Machine Learning Theory (CSC 482A/581B) - Lecture 2

Nishant Mehta

1 Halving algorithm

When the concept class is finite, there is a surprisingly simple algorithm that obtains a mistake bound of $\log_2 |\mathcal{C}|$. This algorithm is called the Halving algorithm, and it uses two key ideas.

The first idea is that of a *version space*. The version space is the set of hypotheses that are consistent with the data observed thus far. Thus, at the start of round t, the version space $\mathcal{V}_t$ is the subset of hypotheses from $\mathcal{C}$ which are consistent with $(x_1, y_1), \ldots, (x_{t-1}, y_{t-1})$.

The second idea is to predict according to a majority vote. For a set of hypotheses $\mathcal{F}$, define the majority vote based on $\mathcal{F}$ as

$$\textsc{mv}_{\mathcal{F}}(x) = \begin{cases} 1 & \text{if } |\{f \in \mathcal{F} : f(x) = 1\}| \geq |\mathcal{F}|/2; \\ 0 & \text{otherwise.} \end{cases}$$

The Halving algorithm simply predicts according to the majority vote with respect to the version space in every round.

Algorithm 1: Halving Algorithm

$\mathcal{V}_1 \leftarrow \mathcal{C}$
for $t = 1 \to T$ **do**
  Observe x_t
  $f_t \leftarrow \textsc{mv}_{\mathcal{V}_t}$ (and predict $\hat{y}_t = \textsc{mv}_{\mathcal{V}_t}(x_t)$)
  Observe true label $y_t = c(x_t)$
  Set $V_{t+1} \leftarrow \{f \in V_t : f(x_t) = y_t\}$
end

How many mistakes does this algorithm make? Because it predicts according to the majority vote, wherever the algorithm makes a mistake it is guaranteed that at least half the hypotheses in the version space were wrong; thus, the version space is halved on each mistake. Formally, if the algorithm makes a mistake in round t, it holds that $|\mathcal{V}_{t+1}| \leq |\mathcal{V}_t|/2$. We initially have $\mathcal{V}_1 = \mathcal{C}$, and so if we have made M_t mistakes at the beginning of round t, it follows that $|V_t| \leq |\mathcal{C}|/2^{M_t}$. Since there exists a perfect hypothesis $c \in \mathcal{C}$, the algorithm can make at most $\log_2 |\mathcal{C}|$ mistakes.

We have just shown that any finite concept class is learnable in the mistake bound model using the Halving algorithm.

Theorem 1. *The Halving algorithm learns any finite concept class $\mathcal{C}$ in the mistake bound model and makes at most $\log_2 |\mathcal{C}|$ mistakes.*

Unfortunately, the runtime of the Halving algorithm is linear in $|\mathcal{C}|$, which can be exorbitant. Why is this bad? In many situations, the size of the concept class $|\mathcal{C}|$ can be exponential in the dimension of the data, in which case the runtime of the Halving algorithm is *exponential* in d (!). For instance, the class of monotone conjunctions has cardinality 2^d. In other cases, such as the case of linear separators, the concept class can even be infinite.

2 Learning linear separators in the mistake bound model

We next consider the problem of linear classification in the realizable case. Specifically, we will look at the subclass of linear classifiers known as *homogenous linear separators*.

Let $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{-1, +1\}$. The concept class of homogenous linear separators is defined as $\mathcal{C} = \{f_w : w \in \mathbb{R}^d\}$, for hypotheses $f_w(x) = \text{sgn}(\langle w, x\rangle)$. Here, sgn is the sign function, defined[1] as the map

$$\text{sgn}(z) = \begin{cases} +1 & \text{if } z \geq 0; \\ -1 & \text{if } z < 0. \end{cases}$$

The reason these linear separators are called homogenous is because the concept class lacks a bias term; consequently, the linear separator corresponding to a vector w is the hyperplane normal to w that passes through the origin, i.e. $\{x \in \mathbb{R}^d : \langle w, x\rangle = 0\}$. If we also allowed for a bias term $b \in \mathbb{R}$, the class would be upgraded to the set of non-homogenous linear separators (which do not necessarily pass through the origin); this class also is commonly referred to as *halfspaces*, and each hypothesis is of the form $x \mapsto \text{sgn}(\langle w, x\rangle + b)$.

As before, we make the realizability assumption with respect to concept class $\mathcal{C}$. The sequence of examples is thus linearly separable, meaning that there exists a vector $w^* \in \mathbb{R}^d$ for which

$$y_t = \text{sgn}(\langle w^*, x_t) \qquad \text{for all } t \in [T].$$

Since the output of any classifier f_w is invariant to scaling of w, without loss of generality we assume that w^* has unit ℓ_2 norm.

