Introduction to Machine Learning **Fall Semester**

Recitation 4

Lecturer: Regev Schweiger *Scribe: Yishay Mansour*

4.1 Perceptron - Review

Assume that all examples $S = \mathbf{x}_1, \ldots$ are normalized to have $\|\mathbf{x}\| = 1$, and there exists a unit vector $\mathbf{w}^*$ ($\|\mathbf{w}^*\| = 1$) such that the true label of $\mathbf{x}$ is

$$c^*(\mathbf{x}) = sign(\mathbf{x} \cdot \mathbf{w}^*)$$

We define the margin as:

$$\gamma = \min_{\mathbf{x} \in \mathcal{S}} |\mathbf{w}^* \cdot \mathbf{x}|.$$

Recall the perceptron algorithm:

Perceptron:

1. Initialize $\mathbf{w}_1 = \mathbf{0}$.
2. Predict positive if $\frac{\mathbf{w}_t \cdot \mathbf{x}}{||\mathbf{w}_t||} \geq 0$, predict negative if $\frac{\mathbf{w}_t \cdot \mathbf{x}}{||\mathbf{w}_t||} \leq 0$.
3. On a mistake: $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + c^*(\mathbf{x})\mathbf{x}$; $t \leftarrow t + 1$.

We saw that the perceptron algorithm makes at most $1/\gamma^2$ mistakes on any sequence of examples that is linearly-separable by margin γ (i.e., any sequence for which there exists a unit-length vector $\mathbf{w}^*$ such that all examples $\mathbf{x}$ satisfy $c^*(\mathbf{x})(\mathbf{w}^* \cdot \mathbf{x})/||\mathbf{x}|| \geq \gamma$, where $c^*(\mathbf{x}) \in \{-1, 1\}$ is the label of $\mathbf{x}$).

4.2 Margin Perceptron

There can be many possible separating hyperplanes, however - some are much better than others. Suppose we are handed a set of examples $\mathcal{S}$ and we want to actually find a *large-margin* separator for them. One approach is to directly solve for the maximum-margin separator using convex programming (which is what is done in the SVM algorithm). However, if we only need to *approximately* maximize the margin, then another approach is to use the perceptron algorithm. In particular, suppose we cycle through the data using the perceptron algorithm, updating not only on mistakes, but also on examples $\mathbf{x}$ that our current hypothesis gets correct by margin less than $\frac{\gamma}{2}$. Assuming our data is separable by margin γ,

then we can show that this is guaranteed to halt in a number of rounds that is polynomial in $\frac{1}{\gamma}$. (In fact, we can replace $\frac{\gamma}{2}$ with $(1-\varepsilon)\gamma$ and have bounds that are polynomial in $\frac{1}{\varepsilon\gamma}$.)

The Margin Perceptron Algorithm(γ)**:**

1. Initialize $\mathbf{w}_1 = \mathbf{0}$.
2. Predict positive if $\frac{\mathbf{w}_t \cdot \mathbf{x}}{||\mathbf{w}_t||} \geq \frac{\gamma}{2}$, predict negative if $\frac{\mathbf{w}_t \cdot \mathbf{x}}{||\mathbf{w}_t||} \leq -\frac{\gamma}{2}$, and consider an example to be a margin mistake when $\frac{\mathbf{w}_t \cdot x}{||\mathbf{w}_t||} \in (-\frac{\gamma}{2}, \frac{\gamma}{2})$.
3. On a mistake (incorrect prediction or margin mistake), update as in the standard Perceptron algorithm: $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + c^*(\mathbf{x})\mathbf{x}$; $t \leftarrow t+1$.

Theorem 4.1 *Let $\mathcal{S}$ be a sequence of labeled examples consistent with a linear threshold function $\mathbf{w}^* \cdot \mathbf{x} \geq 0$, where $\mathbf{w}^*$ is a unit-length vector, and let*

$$\gamma = \min_{\mathbf{x} \in \mathcal{S}} |\mathbf{w}^* \cdot \mathbf{x}|.$$

Then the number of mistakes (including margin mistakes) made by Margin Perceptron(γ) on $\mathcal{S}$ is at most $\frac{16}{\gamma^2}$.

Proof: The argument for this new algorithm follows the same lines as the argument for the original perceptron algorithm. As before, we can show that each update increases $\mathbf{w}_t \cdot \mathbf{w}^*$ by at least γ:

$$\mathbf{w}_{t+1} \cdot \mathbf{w}^* = (\mathbf{w}_t + c^*(\mathbf{x})\mathbf{x}) \cdot \mathbf{w}^* = \mathbf{w}_t \cdot \mathbf{w}^* + c^*(\mathbf{x})\mathbf{x} \cdot \mathbf{w}^* \geq \mathbf{w}_t \cdot \mathbf{w}^* + \gamma$$

This is because γ is chosen to be smaller every $|\mathbf{x} \cdot \mathbf{w}^*|$, and $c^*(x)$ fixes the sign. Note that if $\mathbf{x}$ is a margin error, $|\frac{\mathbf{w}_t \cdot \mathbf{x}}{||\mathbf{w}_t||}| \leq \frac{\gamma}{2}$, but still $|\mathbf{w}^* \cdot \mathbf{x}| \geq \gamma$.

