Notes 9: Occam’s Razor

1. Hypothesis class

**Hypothesis class** $\mathcal{H} = \text{set of hypotheses the learning algorithm may output}

Usually $\mathcal{H} \supseteq \mathcal{C}$, but can sometimes be bigger

e.g. Winnow1 learns $\mathcal{C} = \{k\text{-sparse monotone disjunctions}\}$ using $\mathcal{H} = \{\text{LTFs with } \geq 0\text{ weights}\}$

**Proper learning**: Algorithm required to output $h \in \mathcal{C}$, i.e. $\mathcal{H} = \mathcal{C}$

**Improper learning**: Algorithm allowed to output $h \not\in \mathcal{C}$, i.e. $\mathcal{H} \supseteq \mathcal{C}$

2. Consistent hypotheses

Fix concept class $\mathcal{C}$ and finite hypothesis class $\mathcal{H}$

**Consistent Hypothesis Algorithm**

Given labelled samples, output any $h \in \mathcal{H}$ consistent with all samples

Call hypothesis $h$ bad if $\text{err}_D(h, c) \geq \varepsilon$

**Theorem 2.1.** For any distribution $\mathcal{D}$ over instance space $X$, given $m$ independent samples from $EX(c, D)$, if $m \geq \frac{1}{\varepsilon} \ln(|\mathcal{H}|/\delta)$, then

$$\Pr[\text{some bad hypothesis consistent with all samples}] \leq \delta$$

Better bound than Halving Algorithm + Online-to-PAC conversion

**Proof.** For any bad $h \in \mathcal{H}$

$$\Pr[h \text{ consistent with all } m \text{ samples}] \leq (1 - \varepsilon)^m \leq e^{-\varepsilon m} = \delta/|\mathcal{H}|$$

Union bound:

$$\Pr[\text{some bad hypothesis consistent with all samples}] \leq |\mathcal{H}| \cdot (\delta/|\mathcal{H}|) = \delta$$

In other words, $|\mathcal{H}| \leq \delta e^{\varepsilon m}$

**Occam’s Razor**: Scientific principle to favour simpler hypotheses

PAC learning algorithm due to small hypothesis class

Simple hypothesis $\approx$ hypothesis with short description $\approx$ small number of hypotheses

3. PAC learning sparse disjunctions

$\mathcal{C} = \{\text{disjunctions}\}$ over $X = \{0, 1\}^n$ \quad $s \overset{\text{def}}{=} \text{size}(c)$

How to PAC learn $\mathcal{C}$ efficiently?

1. Elimination Algorithm + Online-to-PAC conversion: $O\left(\frac{n}{\varepsilon} \ln \frac{n}{\delta}\right)$ samples
   $\approx \frac{n}{\varepsilon}$ ignoring log factors

2. Winnow1 + Online-to-PAC conversion: $O\left(\frac{s \ln n}{\varepsilon} \ln \frac{s \ln n}{\delta}\right)$ samples
   $\approx \frac{s}{\varepsilon}$ ignoring log factors

Better dependence on $n$; Good for small $s$

But improper

3. Consistent Hypothesis Algorithm: $O\left(\frac{s}{\varepsilon} \ln \frac{n}{\delta}\right)$ samples

Because $|\mathcal{H}| = \binom{n}{s} 2^s \leq (2n)^s$ \quad ($\mathcal{H} \overset{\text{def}}{=} \{s\text{-sparse disjunctions}\}$)

Even better dependence on $n$ and $s$

But inefficient! (need $|\mathcal{H}| \approx n^s$ time, not poly$(n, 1/\varepsilon, 1/\delta, s)$)

4. (Below) efficient algorithm using $O\left(\frac{1}{\varepsilon} (\ln \frac{1}{\delta} + s \ln \frac{1}{\delta} \ln n)\right)$ samples

$\approx \frac{s}{\varepsilon}$ ignoring log factors; Good dependence on $s$ and $n$

Idea 1: Find consistent disjunction quickly using Greedy Heuristic for Set Cover

Idea 2: Further reduce $|\mathcal{H}|$ by hypothesis testing
4. Set Cover

A computational problem (not originated from learning)

**Input:** Universe $U = \{1, \ldots, m\}$ of $m$ elements and subsets $S_1, \ldots, S_k \subseteq U$

**Goal:** Find smallest collection $S_{i_1}, \ldots, S_{i_k}$ of given subsets to cover $U$ (i.e. $S_{i_1} \cup \cdots \cup S_{i_k} = U$)

Set Cover is NP-hard (as hard as thousands other problems conjectured to be intractable)

