# COMPUTATIONAL LEARNING THEORY: ANSWERS TO HOMEWORK 2

VISHVAS VASUKI

## 1

*Definition* 1.0.1. A sober algorithm in the Mistake bounded model is an algorithm which, whenever it makes a mistake, alters its hypothesis to be consistent with all the examples it has seen so far.

**Lemma 1.0.2.** *If a Concept class C is efficiently learnable in the mistake bound model, it can be learnt by an efficient sober algorithm.*

*Proof.* Let L be a non-sober algorithm which learns C with mistake bound b. So, upon making a mistake on a certain example, L does not update its hypothesis to be consistent with all the examples it has seen so far.

Suppose that, at the end of a certain sequence S of examples, L makes a mistake on x. Also suppose that L does not update its hypothesis at all; or even if it does, its new hypothesis is still not consistent with S. Then, we can conceive of L repeatedly receiving the same sequence S, causing it to exceed its claimed mistake bound b.

Hence, after seeing S a some finite number of times, if L is to stay within its mistake bound, it must update its hypothesis to be consistent with S.

The efficient sober algorithm L' is obtained by executing such an update whenever a mistake is seen. □

**Theorem 1.0.3.** *If C is efficiently learnable in the mistake bounded learning model, it is efficiently learnable in the PAC learning model.*

*Proof.* If C is efficiently learnable in the mistake bounded model, then there exists an efficient sober algorithm L which learns it.

Suppose that we run L on a set S of $m = \Omega(\epsilon^{-1}\log(\delta^{-1}) + \frac{d}{\epsilon}\log(\epsilon^{-1}))$ examples. Then, L ends up with a hypothesis h consistent with S. Now, we use one of the VCD/ Occam razor theorems to conclude that, with probabilility $\delta^{-1}$, h is $\epsilon$ close to the target concept. Further more, L is efficient. □

## 2

*Acknowledgement.* The lower bound is due to Sungho Yun.

**Theorem 2.0.4.** *Let C be the conjunctions on n literals. VCD for C is n.*

*Remark* 2.0.5. C includes the conjunction of size 0, which accepts every possible example.

*Proof.* Let $e_i$ represent a sample with the bit $x_i = 1$, and with all other bits $x_j = 0$. All conjunctions are assumed to have no redundant variables.

The set $S = \{e_1, ..e_n\}$ is shattered by C. The reason is as follows: Consider any $S' \subset S$. This subset has exactly $n - |S'|$ bit values in common; and all $e_i \in S - S'$

disagrees with atleast one of these bit values. Corresponding to any set of bit values, there is a conjunction. Hence the dichotomy which labels S' and S-S' differently is realizable.

Suppose that a set $S = \{s_1, ..s_t\}$ was shattered by C. Let the conjunction which selects only $s_i$ in S be called $c_i$. Think of conjunctions as a set of variables involved in the conjunction's expression. Consider the conjunction which admits $s_1, ..s_i$ and rejects everything else in S. This conjunction is $C_i = \bigcap_{j=1}^{i} c_j$. Suppose that the number of literals in $C_i$ is $L_i$. We note that $L_{i+1} \leq L_i$. But, if $L_i = L_{i-1}$, we will be violating the assumtion that $C_i$ admits $s_1, ..s_i$ and rejects $s_{i+1}$. Hence, we observe that $L_{i+1} \leq L_i - 1$.

But, $L_1 \leq n$. And we know that $L_t \geq 0$. Hence, we conclude that $t \leq n$. □

3

**Theorem 3.0.6.** *$C$ = axis aligned rectangles. $C$ is efficiently agnostically learnable.*

*Proof.* Consider an algorithm which samples a set of S points, uses exhaustive search to find the rectangle with the least error rate. It then uses it as the hypothesis.

To bound the size of S, we use one of the VCD-style bounds: $|S| = \Omega(\epsilon^{-2}(4 + \log \frac{1}{\delta}))$. Then, with probability $1 - \delta$, we are $\epsilon$ close to the rectangle with the least error over all possible samples.

This algorithm is efficient as it takes time polynomial in $|S|$. □

4

Distribution used: Uniform distribution on the unit sphere in $R^n$: $S^{n-1}$.

Let c be the target origin centered halfspace defined by vector u and hyperplane $u_\perp$.

Algorithm L for classifying point p:

**Input**: p
**Output**: The label h(p)
Choose set S of labeled points uniformly at random.
Use this net of points to classify a future random point p via m-nearest-neighbor. Let $S' = \{p' : p' \in S, \langle p', p\rangle \geq 0\}$. Return $h(p) = maj_{p' \in S'}(sgn(\langle p', u\rangle))$. Let $|S'| = m$

**Theorem 4.0.7.** *Let $m = poly(n, \epsilon^{-1})$. $Pr_{p,S}(h(p) \neq sgn(\langle p, u\rangle)) \leq \epsilon$.*

*Proof.* Without loss of generality, suppose that p is in the halfspace. The analysis for p not in the halfspace will be identical.

Let t be the angle between p and the hyperplane $u_\perp$.

Consider some $p' \in S'$; so its angle with p is less than $\frac{\pi}{2}$. $sgn(\langle p', u\rangle) < 0$ iff its angle with u $> \frac{\pi}{2}$; that is, if it is in the space $p_\perp - u_\perp$. We use the following fact:

*Fact* 4.0.8. If angle between u, p is $\frac{\pi}{2} - t$, $Pr(sgn(\langle u, x\rangle) \neq sgn(\langle p, x\rangle)) = \frac{\frac{\pi}{2} - t}{\pi} = \frac{\pi - 2t}{2\pi}$.

So, using symmetry, we find that Pr(p' is misleading) $= Pr_{S'}(sgn(\langle p', u\rangle) < 0) = \frac{\pi - 2t}{4\pi}$.

For our algorithm to predict erroneously, at least m/2 points in S' should be misleading. The expected number of misleading points in S' is $\frac{\pi-2t}{4\pi}$.

Using the Chernoff bound and some algebra, we find the probability that this is over $\frac{m}{2}$ to be $e^{(\frac{\pi+2t}{4\pi}-\frac{1}{2}\ln(\frac{2\pi}{\pi-2t}))m}$. We see that this quantity is less than $\epsilon/4$ when $m \geq \frac{\ln\frac{\epsilon}{4}}{(-\frac{\pi+2t}{4\pi}+\frac{1}{2}\ln(\frac{2\pi}{\pi-2t}))}$.

*Fact* 4.0.9. For small a: $Pr_p(t < a) \leq an^{0.5}$.

So, for $a = \frac{\epsilon}{4n^{0.5}}$, $Pr_p(t < a) \leq \frac{\epsilon}{4}$. By the way of worst case analysis, let us assume that whenever $t < a$, L fails.

Hence, using the union bound, we see that Pr(L fails when given p inside the halfspace) $\leq \frac{\epsilon}{2}$ if we choose $m = \frac{\ln\frac{\epsilon}{4}}{(-\frac{\pi+2a}{4\pi}+\frac{1}{2}\ln(\frac{2\pi}{\pi-2a}))}$.

By symmetry, Pr(L fails when given p outside the halfspace) $\leq \epsilon/2$ for the same value of m.

Thus, for m polynomial in n and $\epsilon^{-1}$, L fails with probability $\epsilon$. □