CS 395T Computational Learning Theory

Lecture 4: September 21, 2005 — Lecturer: Adam Klivans

Scribe: Doo Soon Kim

4.1 Halfspace

We are given a domain(e.g. $\Re^n$, $\{0,1\}^n$,etc) and a linear threshhold function that cuts the domain into two sides. Positive points on one side of the linear function corresponds to a halfspace. More formally, a halfspace is defined by the linear threshold function h such that
h(x) = $sign$(l(x)-θ) where a linear function l = $w_1x_1 + ... + w_nx_n$ and $\theta \in \Re$.

$$sign(a) = \begin{cases} +1 & \text{if } a > 0 \\ -1 & \text{otherwise} \end{cases}$$

$\Re^n$, $\{0,1\}^n$, and S^{n-1}(the set of all $\mathbf{x}$ s.t. $|\mathbf{x}| = 1$) will be the 3 most common domains that we will look at. As we showed in a previous class, the VC-dimension of a halfspace in $\Re^d$ is d+1.

4.2 Learning Halfspace

4.2.1 PAC algorithm for learning halfspace

Here is the general strategy for PAC algorithm for learning halfspaces. It guarantees that if we have O($\frac{n}{\epsilon}$) examples, we have the ϵ-accurate hypothesis.

1. Draw a bunch of random example (e_1,l_1),...,(e_n,l_n)
2. Apply any polynomial time algorithm for solving a linear program to find weights w_1,w_2,...,w_n, θ

$$\begin{aligned} w_1e_{11} + w_2e_{12} + \ldots + w_ne_{1n} &\geq (or \leq) \quad \theta \\ w_1e_{21} + w_2e_{22} + \ldots + w_ne_{2n} &\geq (or \leq) \quad \theta \\ &\vdots \\ w_1e_{n1} + w_2e_{n2} + \ldots + w_ne_{nn} &\geq (or \leq) \quad \theta \end{aligned}$$

3. By the VC-theorem, we could guarantee to have an ϵ-accurate hypothesis if the sample size $m >> \frac{VCD(h)}{\epsilon}$ (VCD(h) = n+1)

4.2.2 Perceptron Algorithm

In 1958, Frank Rosenblatt, who was a neuroscientist, invented the *perceptron algorithm* which is the learning algorithm for a linear threshold function(halfspace). Under some circumstances, it is a polynomial time algorithm even though it is generally not.

Initially set the all weight as 0
On input $\mathbf{x}$, if $\mathbf{w}\cdot\mathbf{x} \geq 0$, output1. Otherwise, output -1.
If it makes a mistake, update as the following:
if x is a positive example **then**
  w' = w + x
else if x is a negative example **then**
  w' = w - x
end if

Blak and Novikoff proved that the mistake bound of the perceptron algorithm depends on only the geometric margin, not on the dimension.

4.2.3 Perceptron Convergence Theorem

Let's assume that the domain is the n-dimensional unit length sphere, S^{n-1}, and there is a halfspace that classifies the data, $\mathbf{w}^*$ s.t. $|\mathbf{w}^*| = 1$. The geometric margin, σ, is defined as $min(\frac{|\mathbf{w}^*\cdot\mathbf{x}|}{|\mathbf{x}|})$. In other words,$\mathbf{w}\cdot\mathbf{x} \geq \sigma$ for all $\mathbf{x}$. It just says that all $\mathbf{x}$ are distant from $\mathbf{w}^*$ by the σ. Without loss of generality, we could assume that θ is 0. Otherwise, we could introduce an additional input $\mathbf{x_{n+1}} = 1$ and set $\mathbf{w_{n+1}} = -\theta$.

Lemma 1 *On every mistake,* $\mathbf{w}\cdot\mathbf{w}^*$ *increases by at least* σ.

Proof: Let $\mathbf{w}'$ to be the adjusted weight after an integration, and $\mathbf{w}$ be the weight before the interation. So, $\mathbf{w}' = \mathbf{w} + y_i\mathbf{x_i}$.
$\mathbf{w}'\cdot\mathbf{w}^* = (\mathbf{w} + y_i\mathbf{x_i})\cdot\mathbf{w}^* = \mathbf{w}\cdot\mathbf{w}^* + y_i\mathbf{x}\cdot\mathbf{w}^*$. By the definition of σ, $y_i\mathbf{x}\cdot\mathbf{w}^* \geq \sigma$. So, $\mathbf{w}'\cdot\mathbf{w}^* \geq \mathbf{w}\cdot\mathbf{w}^* + \sigma$ ■

