Perceptron

Perceptron is a more sophisticated learning algorithm for producing linear classifiers.

Input: D_n, T
$\theta = \overline{\theta}$
$\theta_0 = 0$
for $t = 1$ to T **do**
  for $i = 1$ to n **do**
    if $y^{(i)}(\theta^T x^i + \theta_0) \leq 0$ **then**
      $\theta = \theta + y^{(i)} x^{(i)}$
      $\theta_0 = \theta_0 + y^{(i)}$
    end if
  end for
end for
return θ, θ_0

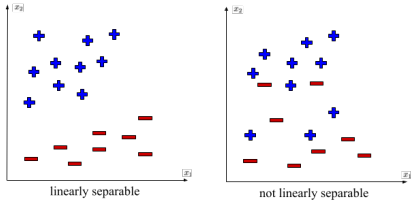
The algorithm iterates through the dataset D_n T times (T is a hyperparameter). At each point in the dataset, it checks if $y^{(i)}(\theta^T x^{(i)} + \theta_0) \leq 0$. The if statement basically determines whether or not the current θ classified the point at index i correctly. If $(x^{(i)}, y^{(i)})$ were correctly classified, then $y^{(i)}(\theta^T x^{(i)} + \theta_0)$ would be positive. Recall that $\theta^T x^{(i)} + \theta_0$ is the input to the sign function in the linear classifier hypothesis. Thus, if it is negative and $y^{(i)}$ is also negative, their product should be positive. If both are positive, their product is positive too. If they are of different signs (an incorrect classification occurred), then the if statement would trigger some modifications. In particular, it modifies θ to $\theta + y^{(i)} x^{(i)}$ and θ_0 to $\theta_0 + y^{(i)}$.

The modifications that perceptron makes to the parameters θ and θ_0 probably are not the obvious "right move." One is probably left wondering why Rosenblatt, the inventor of perceptron, chose to set $\theta = \theta + y^{(i)} x^{(i)}$ and $\theta_0 = \theta_0 + y^{(i)}$. This brings up another interesting discussion point. Whereas the inner workings of most algorithms have been built intuitively around the problem to be solved, perceptron was simply introduced and left for scholars to analyze over the decades. Years of papers have determined that Rosenblatt's modifications are quite functional.

To better analyze how perceptron works and even introduce a theorem about it, exploring a simpler version of the algorithm is useful. Think of perceptron-through-origin as perceptron without any offset (θ_0) parameter. Some playing with dimensions later on will show that what applies to perceptron-through-origin applies to perceptron with an offset.

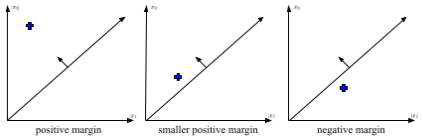
1 Linear Separability

Linear separability is a property of a dataset D_n. D_n is linearly separable when there exists some θ (no θ_0 because the current discussion is about perceptron-through-origin) such that $y^{(i)}(\theta^T x^{(i)}) > 0$ for all i. In other words, all points in the dataset D_n are correctly classified.



2 Margin

The margin of a data point (x, y) with repsect to a linear separator (a hyperplane) is $y\frac{\theta^T x}{||\theta||}$. $\frac{\theta^T x}{||\theta||}$ is the signed distance from the point to the separator. If y is -1 and (x, y) is correctly classified, then $\frac{\theta^T x}{||\theta||}$ should be negative. y is the target label (either $+1$ or -1). Therefore, a correctly classified point should have a positive margin. An incorrectly classified point should have a negative margin. A higher margin is better because the farther a point is away from the classifier, the better classified it is (not a close call).



The margin of an entire dataset with respect to a linear separator is equal to the margin of the point it contains that has the lowest margin: $\min_i y^{(i)} \frac{\theta^T x^{(i)}}{||\theta||}$.

3 Perceptron Convergence Theorem

The perceptron convergence states that if (a) there exists some θ^* such that $y^{(i)} \frac{\theta^T x^{(i)}}{||\theta||} > \gamma > 0$ for all i (in other words, if the margin of the dataset with respect to θ^*) is greater than or equal to some constant γ) and (b) $||x^{(i)}|| \leq R$ for all i (in other words, when graphed, the data points are contained in a circle of radius R), then the perceptron will make at most $(\frac{R}{\gamma})^2$ modifications during training.

4 Proof of the Perceptron Convergence Theorem

Say that $\theta^{(k)}$ is the hypothesis produced after k modifications during training and θ^* is the parameter such that $y^{(i)} \frac{\theta^{*T} x^{(i)}}{||\theta^*||} \geq \gamma > 0$ for all i. The angle between $\theta^{(k)}$ and θ^* is α. To show that $\theta^{(k)}$ will converge to become θ^* as k increases, one must show that the angle α becomes smaller and smaller. As the angle becomes smaller, its cosine should become greater (i.e. $\cos\alpha$ should increase).

1. There is a formula that describes the cosine of the angle α between two vectors: $\cos\alpha = \frac{a \cdot b}{||a||||b||}$. Thus, $\cos\alpha = \frac{\theta^* \cdot \theta^{(k)}}{||\theta^*||||\theta^{(k)}||}$.

2. One can break $\frac{\theta^* \cdot \theta^{(k)}}{||\theta^*||||\theta^{(k)}||}$ down into $\frac{\theta^* \cdot \theta^{(k)}}{||\theta^*||} \cdot \frac{1}{||\theta^{(k)}||}$.

3. First, one should analyze $\frac{\theta^* \cdot \theta^{(k)}}{||\theta^*||}$. $\theta^{(k)} = (\theta^{(k-1)} + y^{(i)} x^{(i)})$ when i is the index of the point at which perceptron made its last modification to θ. Using this information, $\frac{\theta^* \cdot \theta^{(k)}}{||\theta^*||} = \frac{(\theta^{(k-1)} + y^{(i)} x^{(i)})\theta^*}{||\theta^*||} = \frac{\theta^{(k-1)} \cdot \theta^*}{||\theta^*||} + \frac{y^{(i)} x^{(i)} \cdot \theta^*}{||\theta^*||}$. Recall that $\frac{y^{(i)} x^{(i)} \cdot \theta^*}{||\theta^*||}$ is the margin, which is greater than or equal to γ. Thus, $\frac{\theta^{(k-1)} \cdot \theta^*}{||\theta^*||} + \frac{y^{(i)} x^{(i)} \cdot \theta^*}{||\theta^*||} \geq k\gamma$.

4. Now, one may analyze $\frac{1}{||\theta^{(k)}||}$. $||\theta^{(k)}||^2 = ||\theta^{(k-1)} + y^{(i)} x^{(i)}||^2 = ||\theta^{(k-1)}||^2 + 2y^{(i)}\theta^{(k-1)} \cdot x^{(i)} + ||x^{(i)}||^2$. It is given in the perceptron convergence theorem that $||x^{(i)}||^2 \leq R^2$. Also, $2y^{(i)}\theta^{(k-1)} \cdot x^{(i)}$ is negative because θ^k made a mistake. Therefore, $||\theta^{(k-1)}||^2 + 2y^{(i)}\theta^{(k-1)} \cdot x^{(i)} + ||x^{(i)}||^2 \leq kR^2$. Thus, $\frac{1}{||\theta^{(k)}||} \geq \frac{1}{\sqrt{k}R}$.

5. Putting everything together, $\cos\alpha \geq \sqrt{k}\frac{\gamma}{R}$. Provided that cosine's greatest output is 1, $1 \geq \frac{\sqrt{k}\gamma}{R}$, so $k \leq (\frac{R}{\gamma})^2$.