Neural Networks

- Decision trees are good at modeling nonlinear interactions among a small subset of attributes
- Sometimes we are interested in linear interactions among all attributes
- Simple neural networks are good at modeling such interactions
- The resulting models have close connections with naïve Bayes.
Learning Threshold Functions

• Outline
• Background
• Threshold logic functions
• Connection to logic
• Connection to geometry
• Learning threshold functions – perceptron algorithm and its variants
• Perceptron convergence theorem

Background – Neural computation

• 1900 – Birth of neuroscience – Ramon Cajal et al.
• 1913 – Behaviorist or stimulus response psychology
• 1930-50: Theory of Computation, Church-Turing Thesis
• 1943: McCulloch & Pitts “A logical calculus of neuronal activity”
• 1949: Hebb – Organization of Behavior
• 1956 – Birth of Artificial Intelligence – “Computers and Thought”
• 1960-65: Perceptron model developed by Rosenblatt
Background – Neural computation

- 1969: Minsky and Papert criticize Perceptron
- 1969: Chomsky argues for universal innate grammar
- 1970: Rise of cognitive psychology and knowledge-based AI
- 1975: Learning algorithms for multi-layer neural networks
- 1985: Resurgence of neural networks and machine learning
- 1988: Birth of computational neuroscience
- 1990: Successful applications (stock market, OCR, robotics)
- 1990-2000 New synthesis of behaviorist and cognitive or representational approaches in AI and psychology
- 2000-2006 Synthesis of logical and probabilistic approaches to representation and learning

Background – Brains and Computers

- Brain consists of $10^{11}$ neurons, each of which is connected to $10^4$ neighbors
- Each neuron is slow (1 millisecond to respond to a stimulus) but the brain is astonishingly fast at perceptual tasks (e.g. face recognition)
- Brain processes and learns from multiple sources of sensory information (visual, tactile, auditory…)
- Brain is massively parallel, shallowly serial, modular and roughly hierarchical with recurrent and lateral connectivity within and between modules
- If cognition is -- or at least can be modeled by -- computation, it is natural to ask how and what brains compute
Brain and information processing

- Primary somato-sensory cortex
- Motor association cortex
- Primary motor cortex
- Sensory association area
- Auditory cortex
- Speech comprehension
- Visual association area
- Primary visual cortex
- Auditory association area
- Prefrontal cortex

Neural Networks

Ramon Cajal, 1900
Neurons and Computation

McCulloch-Pitts computational model of a neuron

\[ y = \begin{cases} 1 & \text{if } \sum_{i=0}^{n} w_i x_i > 0 \\ -1 & \text{otherwise} \end{cases} \]
Threshold neuron – Connection with Geometry

\[ w_1 x_1 + w_2 x_2 + w_0 = 0 \]

Describes a hyperplane which divides the instance space \( \mathbb{R}^n \) into two half-spaces \( C_1 \) and \( C_2 \):

\[ \{ x_p \in \mathbb{R}^n \mid W \cdot X_p + w_0 > 0 \} \]

and

\[ \{ x_p \in \mathbb{R}^n \mid W \cdot X_p + w_0 < 0 \} \]

McCulloch-Pitts Neuron or Threshold Neuron

\[ y = \text{sign} \left( W \cdot X + w_0 \right) \]

\[ = \text{sign} \left( \sum_{i=0}^{n} w_i x_i \right) \]

\[ = \text{sign} \left( W^T X + w_0 \right) \]

\[ \text{sign}(v) = 1 \text{ if } v > 0 \]

\[ = 0 \text{ otherwise} \]
Threshold neuron – Connection with Geometry

Instance space \( \mathbb{R}^n \)

Hypothesis space is the set of \((n-1)\)-dimensional hyperplanes defined in the \(n\)-dimensional instance space

A hypothesis is defined by

\[
\sum_{i=0}^{n} w_i x_i = 0
\]

