What now…

- We have explored many ways of learning from data
- But…
  - How good is our classifier, really?
  - How much data do I need to make it “good enough”?
A simple setting…

- Classification
  - N data points
  - Finite number of possible hypothesis (e.g., dec. trees of depth d)
- A learner finds a hypothesis \( h \) that is consistent with training data
  - Gets zero error in training – \( \text{error}_{\text{train}}(h) = 0 \)
- What is the probability that \( h \) has more than \( \varepsilon \) true error?
  - \( \text{error}_{\text{true}}(h) \geq \varepsilon \)

How likely is a bad hypothesis to get \( N \) data points right?

- Hypothesis \( h \) that is consistent with training data \( \rightarrow \)
  - got \( N \) i.i.d. points right
  - \( h \) “bad” if it gets all this data right, but has high true error
- Prob. \( h \) with \( \text{error}_{\text{true}}(h) \geq \varepsilon \) gets one data point right
- Prob. \( h \) with \( \text{error}_{\text{true}}(h) \geq \varepsilon \) gets \( N \) data points right
But there are many possible hypothesis that are consistent with training data.

How likely is learner to pick a bad hypothesis?

- Prob. $h$ with $\text{error}_{\text{true}}(h) \geq \epsilon$ gets $N$ data points right

- There are $k$ hypothesis consistent with data
  - How likely is learner to pick a bad one?
Union bound

- $P(A \lor B \lor C \lor D \lor \ldots)$

How likely is learner to pick a bad hypothesis

- Prob. a particular $h$ with $\text{error}_{\text{true}}(h) \geq \varepsilon$ gets $N$ data points right

- There are $k$ hypothesis consistent with data
  - How likely is it that learner will pick a bad one out of these $k$ choices?
Generalization error in finite hypothesis spaces [Haussler ’88]

**Theorem:** Hypothesis space $H$ finite, dataset $D$ with $N$ i.i.d. samples, $0 < \varepsilon < 1$: for any learned hypothesis $h$ that is consistent on the training data:

$$P(error_{true}(h) > \varepsilon) \leq |H| e^{-N\varepsilon}$$

Using a PAC bound

Typically, 2 use cases:

- 1: Pick $\varepsilon$ and $\delta$, give you $N$
- 2: Pick $N$ and $\delta$, give you $\varepsilon$
Summary: Generalization error in finite hypothesis spaces [Haussler ’88]

**Theorem:** Hypothesis space $H$ finite, dataset $D$ with $N$ i.i.d. samples, $0 < \varepsilon < 1$: for any learned hypothesis $h$ that is consistent on the training data:

$$P(\text{error}_\text{true}(h) > \varepsilon) \leq |H|e^{-N\varepsilon}$$

Even if $h$ makes zero errors in training data, may make errors in test

Limitations of Haussler ‘88 bound

- Consistent classifier
- Size of hypothesis space
What if our classifier does not have zero error on the training data?

- A learner with zero training errors may make mistakes in test set
- What about a learner with $\text{error}_{\text{train}}(h)$ in training set?

Simpler question: What’s the expected error of a hypothesis?

- The error of a hypothesis is like estimating the parameter of a coin!

- Chernoff bound: for $N$ i.i.d. coin flips, $x^1, \ldots, x^N$, where $x^j \in \{0,1\}$. For $0 < \epsilon < 1$:

$$P \left( \theta - \frac{1}{N} \sum_{j=1}^{N} x^j > \epsilon \right) \leq e^{-2N\epsilon^2}$$
Using Chernoff bound to estimate error of a single hypothesis

\[
P\left( \theta - \frac{1}{N} \sum_{j=1}^{N} x^j > \epsilon \right) \leq e^{-2N\epsilon^2}
\]

But we are comparing many hypothesis: **Union bound**

For each hypothesis \( h_i \):

\[
P(\text{error}_{true}(h_i) - \text{error}_{train}(h_i) > \epsilon) \leq e^{-2N\epsilon^2}
\]

What if I am comparing two hypothesis, \( h_1 \) and \( h_2 \)?
Generalization bound for $|H|$ hypothesis

**Theorem**: Hypothesis space $H$ finite, dataset $D$ with $N$ i.i.d. samples, $0 < \varepsilon < 1$ : for any learned hypothesis $h$:

$$P(error_{true}(h_i) - error_{train}(h_i) > \varepsilon) \leq e^{-2N\varepsilon^2}$$

PAC bound and Bias-Variance tradeoff

$$P(error_{true}(h) - error_{train}(h) > \varepsilon) \leq e^{-2N\varepsilon^2}$$

or, after moving some terms around, with probability at least $1-\delta$:

$$error_{true}(h) \leq error_{train}(h) + \sqrt{\frac{\ln |H| + \ln \frac{1}{\delta}}{2N}}$$

- Important: PAC bound holds for all $h$, but doesn’t guarantee that algorithm finds best $h$!!!