# COMPUTATIONAL LEARNING THEORY: ANSWERS TO HOMEWORK 3

VISHVAS VASUKI

*Remark* 0.0.1. Thank you for the wonderful assignment. The solution to the first two parts of problem 4 were especially pleasing.

## 1

**Theorem 1.0.2.** *C is efficiently PAC learnable. Then, there is an efficient algorithm A such that: given a sample of size m labelled according to $c \in C$, A, with probability 1-d finds a hypothesis h that is consistent with all the m points and has size poly(n, size(c), log(m)).*

*Proof.* Proof by construction.

As C is efficiently PAC learnable, there exists an efficient algorithm L which can PAC learn C. So, given error parameter $\epsilon = 0.25$ and confidence parameter $d'$; L is capable of producing, with probability $1 - d'$, a hypothesis h with $error(h) \leq \epsilon$. The advantage of h over random guessing is $g = 0.5 - error(h) \geq 0.5 - \epsilon$.

Given a sample S and confidence parameter d, the algorithm A uses ADAboost algorithm to repeatedly invoke algorithm L and produce a hypothesis consistent with S.

During this process, A always uses L to get a hypothesis h with $error(h) \leq \epsilon = 0.25$ with confidence $1 - d'$, where $d'$ will be specified later. Note that the specified $\epsilon$ determines g.

After $\frac{\ln m}{2g^2}$ steps, A produces a hypothesis h of size $|h| = O(\frac{\ln m}{2g^2})|c|$ which is consistent with S. (This follows from the analysis of Adaboost done in class.)

Now we find out what $d'$ must be in order for A to succeed with probability 1-d. A can fail only when one or more of the $\frac{\ln m}{2g^2}$ invocations of L fail to produce an $\epsilon$ close hypothesis. From the union bound, this happens with probability $\frac{\ln m}{2g^2}d'$. Thus, when $d' = \frac{2dg^2}{\ln m}$, A succeds with probability 1-d.

□

## 2

*Remark* 2.0.3. f and h are $\{\pm 1\}$ boolean functions.

### 2.1. a.

**Theorem 2.1.1.** *For any distribution D, h is a weak hypothesis for f with advantage g if and only if $E_{x\sim D}[h(x)f(x)] \geq 2g$.*

*Proof.*

$$\begin{aligned} E_{x\sim D}[h(x)f(x)] &= Pr_{x\sim D}[h(x)f(x) = 1] - Pr_{x\sim D}[h(x)f(x) = -1] \\ &= 1 - 2Pr_{x\sim D}[h(x)f(x) = -1] \end{aligned}$$

h is a weak hypothesis for f if and only if $Pr_{x\sim D}[h(x)f(x) = -1] \leq 2^{-1} + g$; which happens if and only if $1 - 2Pr_{x\sim D}[h(x)f(x) = -1] \geq 2g$; which happens if and only if $E_{x\sim D}[h(x)f(x)] \geq 2g$. ☐

2.2. **b.**

**Theorem 2.2.1.** *f is an LTF.* $f = sgn(w.x)$; $w \in Z^n$; $W = \sum |w_i|$. *Assume:* $\forall x, \langle w, x\rangle \neq 0$. *For any distribution D, there exists an* $x_i$ *such that* $|E_{x\sim D}[f(x)x_i]| \geq W^{-1}$.

*Proof.* Assumption: $w_i \neq 0$. This assumption can be removed by not considering any $x_i$ corresponding to $w_i = 0$.

Now, we will see that an absurd thing will happen if we suppose that $\forall i$ : $|E_{x\sim D}[f(x)x_i]| < W^{-1}$.