- Orientation of the hyperplane is governed by \( \begin{pmatrix} w_1 & \ldots & w_n \end{pmatrix}^T \)
- and the perpendicular distance of the hyperplane from the origin is given by

\[
\begin{pmatrix}
|w_0| \\
\sqrt{w_1^2 + w_2^2 + \ldots + w_n^2}
\end{pmatrix}
\]
Threshold neuron as a pattern classifier

- The threshold neuron can be used to classify a set of instances into one of two classes $C_1$, $C_2$
- If the output of the neuron for input pattern $X_p$ is +1 then $X_p$ is assigned to class $C_1$
- If the output is -1 then the pattern $X_p$ is assigned to $C_2$

Example

$$
[w_0 \ w_1 \ w_2]^T = [-1 -1 1]^T \\
X_p^T = [1 \ 0]^T \ W \cdot X_p + w_0 = -1 + (-1) = -2
$$

$X_p$ is assigned to class $C_2$

Threshold neuron – Connection with Logic

- Suppose the input space is $\{0,1\}^n$
- Then threshold neuron computes a Boolean function $f : \{0,1\}^n \rightarrow \{-1,1\}$

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$g(X)$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>-1.5</td>
<td>-1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>-0.5</td>
<td>-1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>-0.5</td>
<td>-1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.5</td>
<td>1</td>
</tr>
</tbody>
</table>

Example

Let $w_0 = -1.5$; $w_1 = w_2 = 1$

In this case, the threshold neuron implements the logical AND function
Threshold neuron – Connection with Logic

• A threshold neuron with the appropriate choice of weights can implement Boolean AND, OR, and NOT function

• **Theorem**: For any arbitrary Boolean function \( f \), there exists a network of threshold neurons that can implement \( f \).

• **Theorem**: Any arbitrary finite state automaton can be realized using threshold neurons and delay units

• Networks of threshold neurons, given access to unbounded memory, can compute any Turing-computable function

• **Corollary**: Brains if given access to enough working memory, can compute any computable function

Threshold neuron: Connection with Logic

**Theorem**: There exist functions that cannot be implemented by a single threshold neuron.

**Example** Exclusive OR

Why?
Threshold neuron – Connection with Logic

- **Definition:** A function that can be computed by a single threshold neuron is called a threshold function.
- Of the 16 2-input Boolean functions, 14 are Boolean threshold functions.
- As $n$ increases, the number of Boolean threshold functions becomes an increasingly small fraction of the total number of $n$-input Boolean functions.

\[
N_{\text{Threshold}}(n) \leq 2^n^2 \quad \quad N_{\text{Boolean}}(n) = 2^{2^n}
\]

Terminology and Notation

- **Synonyms:** Threshold function, Linearly separable function, linear discriminant function.
- **Synonyms:** Threshold neuron, McCulloch-Pitts neuron, Perceptron, Threshold Logic Unit (TLU).
- We often include $w_0$ as one of the components of $W$ and incorporate $x_0$ as the corresponding component of $X$ with the understanding that $x_0=1$. Then $y=1$ if $W.X > 0$ and $y=-1$ otherwise.
Learning Threshold functions

A training example $E_k$ is an ordered pair $(X_k, d_k)$ where

$$X_k = [x_{0k} \ x_{1k} \ldots x_{nk}]^T$$

is an $(n+1)$ dimensional input pattern, $d_k = f(X_k) \in \{-1, 1\}$

is the desired output of the classifier and $f$ is an unknown target function to be learned.

A training set $E$ is simply a multi-set of examples.

---

Learning Task: Given a linearly separable training set $E$, find a solution $W^*$ such that

- $\forall X_p \in S^+, W^* \cdot X_p > 0$
- $\forall X_p \in S^-, W^* \cdot X_p < 0$

where

$$S^+ = \{X_k | (X_k, d_k) \in E \text{ and } d_k = 1\}$$

$$S^- = \{X_k | (X_k, d_k) \in E \text{ and } d_k = -1\}$$
Rosenblatt’s Perceptron Learning Algorithm

