COMS 6998-4 F17 Homework 2 (due Monday October 30)

Daniel Hsu (djh2164)

Instructions

Solve **one** of the assigned problems. If you solve both problems *perfectly*, then you will receive some "extra credit". But only if both solutions are *perfect*. Also remember homework is only 10% of the total grade. So don't fret too much.

Submit the assignment on Gradescope as a PDF document by 11:59 PM of the due date.

Make sure the following appears at the top of the first page of your write-up:

- your name,
- your UNI, and
- the names and UNIs of any students with whom you discussed the assignment.

You are welcome to use the Markdown or LaTeX source for the assignment as a template for your write-up. I use Pandoc http://pandoc.org to translate the Markdown to LaTeX and ultimately to PDF.

Problem 1

Assume you are given: (i) an algorithm for the online classification problem with mistake bound M, (ii) target error rate ϵ, and (iii) confidence parameter δ.

Consider the following "online-to-batch" conversion procedure.

1. Simulate the online learner on the first $n_1(M, \epsilon, \delta)$ rounds, and save the collection of hypotheses produced by the online learner. You can assume that if the learner doesn't make a mistake in some round, then it uses the same hypothesis in the following round.
2. Evaluate each of the hypotheses saved from the first step in the next $n_2(M, \epsilon, \delta)$ rounds, and return the hypothesis that makes the fewest mistakes.

Determine the sample sizes $n_1(M, \epsilon, \delta)$ and $n_2(M, \epsilon, \delta)$ so that with probability at least $1 - \delta$, a hypothesis of error rate at most ϵ is returned. Give a detailed analysis that proves your claim. The total sample size should be an improvement over that of the online-to-batch procedure we discussed in lecture at least for large enough M and small enough ϵ and δ.

Your solution:

Problem 2

Recall the "selective ERM" algorithm from lecture on October 4, reproduced below.

$S_0 := \emptyset$, $h_0 := \mathcal{A}(S_0)$, $\beta_0 := \infty$.

For $n = 1, 2, \ldots$:

- Get x_n.
- Predict $a_n := h_{n-1}(x_n)$.
- Get $h'_{n-1} := \mathcal{A}(S_{n-1}, (x_n, a_n))$.
- If $\mathrm{err}_{S_{n-1}}(h'_{n-1}) \leq \mathrm{err}_{S_{n-1}}(h_{n-1}) + \beta_{n-1}$, then:
  - Get y_n.
  - Let $S_n := S_{n-1} \cup \{(x_n, y_n)\}$.
- Else: Let $S_n := S_{n-1} \cup \{(x_n, a_n)\}$.
- Let $h_n := \mathcal{A}(S_n)$, $\beta_n := \mathrm{Rad}_{n,P}(H \cup -H) + \sqrt{\frac{2\log(n(n+1)/\delta)}{n}}$.

Let $h^* \in H$ be a fixed hypothesis of minimum error rate with respect to P. Also define:

$$\rho(h, h') := \Pr_{(x,y)\sim P}(h(x) \neq h'(x)), \qquad \tilde{\rho}(h^*, h) := \mathrm{err}_P(h) - \mathrm{err}_P(h^*),$$

$$B(h, r) := \{h' \in H : \rho(h, h') \leq r\}, \qquad \tilde{B}(h^*, r) := \{h \in H : \tilde{\rho}(h^*, h) \leq r\},$$

$$D(h^*, r) := \{x \in \mathcal{X} : \exists h \in B(h^*, r) \centerdot h(x) \neq h^*(x)\}, \quad \tilde{D}(h^*, r) := \{x \in \mathcal{X} : \exists h \in \tilde{B}(h^*, r) \centerdot h(x) \neq h^*(x)\}.$$

Let $\mathcal{X} := \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ be the unit sphere in $\mathbb{R}^d$, and let H be the class of homogeneous linear separators in $\mathbb{R}^d$. Furthermore, assume the marginal of P over $\mathcal{X}$ is the uniform distribution on $\mathcal{X}$.

- Part 1. Prove that $\tilde{D}(h^*, r) \subseteq D(h^*, 2\,\mathrm{err}_P(h^*) + r)$.
- Part 2. Prove a bound on the expected number of labels queried by a learner using this algorithm after n rounds. The bound should be given only in terms of d, n, and $\mathrm{err}_P(h^*)$.
- Part 3. Suppose a random draw $(x, y) \sim P$ is produced using the following generative process. First, draw x from the uniform distribution on $\mathcal{X}$. Then, a coin with heads probability η is tossed; if heads, set $y := -h^*(x)$, and if tails, set $y := h^*(x)$. Here, $\eta \in (0, 1/2)$ is a fixed constant. Prove a bound on the expected number of labels queried by a learner using this algorithm after n rounds. The bound should be given only in terms of d, n, and η. Presumably it should be much better than the bound you get in Part 2.

Your solution: