Introduction to Machine Learning (CS419M)

Lecture 10:

- Perceptron
- Convergence Proof
- Hinge Loss

Mar 4, 2020

Course Roadmap

Perceptron Learners	March 4	
Neural Networks (I)	March 6	Project abstracts due
Neural Networks (II)	March 11	
Neural Networks (III)	March 13	
SVMs and Kernel methods	March 18	Assignment 2 released
SVMs and Kernel methods	March 20	
Clustering + EM	March 25	
Clustering + EM	March 27	
Nearest neighbour classifiers	April 1	
Quiz 2	**April 3**	
Generalization bounds	April 8	Assignment 2 due
Dimensionality Reduction	April 15	
Ensemble learning	April 17	Project preliminary report due
Ensemble learning	April 22	
Buffer	April 24	

Perceptron Algorithm

Goal: To learn a weight vector $\mathbf{w}$ such that sign($\mathbf{w}^T\mathbf{x}$) is correct for all $\mathbf{x} \in \mathscr{D}$.

$$\text{sign}(\mathbf{w}^T\mathbf{x}) = \begin{cases} +1 \text{ if } \mathbf{w}^T\mathbf{x} \geq 0 \\ -1 \text{ otherwise} \end{cases}$$

Algorithm:

- Start with zero-weights vector, $\mathbf{w} \leftarrow \bar{0}$
- For a fixed number of iterations
    - For a training instance, ($\mathbf{x}$, y) $\in \mathscr{D}$
        - if ($y\mathbf{w}^T\mathbf{x} \leq 0$)
            - $\mathbf{w} \leftarrow \mathbf{w} + y\mathbf{x}$

The perceptron weight update rule makes the classifier more correct on a misclassified example:

$$\begin{aligned} y\mathbf{w}_{\text{new}}^T\mathbf{x} &= y(\mathbf{w}_{\text{old}} + y\mathbf{x})^T\mathbf{x} \\ &= y\mathbf{w}_{\text{old}}^T\mathbf{x} + y^2||\mathbf{x}||_2^2 \\ &> y\mathbf{w}_{\text{old}}^T\mathbf{x} \end{aligned}$$

Mistake Bounds for the Perceptron Algorithm

Consider the case when data is linearly separable i.e. there exists a weight vector $\mathbf{u}$ s.t. $y = \text{sign}(\mathbf{u}^T\mathbf{x}) \ \forall \mathbf{x}, y \in \mathscr{D}$. Without loss of generality, we assume that $\mathbf{u}$ is a unit-length vector. We also assume that data is scaled to lie in a Euclidean ball of radius 1, i.e., $||\mathbf{x}|| \leq 1 \ \ \forall \mathbf{x} \in \mathscr{D}$.

We define the *margin of separation*, $\gamma = \min_{\mathbf{x} \in \mathscr{D}} |\mathbf{u}^T\mathbf{x}|$

Theorem: If there exists a unit vector $\mathbf{u}$ such that $y\mathbf{u}^T\mathbf{x} \geq \gamma$ for all $\mathbf{x}$, then the number of weight updates (or number of mistakes) made by the perceptron algorithm is at most $\dfrac{1}{\gamma^2}$.

Proof of the mistake bound

We will track two quantities: $\mathbf{w}^T\mathbf{u}$ and $||\mathbf{w}||^2$

Claim 1: $\mathbf{w}_{t+1}^T\mathbf{u} \geq \mathbf{w}_t^T\mathbf{u} + \gamma$

For a positive example that is misclassified,
$\mathbf{w}_{t+1}^T\mathbf{u} = (\mathbf{w}_t + \mathbf{x})^T\mathbf{u} \geq \mathbf{w}_t^T\mathbf{u} + \gamma$ (by definition of γ)
(Similar argument holds for a negative example)

Claim 2: $||\mathbf{w}_{t+1}||^2 \leq ||\mathbf{w}_t||^2 + 1$

For a positive example that is misclassified,

$$||\mathbf{w}_{t+1}||_2 = (\mathbf{w}_t + \mathbf{x})^T(\mathbf{w}_t + \mathbf{x})$$

$$= ||\mathbf{w}_t||_2 + 2\mathbf{w}_t^T\mathbf{x} + ||\mathbf{w}||^2 \leq ||\mathbf{w}_t||_2 + 1$$