Then:

$$\begin{aligned}
W^{-1} &> |E_{x\sim D}[f(x)x_i]| \\
|w_i|W^{-1} &> |w_i||E_{x\sim D}[f(x)x_i]| \\
&\geq w_i E_{x\sim D}[f(x)x_i] \\
&= E_{x\sim D}[f(x)w_i x_i] \\
\therefore \sum_i |w_i|W^{-1} &> \sum_i E_{x\sim D}[f(x)w_i x_i] \\
1 &> E_{x\sim D}[f(x)\sum_i w_i x_i]
\end{aligned}$$

$|\langle w, x\rangle| \geq 1$ due to $w_i$ being non zero integers, and due to the fact that $\forall x, \langle w, x\rangle \neq 0$.

So, $E_{x\sim D}[f(x)\sum_i w_i x_i] \geq 1$ as $\langle w, x\rangle$ and f(x) always agree on sign. Thus we have reached an absurdity. ☐

2.3. **c.**

*Question* 2.3.1. What is the weak learner?

*Answer* 2.3.2. Let WL be the weak learner. We have shown earlier that $|E_{x\sim D}[f(x)x_i]| \geq W^{-1}$. So, for each of the n bits, WL finds the bit $argmax_i|E_{x\sim D}[f(x)x_i]|$. If $\max_i E_{x\sim D}[f(x)x_i] > 0$, it uses the corresponding bit $x_i$ as its weak hypothesis h; otherwise it uses $-x_i$.

Note that h has advantage $(2W)^{-1}$. Also, using the Hoeffding ineuqality, for any i, $E_{x\sim D}[f(x)x_i] = Pr_{x\sim D}[f(x)x_i = 1] - Pr_{x\sim D}[f(x)x_i = -1]$ can be estimated whp to arbitrarily high accuracy by taking a sufficiently large polynomial sized sample.

*Question* 2.3.3. How do we apply a boosting algorithm?

*Answer* 2.3.4. We take WL as a black box, and simply apply any convinient boosting algorithm. Eg: ADABoost.

*Question* 2.3.5. What is the output hypothesis?

*Answer* 2.3.6. If ADABoost is used for boosting, the output hypothesis reduces to a halfspace; as all the hypotheses the weak learner returns are of the form $\pm x_i$!

*Question* 2.3.7. For what values of W do we get a polynomial time algorithm?

*Answer* 2.3.8. The individual hypotheses produced by WL are guaranteed to have an advantage of $(2W)^{-1}$. So, boosting makes sense only when W is sub exponential. Also, when ADA boost is used, the number of iterations required is polynomial in W. So, for values of W which are polynomial, we get a polynomial time algorithm.

## 3

*Notation.* $p_S$ denotes the parity function $\chi_S$.

**Theorem 3.0.9.** *$\sum_{|S|\geq d}\hat{f}(S)^2 \leq \epsilon$. $Pr_x(g(x) \neq f(x)) = \eta$. Then, $E[(g(x) - \sum_{|S|<d}\hat{g}(S)p_S(x))^2] \leq E[(g(x) - \sum_{|S|<d}\hat{f}(S)p_S(x))^2]$.*

*Proof.* Using the Fourier expansion of f and g, we see the following:

$$\begin{aligned} E[(g(x) - \sum_{|S|<d}\hat{g}(S)p_S(x))^2] &= \left\| \sum_{|S|\geq d}\hat{g}(S)p_S(x))\right\|^2 \\ &= \sum_{|S|\geq d}\hat{g}(S)^2 \end{aligned}$$

$$\begin{aligned} E[(g(x) - \sum_{|S|<d}\hat{f}(S)p_S(x))^2] &= \left\| \sum_{|S|\geq d}\hat{g}(S)p_S(x)) + \sum_{|S|<d}(\hat{g}(S) - \hat{f}(S))p_S(x)\right\|^2 \\ &= \sum_{|S|\geq d}\hat{g}(S)^2 + \left\| \sum_{|S|<d}(\hat{g}(S) - \hat{f}(S))p_S(x)\right\|^2 \end{aligned}$$

Thence the result. □

**Corollary 3.0.10.** *$\sum_{|S|\geq d}\hat{g}(S)^2 \leq O(\eta + \epsilon)$.*

*Definition* 3.0.11. $f_{<d} = \sum_{|S|<d}\hat{f}(S)p_S(x)$. $f_{\geq d} = \sum_{|S|\geq d}\hat{f}(S)p_S(x)$.

