Recap: sample complexity (Chap 3)

We proved: For any class H, distribution D, if

$$\#\text{examples} > \frac{2}{\varepsilon}[\log_2(2H[2m]) + \log_2(1/\delta)]$$

then w.h.p., all bad (error $> \varepsilon$) hypotheses in H are inconsistent with data.

Proof idea:

- Event A: there exists bad h with 0 error on training set.
- Event B: there exists h with 0 error on training set and error $> \varepsilon/2$ on subsequent test set of m points.
- $A \Rightarrow B$ w.h.p., so suffices to show B is unlikely.
- Save randomness by postponing training/test split:
  - First draw $2m$ points. This defines at most $H[2m]$ hypotheses.
  - Then randomly split into training and test sets.

Recap: Hoeffding/Chernoff bounds (Appendix)

For a fixed hypothesis, what is the probability that the observed error is off by additive or multiplicative ε from the true error?

Let $m =$ number of examples, $E = pm =$ expected number of mistakes.

- (Hoeffding) The probability the observed error is off by an additive ε from its expectation is at most $2e^{-2m\varepsilon^2}$.
- (Chernoff) The probability the observed error is greater than $(1+\varepsilon)$ times its expectation is at most $e^{-E\varepsilon^2/3}$. The probability that it is less than $(1-\varepsilon)$ times its expectation is at most $e^{-E\varepsilon^2/2}$.

Can combine with previous argument to get results of the form "how much data do I need so that all $h \in H$ have their observed error near to their true error?" (see homework).

Recap: Weak and Strong learning (Chap 4)

If we can get a weak bias over *every* distribution, then can get a strong bias over every distribution.

Idea: keep modifying distribution to extract new information.

Adaboost:

- Modify the distribution so that the previous hypothesis is 50/50. Do this by multiplying weight on correct points by $\beta = error/(1 - error)$.

- Then take "weighted vote" over the hypotheses. Weight of h_t is proportional to $\ln(1/\beta_t)$.

- Use upper and lower bounds on total weight of example points to bound the error of the combination. (We analyzed fixed β in class.)

Learning when there is no perfect hypothesis

- Hoeffding/Chernoff bounds allow "information-theoretic" bounds to carry over. Give guarantees for when minimizing training error will also approximately minimize true error.

- What about polynomial-time algorithms? Seems harder.
  - Given data set S, finding conjunction with fewest mistakes is NP-hard.
  - Open problem: can you weak-learn (find a poly-time-computable hypothesis with error $< 1/2-\varepsilon$) given only the assumption that the data is 90% consistent with some conjunction?

- One way to make progress: make assumptions on the "noise" in the data. E.g., Random Classification Noise model.

Learning from Random Classification Noise (Ch 5)

- PAC model, but assume labels coming from noisy channel.
- "noisy" Oracle $EX^\eta(c, D)$. η is the *noise rate*.

  Example x is drawn from D.

  With probability $1 - \eta$ see correct label $c(x)$.

  With probability η see incorrect label $1 - c(x)$.
- E.g., if h has non-noisy error p, what is the noisy error rate?

Algorithm A **PAC-learns** C **with random classification noise** if for any $c \in C$, any distribution D, any $\eta < 1/2$, any $\varepsilon, \delta > 0$, given access to noisy examples from $EX^\eta(c, D)$, A find a hypothesis h that is ε-close to c, with probability $1 - \delta$. Allowed time

$$poly\left(\frac{1}{\varepsilon}, \frac{1}{\delta}, \frac{1}{1-2\eta}, n, size(c)\right).$$

Q: Is this a plausible goal? I mean, we are asking the learner to get closer to c than the data is.

A: OK because noisy error rate is linear in true error rate (squashed by $1 - 2\eta$).

Example: Learning monotone conjunctions

Let's assume η is known.

Any ideas?

Example: Learning monotone conjunctions

- **Idea:** Let
$$p_i = \Pr_{x \leftarrow EX(c,D)}[c(x) = 1 \wedge x_i = 0].$$
Any hypothesis conjunction that includes all x_i such that $p_i = 0$ and no x_i such that $p_i > \varepsilon/n$ is good.

- So, just need to estimate this probability to additive error $\varepsilon/2n$. Can rewrite as
$$\Pr_{x \leftarrow EX(c,D)}[c(x) = 1 | x_i = 0] \cdot \Pr[x_i = 0]$$

- Second part is unaffected by noise.

- Let q_i be the first part. Use fact that:
$$\begin{aligned} \Pr_{x \leftarrow EX^{\eta}(c,D)}[\ell(x) = 1 | x_i = 0] &= q_i(1-\eta) + (1-q_i)\eta \\ &= \eta + q_i(1-2\eta). \end{aligned}$$
So, can approximate q_i from observations (assuming $\Pr[x_i = 0]$ is not too tiny).

Open problem

Can noise tolerance be boosted?

- Say for concept class C there exists alg A such that for any $c \in C$, any distribution $\mathcal{D}$, any $\eta < 0.1$, A PAC-learns from $EX^{\eta}(c, D)$.

  Does this imply there must exist an algorithm B that succeeds for all $\eta < 1/2$, with running time $poly(\frac{1}{1-2\eta})$?

- Or, could ask the same question for fixed $\mathcal{D}$ too. E.g., a counterexample would be a C and D under which it is possible to achieve the former but not the latter.

What was the basic idea?

- See how we could learn in non-noisy model by an algorithm that just asked about probabilities of certain events with some "slop".
- Try to estimate these probabilities from noisy data by breaking into
  - parts predictably affected by noise.
  - parts unaffected by noise

Next topic:

- Formalize this in the notion of a "statistical query algorithm".
- Show how any SQ algorithm can be used to learn with classification noise.
- Can actually characterize the kinds of things that can or can't be done with SQ algorithms.

the Statistical Query Model

- No noise.

- Algorithm asks "what is the probability a labeled example will have property χ? Please tell me up to additive error τ."

- Formally, χ is a $\{0, 1\}$ function on labeled examples. Must be poly-time computable. $\tau \geq 1/poly(\cdots)$.

  If $P_\chi = \mathrm{Pr}_{c,D}(\chi(x, \ell) = 1)$ then world responds with $\hat{P}_\chi \in [P_\chi - \tau, P_\chi + \tau]$.

- May repeat this $poly(\cdots)$ many times. Algorithm may also ask to see $poly(\cdots)$ many unlabeled examples.

- Algorithm must output a hypothesis with error less than ε. (No δ in this model.)