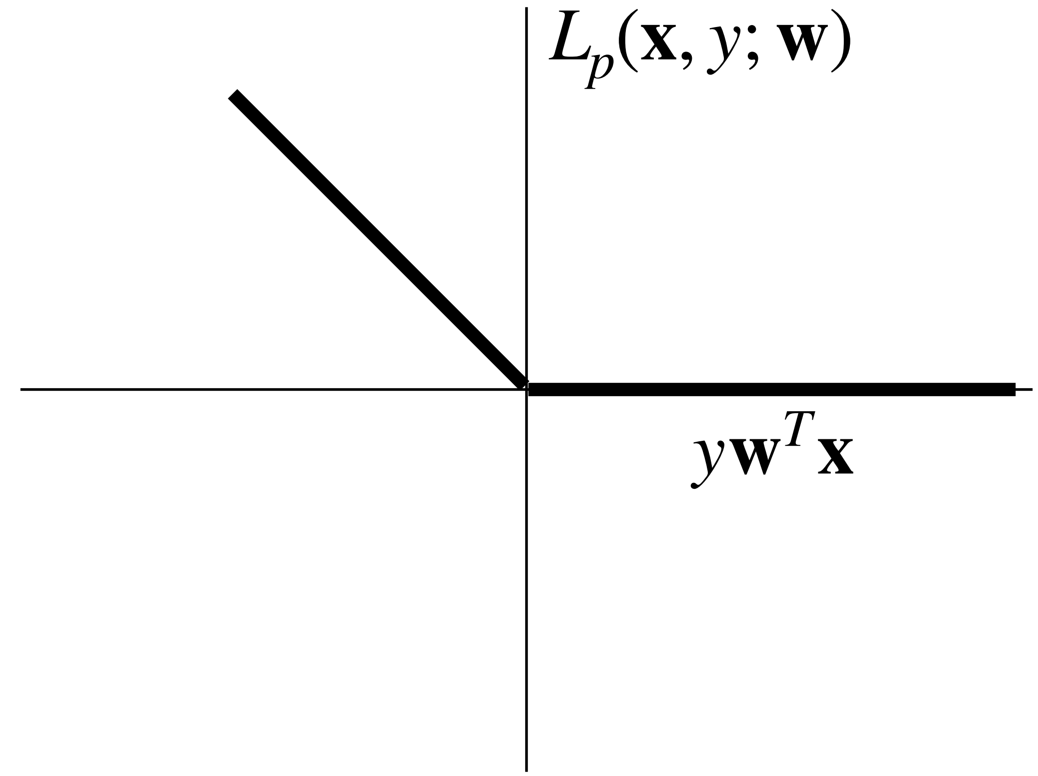
(Similar argument holds for a negative example)

After k updates, we have $\mathbf{w}_k^T\mathbf{u} \geq k\gamma$ and $||\mathbf{w}_k||^2 \leq k$

$$\Rightarrow \sqrt{k} \geq ||\mathbf{w}_k|| \geq \mathbf{w}_k^T\mathbf{u} \geq k\gamma \rightarrow \boxed{k \leq \frac{1}{\gamma^2}}$$

Loss Function of the Perceptron Learner

Hinge Loss: $L_p(\mathbf{x}, y; \mathbf{w}) = \max(0, -y\mathbf{w}^T\mathbf{x})$



A Stochastic Gradient Descent (SGD) weight update on $L_p(\mathbf{x}, y; \mathbf{w}) = \max(0, -y\mathbf{w}^T\mathbf{x})$ gives:

$$\begin{aligned} \mathbf{w} &\leftarrow \mathbf{w} - \nabla_{\mathbf{w}} L_p(\mathbf{x}, y; \mathbf{w}) \\ &\leftarrow \mathbf{w} + y\mathbf{x} \end{aligned}$$

which is exactly the perceptron update rule.