At this stage, we lack the tools to provide a mistake bound for learning the class of homogenous linear separators. Were the concept class finite, we could use our mistake bound for the Halving algorithm, but alas, the concept class is not even countable. However, with one additional assumption and a simple discretization argument, we will be able to apply our result for the Halving algorithm.

Assuming separability with margin. We further assume that the positive and negative examples are linearly separable by some positive margin. To make this precise, for any $w \in \mathbb{R}^d$, define the *margin with respect* $w \in \mathbb{R}^d$ (and the sequence of examples) to be

$$\gamma_w = \min_{t \in [T]} \frac{y_t \langle w, x_t\rangle}{\|w\|_2}.$$

If f_w correctly classifies (x_t, y_t), it is easy to see that γ_w is equal to the Euclidean distance from x_t to the hyperplane $\{x : \langle w, x\rangle = 0\}$. *(In class I explained this statement using a picture and some basic trigonometry)*

[1] The sign function usually maps zero to zero, but we map zero to one so that our classifiers take values in $\{-1, +1\}$.

The *margin* γ is then

$$\gamma = \gamma_{w^*} = \min_{t \in [T]} y_t \langle w^*, x_t \rangle.$$

We now formalize our assumption of separability with margin:

Assumption 1. *There exists a unit vector $w^* \in \mathbb{R}^d$ for which*

$$\gamma = \min_{t \in [T]} y_t \langle w^*, x_t \rangle > 0.$$

With the margin assumption in place, we now form a finite, discretized approximation $\mathcal{C}_\gamma$ of the infinite class $\mathcal{C}$; the key property that our discretized approximation will satisfy is that it still contains a hypothesis that linearly separates the data, and so running the Halving algorithm on C_γ will yield a mistake bound of $\log |\mathcal{C}_\gamma|$.

To define $\mathcal{C}_\gamma$, we first introduce a concept known as a cover and then introduce the related concept of a covering number.

Definition 1. Let $(\mathcal{Z}, \|\cdot\|)$ be a metric space. We say that A is a proper[2] ε-cover for $\mathcal{Z}$ if $A \subset \mathcal{Z}$ and, for every $x \in \mathcal{Z}$, there exists some $x' \in A$ such that $\|x - x'\| \leq \varepsilon$.

Definition 2. For a metric space $(\mathcal{Z}, \|\cdot\|)$, define the proper ε-covering number $\mathcal{N}(\mathcal{Z}, \|\cdot\|, \varepsilon)$ as the minimum cardinality of a proper ε-cover for $\mathcal{Z}$.

Let $S = \{w \in \mathbb{R}^d : \|w\|_2 = 1\}$, i.e. the unit sphere which is the boundary of the d-dimensional Eucidean unit ball. Then it is true that[3]

$$\mathcal{N}(S, \|\cdot\|_2, \varepsilon) \leq \left(\frac{4}{\varepsilon} + 1\right)^d;$$

we will not prove this fact here, but short proofs exist.

Now, define $R := \max_{t \in [T]} \|x_t\|_2$, take S_γ to be a proper ε-cover for S for $\varepsilon = \frac{\gamma}{2R}$, and define

$$\mathcal{C}_\gamma = \{f_w : w \in S_\gamma\}.$$

Then, by the definition of S_γ, there exists $\tilde{w} \in S_\gamma$ for which $\|w^* - \tilde{w}\|_2 \leq \frac{\gamma}{2R}$. For this $\tilde{w}$, for any $t \in [T]$, we have

$$\begin{aligned}
y_t \langle \tilde{w}, x_t \rangle &= y_t \langle w^*, x_t \rangle + y_t \langle \tilde{w} - w^*, x_t \rangle \\
&\geq \gamma + y_t \langle \tilde{w} - w^*, x_t \rangle && \text{(separability with margin)} \\
&\geq \gamma - \|\tilde{w} - w^*\|_2 \cdot \|x_t\|_2 && \text{(Cauchy-Schwarz inequality)} \\
&\geq \gamma - \frac{\gamma}{2R} \cdot R \\
&= \frac{\gamma}{2}.
\end{aligned}$$

Thus, we have $\gamma_{\tilde{w}} \geq \gamma/2 > 0$, and so $f_{\tilde{w}} \in C_\gamma$ perfectly separates the data, as promised.

The following corollary is now immediate.

[2]The modifier *proper* means that we require that $A \subset \mathcal{Z}$. In general, covers need not satisfy this requirement.

[3]The same claim in fact holds for the unit ball itself (and for any norm), and we likely are overpaying quite a bit since the intrinsic dimension of the unit sphere is $d - 1$ rather than d.