*Proof.* Note that:

$$\begin{aligned} \|g - f\|^2 &= E_x[(g(x) - f(x))^2] \\ &= 4Pr_x(f(x) \neq g(x)) \\ &= 4\eta \end{aligned}$$

Using the theorem proved earlier:

$$\begin{aligned}
\sum_{|S|\geq d} \hat{g}(S)^2 &\leq \|g - f_{<d}\|^2 \\
&= \|g - f + f_{\geq d}\|^2 \\
&= \|g - f\|^2 + \|f_{\geq d}\|^2 + 2\langle f_{\geq d}, g - f\rangle \\
&\leq \|g - f\|^2 + \|f_{\geq d}\|^2 + 2\,\|g - f\|\,\|f_{\geq d}\| \\
&\leq 4\eta + \epsilon + 4\sqrt{\eta\epsilon} \\
&\leq 4\eta + \epsilon + 4\max(\eta, \epsilon) \\
&= O(\eta + \epsilon)
\end{aligned}$$

□

## 4

**Theorem 4.0.12.** *f is a monotone boolean function. Flipping a bit $x_i$ from 1 to -1 cannot cause f(x) to flip from -1 to 1. Then, $I_i(f) = \hat{f}(\{x_i\})$.*

*Proof.* Let $x' \in \{1, -1\}^{n-1}$ be a variable corresponding to parts of the string x without $x_i$. Let $g(x_i, x') = f(x)$ for all values of $x_i$ and $x'$.

Note that, for any $x'$, $g(-1, x') = 1 \wedge g(1, x') = -1$ can never happen as g and f are montonic. So, the only way $g(-1, x') = -g(1, x')$ can happen is when $g(-1, x') = -1 \wedge g(1, x') = 1$. Also, for this reason, $E_{x'}[g(1, x') - g(-1, x')] = 2Pr_{x'}(g(1, x') = 1 \wedge g(-1, x') = -1)$.

So,

$$\begin{aligned}
I_i(f) &= Pr_x[f(x) \neq f(x^{(i)})] \\
&= Pr_{x'}[g(1, x') \neq g(-1, x')] \\
&= Pr_{x'}(g(1, x') = 1 \wedge g(-1, x') = -1) \\
&= 2^{-1} E_{x'}[g(1, x') - g(-1, x')] \\
&= 2^{-1}(E_{x'}[g(1, x')] - E_{x'}[g(-1, x')])
\end{aligned}$$

But,

$$\begin{aligned}
\hat{f}(\{i\}) &= E_x[f(x)x_i] \\
&= Pr_x(x_i = 1)E_x[f(x)|x_i = 1] - Pr_x(x_i = -1)E_x[f(x)|x_i = -1] \\
&= 2^{-1}(E_{x'}[g(1, x')] - E_{x'}[g(-1, x')]) \\
&= I_i(f)
\end{aligned}$$

□

**Corollary 4.0.13.** *The sum of influences of any monotone function is at most $\sqrt{n}$.*

*Proof.* This follows from the inequality between 1-norm and 2-norm.

$\sum_i I_i(f) = \sum_i \hat{f}(\{i\}) \leq \sqrt{n} \sum_i (\hat{f}(\{i\})^2)^{1/2} \leq \sqrt{n}$. □

**Corollary 4.0.14.** *Sum of influences of the majority function is* $\approx n(\frac{2}{(n-1)\pi})^{0.5}$.

*Proof.* Consider the influence of a single variable. Flipping a single variable can make a difference only when there is a tie between the votes of the remaining variables. This can happen with probability $2^{-(n-1)}\frac{(n-1)!}{(\frac{n-1}{2})!(\frac{n-1}{2})!}$. Using Stirling's approximation and multiplying by n; we get the above mentioned estimate. □