classifier picks one $h^*$ randomly from $H$

$\text{Prob}[h^* \text{ is bad}] = \text{Prob}[h_1 \text{ bad}] + \ldots + \text{Prob}[h_n \text{ bad}]$

$= \text{Prob}[\text{error}_{\text{true}}(h^*) > \varepsilon] < |H| e^{-m\varepsilon}$

Valiant, 1984

Binary example: sample complexity

Number of data points to reduce chance of false classification, enforce

$\text{Prob}[\text{error}_{\text{true}}(h) \leq \varepsilon] > 1 - \delta$

$1 - \text{Prob}[\text{error}_{\text{true}}(h) \leq \varepsilon] = \text{Prob}[\text{error}_{\text{true}}(h) > \varepsilon] < \delta$

$\text{Prob}[\text{error}_{\text{true}}(h^*) > \varepsilon] < |H| e^{-m\varepsilon} < \delta$

Valiant, 1984

Number of data points to reduce chance of false classification, enforce

$\text{Prob}[\text{error}_{\text{true}}(h) \leq \varepsilon] > 1 - \delta$

$\text{Prob}[\text{error}_{\text{true}}(h^*) > \varepsilon] < |H| e^{-m\varepsilon} < \delta$

$m > \frac{1}{\varepsilon} \ln \frac{|H|}{\delta}$

Valiant, 1984
VC Dimensions

If H not finite, PAC result seems to require $\infty$ data points

• Overly conservative

“Dichotomy” – division of set of points $S$ into two subsets

• “Shattering” – set of points is shattered by $H$ iff there exists $h \in H$ associated with every possible dichotomy

Vapnik-Chervonenkis dimension $VC(H)$ is size of largest finite subset of $S$ that can be shattered by $H$

Shattering example

• $H=$ {rectangles: inside is 1, outside is 0} \hspace{1cm} VC(3)
• $S=$ {3 specified dots}

Shattering example

• $H=$ {rectangles, inside is 1 outside is 0, inside is 0 outside is 1} \hspace{1cm} VC(4)
• $S=$ {4 specified dots}
Shattering example
• H={rectangles, inside is 1 outside is 0} \( VC(4) \)
• S={4 specified dots}

Shattering example 4
• H={rectangle, inside is 1 outside is 0}
• S={8 specified dots}

Shattering example 4
• H={rectangle, inside is 1 outside is 0} \( H(4) \)
• S={8 specified dots}

Shattering infinite points
• H={Linear separators}
• S={Any point in 2D feature space}
• S={Any point in nD space}
PAC result with infinite H

\[ \text{VC}(H) \text{ is size of largest finite subset of } X \text{ that can be shattered by } H \]

- \( d = \text{VC}(H) \)
- \( m \geq O\left( \frac{1}{\varepsilon} \left[ d \log \frac{1}{\varepsilon} + \log \frac{1}{\delta} \right] \right) \sim \frac{1}{\varepsilon} \left[ d \log \frac{1}{\varepsilon} + \log \frac{1}{\delta} \right] \)

Recall: \( m > \frac{1}{\varepsilon} \ln \frac{|H|}{\delta} \) for finite size \( H \)

Intuition for PAC result with infinite H

- \( d = \text{VC}(H) \)
- \( m \geq O\left( \frac{1}{\varepsilon} \left[ d \log \frac{1}{\varepsilon} + \log \frac{1}{\delta} \right] \right) \sim \frac{1}{\varepsilon} \left[ d \log \frac{1}{\varepsilon} + \log \frac{1}{\delta} \right] \)
- Finite \( H \): \( m > \frac{1}{\varepsilon} \ln \frac{|H|}{\delta} \)

\[ d \log \frac{k}{\varepsilon} \rightarrow \log \frac{k^d}{\varepsilon} \]

Can pick \( h \) to shatter at most \( d \) points in one of two classes

\( 2^d \) meaningfully different classifiers \( h \): \( |H| \sim 2^d \)