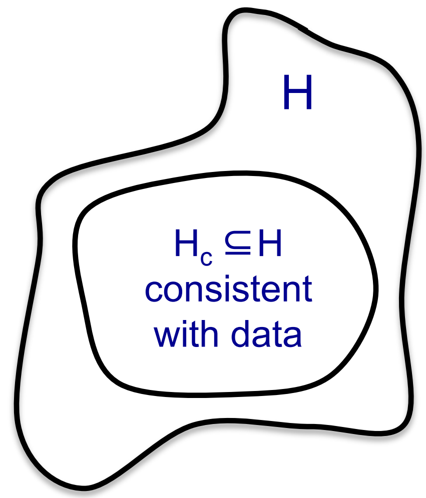
How big should your validation set be?

- In PS1, you tried many configurations of your algorithms (avg vs. regular perceptron, max # of iterations) and chose the one that had smallest validation error
- Suppose in total you tested **|H|=**40 different classifiers on the validation set of **m** held-out e-mails
- The best classifier obtains 98% accuracy on these **m** e-mails!!!
- But, what is the true classification accuracy?
- How large does **m** need to be so that we can guarantee that the best configuration (measured on validate) is truly good?

A simple setting…



- Classification
  - m data points
  - **Finite** number of possible hypothesis (e.g., 40 spam classifiers)
- A learner finds a hypothesis *h* that is **consistent** with training data
  - Gets zero error in training: $error_{train}(h) = 0$
  - I.e., assume for now that one of the classifiers gets 100% accuracy on the **m** e-mails (we'll handle the 98% case afterward)
- What is the probability that *h* has more than ε **true** error?
  - $error_{true}(h) \geq \varepsilon$

How likely is a **bad** hypothesis to get *m* data points right?

- Hypothesis *h* that is **consistent** with validate data
  - got *m* i.i.d. points right
  - h "bad" if it gets all this data right, but has high true error
  - What is the probability of this happening?

- Probability that *h* with $error_{true}(h) \geq \varepsilon$ classifies a randomly drawn data point correctly:

  1. Pr(h gets data point *wrong* | $error_{true}(h) = \varepsilon) = \varepsilon$ — E.g., probability of a biased coin coming up tails

  2. Pr(h gets data point *wrong* | $error_{true}(h) \geq \varepsilon) \geq \varepsilon$

  3. Pr(h gets data point *right* | $error_{true}(h) \geq \varepsilon) = 1$ - Pr(h gets data point *wrong* | $error_{true}(h) \geq \varepsilon)$
     $\leq 1 - \varepsilon$

- Probability that *h* with $error_{true}(h) \geq \varepsilon$ gets *m* iid data points correct:

$$Pr(h \text{ gets m } \textit{iid} \text{ data points right} \mid error_{true}(h) \geq \varepsilon) \leq (1-\varepsilon)^m \leq e^{-\varepsilon m}$$

E.g., probability of m biased coins coming up heads

Are we done?

$$\Pr(h \text{ gets } m \textit{ iid} \text{ data points right} \mid error_{true}(h) \geq \varepsilon) \leq e^{-\varepsilon m}$$

- Says "if h gets m data points correct, then with very high probability (i.e. $1-e^{-\varepsilon m}$) it is close to perfect (i.e., will have error $\leq \varepsilon$)"
- This only considers **one** hypothesis!
- Suppose 1 billion classifiers were tried, and each was a *random* function
- For **m** small enough, one of the functions will classify all points correctly – but all have very large true error

How likely is learner to pick a bad hypothesis?

$\boxed{\Pr(h \text{ gets m } \textit{iid} \text{ data points right} \mid \text{error}_{true}(h) \geq \varepsilon) \leq e^{-\varepsilon m}}$

Suppose there are $|H_c|$ hypotheses consistent with the m data points

- How likely is learner to pick a bad one, i.e. with *true* error $\geq \varepsilon$?
- We need a bound that holds for all of them!

$$P(\text{error}_{true}(h_1) \geq \varepsilon \text{ OR } \text{error}_{true}(h_2) \geq \varepsilon \text{ OR } \ldots \text{ OR } \text{error}_{true}(h_{|H_c|}) \geq \varepsilon)$$

