Learning Theory

Cornell CS 3/5780 · Spring 2026

0. Can You Convince Me of Your Psychic Abilities?

- I think of n bits. If somebody in the class guesses my bit sequence, that person clearly has telepathic abilities – right?
- Students $H = \{h_1, \ldots, h_{|H|}\}$ and any non-psychic student has $1 - p$ probability guessing any single bit correctly.

$$P(h_i \text{ correct} \mid h_i \text{ nonpsychic}) =$$

- How likely is it that at least one student is correct?

$$P(h_1 \text{ correct} \vee \ldots \vee h_{|H|} \text{ correct} \mid \text{all nonpsychic}) =$$

- How large would n need to be? Given some small δ, find n such that the probability above is less than δ:

$$n >$$

1. Setting

- Training data $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ drawn i.i.d. from $P(X, Y)$, binary classification, the 0-1 loss function $\ell(\hat{y}, y) = \mathbf{1}\{\hat{y} = y\}$, and a hypothesis class H
- **Notation**: expected test error of hypothesis h on distribution P:

$$\text{err}_P(h) = E_{(\mathbf{x},y)\sim P}\left[\ell(h(\mathbf{x}), y)\right]$$

- Learning Goal: find h with small expected error $\text{err}_P(h)$
- **Define**: sample error of hypothesis h on sample D:

$$\text{err}_D(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(\mathbf{x}_i), y_i)$$

- Learning Algorithm: empirical risk minimization

$$h_D = \arg\min_{h \in H} \text{err}_D(h)$$

2. Generalization Error of fixed h

- Define the generalization error as $|\text{err}_P(h) - \text{err}_D(h)|$
- **Hoeffding/Chernoff Bound:** For any distribution $P(U)$ where $U \in \{0, 1\}$ and $E[U] = p$, the average of i.i.d. samples deviates from the mean by more than ϵ with bounded probability

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} u_i - p\right| > \epsilon\right) \leq 2e^{-2n\epsilon^2}$$

  I.e., the average *concentrates around the mean* with high probability.

- Apply Hoeffding with $u_i =$_____ to bound

$$P\left(|\text{err}_D(h) - \text{err}_P(h)| > \epsilon\right) \leq$$

3. Generalization Error of ERM (finite H)

- Explain or fill in each step (hint: use union bound)

$$P\left(|\text{err}_D(h_D) - \text{err}_P(h_D)| > \epsilon\right) \leq P\left(\max_{h \in H} |\text{err}_D(h) - \text{err}_P(h)| > \epsilon\right)$$

$$=$$

$$\leq \sum_{h \in H} P\left(|\text{err}_D(h) - \text{err}_P(h)| > \epsilon\right)$$

$$\leq$$

- Fix the probability to be less than δ, and derive expression for ϵ

$$\epsilon =$$

4. Interpretation: Tradeoff

- Derive the following bound:

$$\text{err}_P(h) =$$

$$\leq \text{err}_D(h) + |\text{err}_D(h) - \text{err}_P(h)|$$

- Conclude (previous slide) that with probability at least $1 - \delta$:

$$\text{err}_P(h_D) \leq \underbrace{\text{err}_D(h_D)}_{(a)} + \underbrace{\qquad\qquad}_{(b)}$$

- This PAC ("probably approximately correct") bound reflects the trade-off between
  - (a) Training error (smaller when H is larger)
  - (b) Complexity of H
- Occam's Razor: Prefer the simplest hypothesis that fits the data.

5. Infinite Hypothesis Spaces

- If H is the set of all linear classifiers, how big is H?
- New idea: effective number of hypotheses measures all the ways H can label the training data D
- $H[D]$ = the set of all possible predictions on training data:

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

- Question: what are the maximum and minimum possible sizes of $H[D]$ for n training data points?

6. Generalization Error Bound: Infinite H

- General upper bound in terms of **Vapnik-Chervonenkis (VC) Dimension** of a hypothesis class $d_{VC}(H)$

$$\max_{|D|=n} |H[D]| \leq (ne/d_{VC}(H))^{d_{VC}(H)}$$

- VC Dimension is well known for many H
  - Linear classifiers on d features: $d_{VC} = d$
  - Linear classifiers with bias: $d_{VC} = d + 1$
  - Linear classifiers with margin γ on data with $\|x_i\| \leq R$: $d_{VC} = R^2/\gamma^2$
- Can derive a PAC ("probably approximately correct") bound
- For H with VC dimension d_{VC}, given n training data points D, with probability at least $1 - \delta$:

$$\text{err}_P(h_D) \leq \underbrace{\text{err}_D(h_D)}_{(a)} + \underbrace{\sqrt{\frac{d_{VC}\log(2n/d_{VC}) + 1 + \log(1/\delta)}{4n}}}_{(b)}$$