What is now a little more complicated is to bound the increase in $||\mathbf{w}_t||$. For the original algorithm, we had: $||\mathbf{w}_{t+1}||^2 \leq ||\mathbf{w}_t||^2 + 1$. Using a similar method to the one we use below, we could have actually shown $||\mathbf{w}_{t+1}|| \leq ||\mathbf{w}_t|| + \frac{1}{2||\mathbf{w}_t||}$. However, for the Margin Perceptron algorithm, we can show instead:

$$||\mathbf{w}_{t+1}|| \leq ||\mathbf{w}_t|| + \frac{1}{2||\mathbf{w}_t||} + \frac{\gamma}{2}. \tag{4.1}$$

To see this, note that:

$$||\mathbf{w}_{t+1}||^2 = ||\mathbf{w}_t||^2 + 2c^*(x)\mathbf{w}_t \cdot \mathbf{x} + ||\mathbf{x}||^2 = ||\mathbf{w}_t||^2 \left(1 + \frac{2c^*(x)}{||\mathbf{w}_t||}\frac{\mathbf{w}_t \cdot \mathbf{x}}{||\mathbf{w}_t||} + \frac{1}{||\mathbf{w}_t||^2}\right)$$

Using the inequality $\sqrt{1+\alpha} \leq 1 + \frac{\alpha}{2}$ together with the fact $\frac{c^*(x)\mathbf{w}_t \cdot \mathbf{x}}{||\mathbf{w}_t||} \leq \frac{\gamma}{2}$ (since $\mathbf{w}_t$ made a mistake on $\mathbf{x}$) we get the desired upper bound on $||\mathbf{w}_{t+1}||$, namely:

$$||\mathbf{w}_{t+1}|| = ||\mathbf{w}_t||\sqrt{1 + \frac{2c^*(\mathbf{x})}{||\mathbf{w}_t||}\frac{\mathbf{w}_t\mathbf{x}}{||\mathbf{w}_t||} + \frac{1}{||\mathbf{w}_t||^2}} \leq ||\mathbf{w}_t||\sqrt{1 + \frac{2}{||\mathbf{w}_t||}\frac{\gamma}{2} + \frac{1}{||\mathbf{w}_t||^2}}$$

$$\leq ||\mathbf{w}_t||(1 + \frac{\frac{\gamma}{||\mathbf{w}_t||} + \frac{1}{||\mathbf{w}_t||^2}}{2}) \Rightarrow ||\mathbf{w}_{t+1}|| \leq ||\mathbf{w}_t|| + \frac{1}{2||\mathbf{w}_t||} + \frac{\gamma}{2}$$

Notice we selected $(\frac{\gamma}{||\mathbf{w}_t||} + \frac{1}{||\mathbf{w}_t||^2})$ as α. Note that (4.1) implies that if $||\mathbf{w}_t|| \geq \frac{2}{\gamma}$ then $||\mathbf{w}_{t+1}|| \leq ||\mathbf{w}_t|| + \frac{3}{4}\gamma$. Let T be the smallest index (if such exists) for which, for all $t \geq T$, $||\mathbf{w}_t|| \geq \frac{2}{\gamma}$. Therefore, $||\mathbf{w}_{T-1}|| < \frac{2}{\gamma}$ and $||\mathbf{w}_T|| \geq \frac{2}{\gamma}$. We can see that:

$$||\mathbf{w}_T|| = ||\mathbf{w}_{T-1} + c^*(\mathbf{x}_{T-1})\mathbf{x}_{T-1}|| \leq ||\mathbf{w}_{T-1}|| + ||c^*(\mathbf{x}_{T-1})\mathbf{x}_{T-1}|| \leq \frac{2}{\gamma} + 1$$

Given that, it is easy to see that after M updates we have:

$$||\mathbf{w}_{M+1}|| \leq ||\mathbf{w}_T|| + \frac{3}{4}(M - T + 1)\gamma \leq 1 + \frac{2}{\gamma} + \frac{3}{4}M\gamma.$$

If there is no such T, then it means that $||\mathbf{w}_{M+1}|| < \frac{2}{\gamma}$, so the above inequality holds.

As before, $\gamma M \leq ||\mathbf{w}_{M+1}||$. Solving $M\gamma \leq 1 + \frac{2}{\gamma} + \frac{3}{4}M\gamma$, remembering that $\gamma \leq 1$ (both the samples and $\mathbf{w}^*$ are unit size), we get $M \leq \frac{16}{\gamma^2}$, as desired. □

Comment: Here we saw how the Perceptron algorithm can be modified so that its result approaches the best possible margin. We accomplished this using $\frac{\gamma}{2}$, but we might as well have chosen $(1 - \varepsilon)\gamma$.