$\leq \sum_k P(\text{error}_{true}(h_k) \geq \varepsilon)$ ← Union bound

$\leq \sum_k (1-\varepsilon)^m$ ← bound on individual h_js

$\leq |H|(1-\varepsilon)^m$ ← $|H_c| \leq |H|$

$\leq |H|\, e^{-m\varepsilon}$ ← $(1-\varepsilon) \leq e^{-\varepsilon}$ for $0 \leq \varepsilon \leq 1$

Generalization error of finite hypothesis spaces [Haussler '88]

We just proved the following result:

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

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

Using a PAC bound

Typically, 2 use cases:

- 1: Pick ε and δ, compute *m*
- 2: Pick m and δ, compute ε

Argument: Since for all *h* we know that

$$P(\text{error}_{true}(h) > \epsilon) \le |H|e^{-m\epsilon}$$

… with probability 1-δ the following holds… (either case 1 or case 2)

$$p(\text{error}_{true}(h) \ge \epsilon) \le |H|e^{-m\epsilon} \le \delta$$

Says: we are willing to tolerate a δ probability of having ≥ ε error

$$\ln\left(|H|e^{-m\epsilon}\right) \le \ln\delta$$

$$\ln|H| - m\epsilon \le \ln\delta$$

$\epsilon = \delta = .01, |H| = 40$

Need $m \ge 830$

Case 1

$$m \ge \frac{\ln|H| + \ln\frac{1}{\delta}}{\epsilon}$$

Case 2

$$\epsilon \ge \frac{\ln|H| + \ln\frac{1}{\delta}}{m}$$

Log dependence on |H|, OK if exponential size (but not doubly)

ε has stronger influence than δ

ε shrinks at rate O(1/m)

Limitations of Haussler '88 bound

- There may be no consistent hypothesis h (where $error_{train}(h)=0$)

- Size of hypothesis space
  - What if |H| is really big?
  - What if it is continuous?

- First Goal: Can we get a bound for a learner with $error_{train}(h)$ in the data set?

Question: What's the expected error of a hypothesis?

- The probability of a hypothesis incorrectly classifying: $\sum_{(\vec{x},y)} p(\vec{x},y) 1[h(\vec{x}) \neq y]$
- Let's now let Z_i^h be a random variable that takes two values, 1 if h correctly classifies data point i, and 0 otherwise
- The Z variables are **independent** and **identically distributed** (i.i.d.) with

$$\Pr(Z_i^h = 0) = \sum_{(\vec{x},y)} p(\vec{x},y) 1[h(\vec{x}) \neq y]$$

- Estimating the true error probability is like estimating the parameter of a coin!

- **Chernoff bound**: for *m* i.i.d. coin flips, $X_1,\ldots,X_m$, where $X_i \in \{0,1\}$. For $0<\varepsilon<1$:

$$p(X_i = 1) = \theta$$

$$\boxed{P\left(\theta - \frac{1}{m}\sum_i x_i > \epsilon\right) \leq e^{-2m\epsilon^2}}$$

True error probability

Observed fraction of points incorrectly classified

$$E\left[\frac{1}{m}\sum_{i=1}^m X_i\right] = \frac{1}{m}\sum_{i=1}^m E[X_i] = \theta$$

(by linearity of expectation)

Generalization bound for |H| hypothesis

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

$$\Pr(\mathrm{error}_{true}(h) - \mathrm{error}_D(h) > \epsilon) \leq |H|e^{-2m\epsilon^2}$$

Why? Same reasoning as before. Use the Union bound over individual Chernoff bounds

PAC bound and Bias-Variance tradeoff

for all h, with probability at least 1-δ:

$$\text{error}_{true}(h) \leq \underbrace{\text{error}_D(h)}_{\text{“bias”}} + \underbrace{\sqrt{\frac{\ln|H| + \ln\frac{1}{\delta}}{2m}}}_{\text{“variance”}}$$

- For large |H|
  - low bias (assuming we can find a good h)
  - high variance (because bound is looser)
- For small |H|
  - high bias (is there a good h?)
  - low variance (tighter bound)

PAC bound: How much data?

$$\Pr(\text{error}_{true}(h) - \text{error}_D(h) > \epsilon) \leq |H|e^{-2m\epsilon^2}$$

$$\text{error}_{true}(h) \leq \text{error}_D(h) + \sqrt{\frac{\ln|H| + \ln\frac{1}{\delta}}{2m}}$$

- Given δ,ε how big should m be?

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