We settle for an approximation algorithm that outputs a nearly optimal solution

---

**Theorem 4.1.** Greedy Heuristic always outputs a cover with $\leq \text{OPT} \cdot \ln m$ many sets

**Proof.** Let $T_t \subseteq U$ denote set of uncovered elements after iteration $t$ (initially $T_0 = U$)

**Claim:** Largest subset $S_{i_t}$ at iteration $t$ covers $\geq 1/\text{OPT}$ fraction of $T_{t-1}$

**Reason:** Uncovered elements are covered by OPT sets; largest set must cover $\geq 1/\text{OPT}$ fraction

Using Claim,

$$|T_t| \leq \left(1 - \frac{1}{\text{OPT}}\right)|T_{t-1}| \leq \ldots \leq \left(1 - \frac{1}{\text{OPT}}\right)^t m < e^{-t/\text{OPT}}m$$

$$\leq 1 \quad \text{if } t \geq \text{OPT} \cdot \ln m \quad \square$$

---

Elimination Algorithm + conversion only uses negative samples

Keep removing literals $x_i$ or $\overline{x}_i$ that contradicts a negative sample

All literals in $c$ are also in $h$  (h automatically consistent with all positive samples)

Improved algorithm further uses positive samples to shorten $h$  (and hence shrink $\mathcal{H}$)

Find a few literals to “explain” (i.e. cover) all positive samples

$c$ contains $s$ literals, all positive samples can be “covered” with $s$ literals

Can quickly find a cover using $s \ln m$ literals  ($m = \#\text{positive samples}$)

---

**Improved algorithm**

\[ \{y_1, \ldots, y_r\} = \text{set of literals that are consistent with all negative examples} \]

i.e. if literal $y_i$ is true in some negative sample, then $y_i$ is excluded

For $1 \leq i \leq r$, let $S_i = \text{set of positive samples where } y_i \text{ is true}$

Find a set cover $S_{i_1}, \ldots, S_{i_k}$ using $k = s \ln m$ sets

Hypothesis $h = y_{i_1} \lor \cdots \lor y_{i_k}$

\[ |\mathcal{H}| = \binom{n}{s \ln m} 2^{s \ln m} \leq (2n)^{s \ln m} \]

Need $|\mathcal{H}| \leq \delta e^{cm}$

True if $(2n)^{s \ln m} \leq \delta e^{cm} \iff s(\ln m) \ln 2n + \ln(1/\delta) \leq cm$

Can show that $m \geq \Omega \left(\frac{1}{\epsilon}(\ln(1/\delta) + s(\ln n) \ln(s \ln n))\right)$ suffices  (details omitted)

---

5. Chernoff Bounds

Due to Herman Rubin

$X_1, \ldots, X_m$ independent $\{0,1\}$-valued random variables

s.t. $\mathbb{P}[X_i = 1] = \mathbb{E}[X_i] = p$ for $1 \leq i \leq m$

$X \overset{\text{def}}{=} X_1 + \cdots + X_m$  ($\mathbb{E}[X] = mp$)

**Theorem 5.1** (Multiplicative Chernoff). For all $0 \leq \gamma \leq 1$,

$$\mathbb{P}[X \leq (1 - \gamma)mp] \leq e^{-\frac{1}{2} \gamma^2 mp}$$

$$\mathbb{P}[X \geq (1 + \gamma)mp] \leq e^{-\frac{1}{2} \gamma^2 mp}$$
Also true for $X_1, \ldots, X_m$ independent $[0,1]$-valued (i.e. bounded) random variables
Many proofs; see e.g. Mulzer “Five Proofs of Chernoff’s Bound with Applications” if interested
Exponential decay

6. Hypothesis Testing

Fix $h \in \mathcal{H}$, how can we test whether $h$ is bad? (i.e. $\text{err}_D(h,c) = \mathbb{P}_{x \in \mathcal{D}}[h(x) \neq c(x)] \geq \varepsilon$)

Solution: Draw $m$ independent labelled samples $(x^1, c(x^1)), \ldots, (x^m, c(x^m))$,

Compute (empirical error) $\widehat{\text{err}} \overset{\text{def}}{=} \frac{\# \text{s.t. } h(x) \neq c(x)}{m}$

By Chernoff bound, $\widehat{\text{err}} \approx \text{err}_D(h,c)$

e.g. If $h$ is bad, $p \overset{\text{def}}{=} \text{err}_D(h,c) \geq \varepsilon$,

$$\mathbb{P}\left[\widehat{\text{err}} \leq \frac{\varepsilon}{2}\right] \leq e^{-\frac{1}{2}mp} \leq e^{-\frac{1}{2}\varepsilon m}$$

Further Improved Algorithm: Similar to Improved Algorithm

But only cover $1 - \varepsilon/2$ fraction of positive samples using $S_1, \ldots, S_k$

Number of sets needed $k \leq \text{OPT} \cdot \ln(1/\varepsilon)$ (why?)

Can show that $O\left(\frac{1}{\varepsilon}(\ln\frac{1}{\delta} + s \ln\frac{1}{\varepsilon}\ln n)\right)$ samples suffices (exercise)