Corollary 1. *Let $(x_1, y_1), \ldots, (x_T, y_T)$ be a sequence of examples that is linearly separable with margin $\gamma > 0$ in the sense of Assumption 1. Then on this sequence the Halving algorithm based on class $\mathcal{C}_\gamma$ makes at most $d \log_2 \left(\left\lceil \frac{8R}{\gamma} \right\rceil + 1 \right)$ mistakes, where $R = \max_{t \in [T]} \|x_t\|_2$.*

We thus have a finite mistake bound over an infinite class, at least under the assumption of realizability and boundedness of the data. However, the algorithm (Halving) is inefficient, maintaining C_γ is inefficient (and not even possible if γ is unknown!), and the mistake bound is far from optimal. Next, we will see that there is a simple, efficient algorithm which obtains a mistake bound that is independent of the dimension d, and this algorithm does not need to know γ.

3 Perceptron

We now consider an efficient algorithm for learning linear separators with a finite number of mistakes, under the assumption that the data is separable with margin γ. This algorithm is the *Perceptron* algorithm.

Algorithm 2: PERCEPTRON

$w_0 \leftarrow \mathbf{0}$
$m = 0$
for $t = 1 \to T$ **do**
  Observe x_t
  Predict $\hat{y}_t = \text{sgn}(\langle w_m, x_t \rangle)$
  Observe true label $y_t = c(x_t)$
  if $\hat{y}_t \neq y_t$ **then**
    $w_{m+1} \leftarrow w_m + y_t x_t$
    $m \leftarrow m + 1$
  end
end

Theorem 2. *Let $(x_1, y_1), \ldots, (x_T, y_T)$ be a sequence of examples that is linearly separable with margin $\gamma > 0$. Then on this sequence the Perceptron algorithm makes at most $\frac{R^2}{\gamma^2}$ mistakes.*

Remarkably, this mistake bound is *independent* of the dimension d (note, however, that the dependence on γ is much worse as compared our argument based on the Halving algorithm).

Before showing the proof, let's build some intuition for Perceptron's update rule in the case of a mistake. Suppose that Perceptron makes its m^{th} mistake on a positively labeled example $(x, 1)$, resulting in the update $w_{m+1} = w_m + x$. Then $\langle w_{m+1}, x \rangle = \langle w_m, x \rangle + \|x\|^2 > \langle w_m, x \rangle$, and so the new hypothesis is closer to classifying x as positive. This same intuition works for mistakes on negative examples as well.

Proof. The proof is based on two claims.

The first claim is that the norm of w_m is never too big:

$$\|w_m\| \leq R\sqrt{m}. \tag{1}$$

Let's prove this claim. For $j \geq 1$, let $(\tilde{x}_j, \tilde{y}_j)$ denote the j^{th} example on which Perceptron made a mistake. It therefore holds that

$$w_{m+1} = w_m + \tilde{y}_m \tilde{x}_m.$$

Then

$$\begin{aligned}
\|w_{m+1}\|^2 &= \|w_m + \tilde{y}_m \tilde{x}_m\|^2 \\
&= \|w_m\|^2 + \|\tilde{x}_m\|^2 + 2\tilde{y}_m \langle w_m, \tilde{x}_m \rangle \\
&\leq \|w_m\|^2 + R^2,
\end{aligned}$$

where the inequality follows because $\tilde{y}_m \langle w_m, \tilde{x}_m \rangle \leq 0$ since w_m made a mistake on $(\tilde{x}_m, \tilde{y}_m)$. Repeating this argument all the way back to $w_0 = \mathbf{0}$ yields the claim.

The second claim is that the inner product $\langle w_m, w^* \rangle$ grows quickly with m:

$$\langle w_m, w^* \rangle \geq \gamma \cdot m. \tag{2}$$

To see this, observe that

$$\begin{aligned}
\langle w^*, w_{m+1} \rangle &= \langle w^*, w_m + \tilde{y}_m \tilde{x}_m \rangle \\
&= \langle w^*, w_m \rangle + \tilde{y}_m \langle w^*, \tilde{x}_m \rangle \\
&\geq \langle w^*, w_m \rangle + \gamma.
\end{aligned}$$

Applying this argument recursively yields $\langle w^*, w_{m+1} \rangle \geq \gamma \cdot (m+1)$, proving the claim.

Now, the inner product $\langle w_m, w^* \rangle$ grows *at most* linearly in $\|w_m\|$ (from the Cauchy-Schwarz inequality), which itself grows no faster than the root of m (from the first claim). Consequently, $\langle w_m, w^* \rangle = O(\sqrt{m})$. On the other hand, this inner product also grows *at least* linearly in m (from the second claim), and so it must be the case that m is bounded as otherwise we would arrive at a contradiction. Indeed, applying Cauchy-Schwarz to (2) and using (1) yields

$$\gamma \cdot m \overset{(2)}{\leq} \langle w^*, w_m \rangle = \|w^*\| \cdot \|w_m\| \leq \|w_m\| \overset{(1)}{\leq} R\sqrt{m},$$

and so $m \leq \frac{R^2}{\gamma^2}$. □