Lemma 2 $|\mathbf{w}|^2$ *increases by at most 1 after every mistake.*

Proof: $\mathbf{w}'^2$=$(\mathbf{w}$+$y_i\mathbf{x})^2 = |\mathbf{w}|^2 + 2y_i\mathbf{w}\cdot\mathbf{x} + |\mathbf{x}|^2$. Since, $y_i\mathbf{w}\cdot\mathbf{x} < 0$, $|\mathbf{x}|^2 = 1$. the lemma holds. ■

Theorem 1 (Perceptron Convergence Theorem) *The mistake bound of the perceptron algorithm is less than or equal to* $\frac{1}{\sigma^2}$

Proof: After M mistake, $\mathbf{w}\cdot\mathbf{w}^* \geq \sigma$ M (by lemma1) , and $|\mathbf{w}| \leq \sqrt{M}$ (by lemma2). Since $|\mathbf{w}^*| = 1$, σ M $\leq \sqrt{M}$. Then, M$\leq \frac{1}{\sigma^2}$ ■

The convergence theorem assumed that the data are linearly separable. So, it can't be applied if the data are unseparable. Freund and Schapire extended the theorem into the more general case where data are not linearly separable.

Let's first define a terminology. Given m training example $< (x_1, y_1), .., (x_m, y_m) >$ such that $|x_i| \leq R$, $|w^*|$=1, $\sigma > 0$, the deviation of example x_i is defined as $d_i = max(0, \gamma - y_i(u \cdot x_i))$ (Notice that d_i is 0 if the data are linearly separable).

Theorem 2 (Perceptron Convergence Theorem2) *Let $< (x_1, y_1), .., (x_m, y_m) >$ be a sequence of labeled examples with $|x_i| \leq \Re$. Let w'^* be any vector with $|\mathbf{w}^*|$=1 and let γ¿0. Define the deviation of each example as $d_i = max(0, \gamma - y_i(\mathbf{w}^* \cdot \mathbf{x_i}))$ and define D=$\sqrt{\sum_{i=1}^m d_i^2}$. Then the number of mistakes of the online perceptron algorithm on this sequence is bounded by $\frac{R+D}{\gamma^2}$*

Proof: Since we proved the case D = 0 in the above, we now assume that D > 0. Now, we extend the instance space R^n to $R^n + m$ by adding m new dimensions. Let x' denote the extension of the x_i. We set the first n coordinate of x'_i equal to x_i, and (n+i)'th coordinate to $\triangle$, rest to zero. Next we extend $\mathbf{w}^*$ to $\mathbf{w}'^*$. We set the first n coordinates equal to u/Z, (n+i)'th coordinate to $(y_i d_i)/(\mathrm{Z}\triangle)$ (Z = $\sqrt{1 + \frac{D^2}{\triangle^2}}$).

Now we will show that w'^* classifies x'_i with sufficiently large margin. Consider the value of $y_i(w'^* \cdot x'_i)$.

$$\begin{aligned} y_i(\mathbf{w}'^* \cdot \mathbf{x'_i}) &= y_i(\frac{\mathbf{w}^* \cdot \mathbf{x_i}}{Z} + \triangle\frac{y_i d_i}{Z\triangle}) \\ &= \frac{y_i(\mathbf{w}^* \cdot \mathbf{x_i})}{Z} + \frac{d_i}{Z} \\ &\leq \frac{y_i(\mathbf{w}^* \cdot \mathbf{x_i})}{Z} + \frac{\gamma - y_i(w^* \cdot x_i}{Z} \\ &= \frac{\gamma}{Z} \end{aligned}$$

Thus the extended vector w'^* achieves a margin of $\frac{\gamma}{\sqrt{1+D^2/\triangle^2}}$ on the extended examples. Using the perceptron convergence theorem1, we could get the mistake bound for the extended vectors. Since $|x'_i| \leq R^2 + \triangle^2$, the mistake bound is at most
$\frac{(R^2+\triangle^2)(1+D^2/\triangle^2)}{\gamma^2}$
Setting $\triangle = \sqrt{RD}$ minimizes the bound and yield $\frac{R+D}{\gamma^2}$
To finish the proof, we show that the predictions of the perceptron algorithm in the extended space are equal to the prediction of the perceptron in the original space. The claim follows by induction over 1≤i≤m of the following three claims :

1. The first n cordinates of w'_i are equal to those of w_i
2. The (n+i)'th coordinate of w'_i is equal to zero.
3. sign($\mathbf{w'_i} \cdot \mathbf{x'_i}$)=sign($\mathbf{w_i} \cdot \mathbf{x_i}$)

■

References

[1] F. Rosenblatt: "The Perceptron: A probabilistic model for information storage and organization in the brain". *Psychological Review*, 65:386-407, 1958. Article reprinted in *Neurocomputing* (MIT Press, 1988)

[2] Freund and Schapire, "Large margin classification using the perceptron algorithm"