Perceptron Bounds

Tirgul 4
Perceptron Bound Theorem

Let \( \langle (x_1, y_1), \cdots, (x_m, y_m) \rangle \) be a sequence of labeled examples with \( \|x_i\| \leq R \). Suppose that there exists a vector \( w^* \) such that \( \|w^*\| = 1 \) and \( y_i(w^* \cdot x_i) \geq \gamma \) for all examples in the sequence. Then the number of mistakes made by the online perceptron algorithm on the sequence is at most \( \left( \frac{R}{\gamma} \right)^2 \)

Note: \( \gamma \) is called a “margin”.
Perceptron Bound Theorem: Intuition

• We assumed that the data is linearly separable;
  • Hence, there exists some “perfect” $w^*$

• Perceptron aims to find a $w$ that is parallel to $w^*$
  • (or as close as possible to it)
Reminder:

• Dot product (geometric definition):
  • $\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\|\|\mathbf{b}\| \cos(\theta)$

• Where:
  • $\|\mathbf{x}\|$ is the magnitude of vector $\mathbf{x}$.
  • $\theta$ is the angle between $\mathbf{a}$ and $\mathbf{b}$. 
Perceptron Bound Theorem: Intuition

• To see whether two vectors are parallel, use: $w^* \cdot w$

• $w^* \cdot w = \|w^*\|\|w\|\cos(\theta)$

• If $\theta = 0^\circ$, then $\cos(\theta) = 1$
  • So the size of the dot product is a maximum
Perceptron Bound Theorem: Intuition

\[ \mathbf{w}^* \cdot \mathbf{w} = ||\mathbf{w}^*|| ||\mathbf{w}|| \cos(\theta) \]

• Prove that:

1. At each update, \( \mathbf{w}^* \cdot \mathbf{w} \) increases
   • Which implies that the vectors are getting closer...

2. Make sure that \( ||\mathbf{w}|| \) doesn’t increase too much as well
Perceptron Bound Theorem: Proof

• Claim 1:

\[ w^{(t+1)} \cdot w^* \geq w^{(t)} \cdot w^* + \gamma \]

• This means that after each update, the dot product increases by at least \( \gamma \). After \( M \) updates:

\[ w^{(t+1)} \cdot w^* \geq M\gamma \]
Perceptron Bound Theorem: Proof

• Claim 2:
  • The length of the weight vector after $M$ updates is:

\[ \|w^{(t+1)}\|^2 \leq MR^2 \]
Perceptron Bound Theorem

• Using the fact that (Cauchy-Schwartz inequality):

\[ w^{(t+1)} \cdot w^* \leq \|w^{(t+1)}\|\|w^*\| \]

• Combining claim 1 and 2, we get that:

\[ M \leq \left( \frac{R}{\gamma} \right)^2 \]
Perceptron Bound Theorem: Summary

- If the data is linearly separable, then the algorithm will converge.

- The number of mistakes till it converges \( \leq \left( \frac{R}{\gamma} \right)^2 \), i.e., is a function of the maximal length of \( x_i \) and the minimal margin \( \gamma \).
  - Margin (\( \gamma \)): the distance between the separating hyperplane and the nearest datapoint.

- Perceptron stops learning when it gets all of the training data correct
  - So there is no guarantee that it will find the largest margin.
The XOR Function

<table>
<thead>
<tr>
<th>$\text{In}_1$</th>
<th>$\text{In}_2$</th>
<th>$t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
The XOR Function

• We cannot draw a straight line that separates \textit{true} from \textit{false}.
  • XOR is not \textit{linearly separable}.

• Perceptron is not capable of solving this problem.

• Solution:
  • Make the network more complicated (later in the course).
The XOR Function

• XOR is impossible to solve using a linear function?
  • It could be solved, by adding dimensions!

• Rewrite the problem in 3D
• Find a plane that can separate the two classes
The XOR Function

• It’s always possible to separate two classes with a linear function, provided that you project data to the correct dimension (=like kernel SVM).

<table>
<thead>
<tr>
<th>In₁</th>
<th>In₂</th>
<th>In₃</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

A decision boundary (the shaded plane) solving the XOR problem in 3D with the crosses below the surface and the circles above it.
Mistake Bound for Inseparable Data

• What if only most of the data is separable by a large margin?

• Define the deviation of each example as:

\[ \xi_i = \max\{0, \gamma - y_i(w^T x_i)\} \]
Mistake Bound for Inseparable Data

• Theorem: Let \( (x_1, y_1), \ldots, (x_n, y_n) \) be a sequence of labeled examples with \( \|x_i\| \leq R \). Let \( w^* \) be any vector with \( \|w^*\| = 1 \) and let \( \gamma > 0 \). Define the deviation of each example as

\[
\xi_i = \max\{0, \gamma - y_i (w^T x_i)\}
\]

and define \( D = \sqrt{\sum_{i=1}^{n} \xi_i^2} \). Then the number of mistakes of the online perceptron algorithm on this sequence is bounded by:

\[
\left( \frac{R + D}{\gamma} \right)^2.
\]