Lecture 10. (Sketch.)

Today we'll cover "online concept learning".

- This is a classical online setting where the vector we receive x_i, but then work only with the evaluations of some hypothesis/concept set $\mathcal{H}$:

$$(h_1(x_i), h_2(x_i), \dots, h_d(x_i)) \in \{0,1\}^d.$$

  (Sometimes it's {-1,+1}^d.)

- Sometimes these predictors are called "experts", and sometimes we assume there is a perfect expert, meaning $\exists h \in \mathcal{H}$ with $h(x_i) = y_i$ for all i.

[I discussed a few things at the start of lecture which I'll omit here, for instance a guarantee for Perceptron in the nonseparable case.]

1. Two baseline methods.

Consider the setting that $\exists \bar{h} \in \mathcal{H}$, $\bar{h}(x_i) = y_i$ for all i. This means that each iteration can permanently ignore any $h \in \mathcal{H}$ which makes any mistakes.

CONSISTENT.

1. Initialize $\mathcal{H}_0 = \mathcal{H}$.
2. For $i \in \{1, 2, \dots\}$:
   - 2.1 Receive x_i.
   - 2.2 Choose any $h_i \in \mathcal{H}_{i-1}$, output $\hat{y}_i := h_i(x_i)$.
   - 2.3 Receive y_i, construct $\mathcal{H}_i$:

$$\mathcal{H}_i := \begin{cases} \mathcal{H}_{i-1} & \hat{y}_i = y_i, \\ \mathcal{H}_{i-1} \setminus \{h_i\} & \hat{y}_i \neq y_i. \end{cases}$$

Theorem. If $\exists \bar{h}$, $\bar{h}(x_i) = y_i$, and $(\hat{y}_i)_{i \geq 1}$ are output by CONSISTENT,

$$\sum_{i \geq 1} \mathbb{1}[\hat{y}_i \neq y_i] \leq |\mathcal{H}| - 1.$$

Proof. By the update rule for $\mathcal{H}_i$,

$$|\mathcal{H}_i| = |\mathcal{H}_{i-1}| - \mathbb{1}[\hat{y}_i \neq y_i]$$

which rearranges to give $\mathbb{1}[\hat{y}_i \neq y_i] \leq |\mathcal{H}_{i-1}| - |\mathcal{H}_i|$. Applying $\sum_{i \leq t}$ to both sides,

$$\begin{aligned} \sum_{i \leq t} \mathbb{1}[\hat{y}_i \neq y_i] &= \sum_{i \leq t} (|\mathcal{H}_{i-1}| - |\mathcal{H}_i|) = |\mathcal{H}_0| - |\mathcal{H}_n| \\ &\leq |\mathcal{H}| - |\{\bar{h}\}| = |\mathcal{H}| - 1. \end{aligned}$$

Rather than removing only one hypothesis, we can remove all hypotheses that make a mistake on x_i.

In general, the subset of $\mathcal{H}$ which is consistent with all examples seen up through time i, meaning $((x_j, y_j))_{j \leq i}$, is called the **version space**: specifically,

$$\{h \in \mathcal{H} \ : \ \forall j \leq i \,.\, h(x_i) = y_i\}.$$

We can update CONSISTENT to remove more hypotheses, but without another change we can still guarantee only $|\mathcal{H}_i| \leq |\mathcal{H}_{i-1}| - \mathbb{1}[\hat{y}_i \neq y_i]$.

We can remove more hypotheses by choosing h_i more carefully.

HALVING.

1. Initialize *version space* $\mathcal{H}_0 = \mathcal{H}$.
2. For $i \in \{1, 2, \ldots\}$:
   2.1 Receive x_i.
   2.2 Choose majority label:
   $$\hat{y}_i := \arg\max_y \left| \left\{ h \in \mathcal{H} : h(x_i) = y \right\} \right| .$$
   2.3 Receive y_i, update version space $\mathcal{H}_i$:
   $$\mathcal{H}_i := \left\{ h \in \mathcal{H}_{i-1} : h(x_i) = y_i \right\} .$$

Theorem. If $\exists \bar{h}$, $\bar{h}(x_i) = y_i$, and $(\hat{y}_i)_{i \geq 1}$ are output by HALVING,

$$\sum_{i \geq 1} \mathbb{1}[\hat{y}_i \neq y_i] \leq \lg |\mathcal{H}|.$$

Proof. Since we chose the majority label, on mistake we know remove at least half the hypotheses:

$$|\mathcal{H}_i| \leq |\mathcal{H}_{i-1}| 2^{-\mathbb{1}[\hat{y}_i \neq y_i]},$$

which by induction gives

$$1 = |\{\bar{h}\}| \leq |\mathcal{H}_t| \leq |\mathcal{H}_0| \prod_{i=1}^{t} 2^{-\mathbb{1}[\hat{y}_i \neq y_i]} = |\mathcal{H}| 2^{-\sum_{i=1}^{t} \mathbb{1}[\hat{y}_i \neq y_i]},$$

which rearranges to give the desired bound.

Remark (comparison to Perceptron)**.** Suppose linear separability: $\exists \bar{u}, \gamma$ such that $\|\bar{u}\| = 1$ and $\inf_i \langle \bar{u}, x_i y_i \rangle \geq \gamma > 0$, and $\sup_i \|x_i y_i\| \leq 1$. Recall that perceptron makes at most $1/\gamma^2$ mistakes, and uses $\mathcal{O}(d)$ computation per round.

For CONSISTENT and HALVING, it suffices to choose $\mathcal{H}$ to be (the linear predictors corresponding to) a cover $\mathcal{W}$ of $\{w \in \mathbb{R}^d : \|w\| = 1\}$ at scale $\gamma/2$, since

$$\inf_{w \in \mathcal{W}} \inf_i \langle w, x_i y_i \rangle = \inf_{w \in \mathcal{W}} \inf_i \langle \bar{u}, x_i y_i \rangle + \langle w - \bar{u}, x_i y_i \rangle \geq \inf_{w \in \mathcal{W}} \inf_i \gamma - \|w - \bar{u}\| \cdot$$

Thus $|\mathcal{H}| = \mathcal{O}(1/\gamma)^d$, thus CONSISTENT and HALVING respectively make $\mathcal{O}(1/\gamma)^d$ and $\mathcal{O}(d \lg(1/\gamma))$ mistakes, but both have $\mathcal{O}(1/\gamma)^d$ computation. Also, they must guess γ, for instance with a double (halving) trick.

2. Winnow.

Now let's suppose the prefect predictor is a logical or of k elements of $\mathcal{H}$: $\bar{h}(x) = h_{i_1} \vee \cdots \vee h_{i_k}$.

- We can using HALVING and make only $k \ln(d)$ mistakes, but we still spend $\mathcal{O}(d^k)$ computation.
- Let's make an algorithm which maintains a candidate set of disjunction terms (initially everything). On iteration i:
  - If $y_i = -1$, remove any h which (mistakenly) output $+1$.
  - If $y_i = +1$, we can't be wrong (we started with everything, and never incorrectly remove), and we shouldn't remove anything.

  Even if $k = 1$, this can unfortunately take $d - 1$ not $\mathcal{O}(\ln(d))$ mistakes: suppose the target disjunction is just $h_n(x)$, but the sequence of inputs is $(\mathbf{e}_1, \mathbf{e}_2, \ldots)$, all with label -1.

Winnow will get roughly the same mistake bound as HALVING, while simultaneously being computationally efficient. One of its tricks is to maintain a linear predictor rather than a disjunction.

Remark. Methods that learn a hypothesis outside the target class are called **improper**.

WINNOW.

1. Initialize $w_j = 1$ for $j \in [d]$.
2. For $i \in \{1, 2, \ldots\}$:
   2.1 Receive x_i, predict $\hat{y}_i := \operatorname{sgn}(\sum_j w_j h_j(x_i) - d)$.
   2.2 Receive y_i; if $y_i \neq \hat{y}_i$, update w:
   $$w_j := \begin{cases} 2w_j \mathbb{1}[y_i = +1] & h_j(x_i) = +1, \\ w_j & h_j(x_i) = 0. \end{cases}$$
   (Note, $h_j(x) \in \{0, 1\}$.)

Theorem. If $(\hat{y}_i)_{i \geq 1}$ are computed by WINNOW,

$$\sum_{i \geq 1} \mathbb{1}[y_i \neq \hat{y}_i] \leq 1 + 2k\lceil \lg d \rceil.$$

Proof.

- Each mistake when $y = +1$ doubles w_j for some true disjunction term h_j, and they are never decreased when $y = -1$; together, mistake on $y = +1$ can happen at most $r\lceil \lg(d) \rceil$ times.
- Each mistake when $y = -1$ has $\sum_{j:h_j(x)=1} = \sum_j w_j h_j(x) > d$, thus $\|w\|_1$ decreases by at least d. Initially, $\|w\|_1 = d$, and each (of at most $P \leq r\lceil \lg(d) \rceil$) mistakes on positive add at most d, so the number of mistakes on negative, N, can not decrease $\|w\|_1$ below 0, and so $N \leq 1 + P$.

Remark. This simple proof for disjunctions is by Avrim Blum. In class I discussed learning linear predictors (not just disjunctions), but the math is messy.