Initialize $W = [0 \ 0 \ldots 0]^T$  Set learning rate $\eta > 0$

Repeat until a complete pass through $E$ results in no weight updates

For each training example $E_k \in E$

\[
\begin{align*}
\quad y_k &\leftarrow \text{sign} (W \cdot X_k) \\
W &\leftarrow W + \eta(d_k - y_k)X_k
\end{align*}
\]

$W^* \leftarrow W$; Return $W^*$

---

Perceptron learning algorithm – Example

Let

$S^+ = \{(1, 1, 1), (1, 1, -1), (1, 0, -1)\}$ \\
$S^- = \{(1,-1, -1), (1,-1, 1), (1,0, 1)\}$ \\
$W= (0 \ 0 \ 0)^T$; \\
$\eta = \frac{1}{2}$

<table>
<thead>
<tr>
<th>$X_k$</th>
<th>$d_k$</th>
<th>$W$</th>
<th>$W \cdot X_k$</th>
<th>$y_k$</th>
<th>Update?</th>
<th>Updated W</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1, 1, 1)</td>
<td>1</td>
<td>(0, 0, 0)</td>
<td>0</td>
<td>-1</td>
<td>Yes</td>
<td>(1, 1, 1)</td>
</tr>
<tr>
<td>(1, 1, -1)</td>
<td>1</td>
<td>(1, 1, 1)</td>
<td>1</td>
<td>1</td>
<td>No</td>
<td>(1, 1, 1)</td>
</tr>
<tr>
<td>(1,0, -1)</td>
<td>1</td>
<td>(1, 1, 1)</td>
<td>0</td>
<td>-1</td>
<td>Yes</td>
<td>(2, 1, 0)</td>
</tr>
<tr>
<td>(1,-1, -1)</td>
<td>-1</td>
<td>(2, 1, 0)</td>
<td>1</td>
<td>1</td>
<td>Yes</td>
<td>(1, 2, 1)</td>
</tr>
<tr>
<td>(1,-1, 1)</td>
<td>-1</td>
<td>(1, 2, 1)</td>
<td>0</td>
<td>-1</td>
<td>No</td>
<td>(1, 2, 1)</td>
</tr>
<tr>
<td>(1,0, 1)</td>
<td>-1</td>
<td>(1, 2, 1)</td>
<td>2</td>
<td>1</td>
<td>Yes</td>
<td>(0, 2, 0)</td>
</tr>
<tr>
<td>(1, 1, 1)</td>
<td>1</td>
<td>(0, 2, 0)</td>
<td>2</td>
<td>1</td>
<td>No</td>
<td>(0, 2, 0)</td>
</tr>
</tbody>
</table>
Perceptron (1957)

Copyright Vasant Honavar, 2006.
Perceptron Convergence Theorem (Novikoff)

Theorem Let $E = \{(X_k, d_k)\}$ be a training set where $X_k \in \{1\} \times \mathbb{R}^n$ and $d_k \in \{-1,1\}$.

Let $S^+ = \{X_k | (X_k, d_k) \in E \& d_k = 1\}$ and $S^- = \{X_k | (X_k, d_k) \in E \& d_k = -1\}$.

The perceptron algorithm is guaranteed to terminate after a bounded number $t$ of weight updates with a weight vector $W^*$ such that $\forall X_k \in S^+, W^* \cdot X_k \geq \delta$ and $\forall X_k \in S^-, W^* \cdot X_k \leq -\delta$ for some $\delta > 0$, whenever such $W^* \in \mathbb{R}^{n+1}$ and $\delta > 0$ exist -- that is, $E$ is linearly separable. The bound on the number $t$ of weight updates is given by

$$t \leq \left( \frac{\|W^*\|L}{\delta} \right)^2$$

where $L = \max_{X \in S} \|X\|$ and $S = S^+ \cup S^-$.

Proof of Perceptron Convergence Theorem

Let $W_t$ be the weight vector after $t$ weight updates.

$$W^*$$

$\theta$

$$W_t$$

Invariant: $\forall \theta \ |\cos \theta| \leq 1$
Proof of Perceptron Convergence Theorem

Let $W^*$ be such that
\[
\forall X_k \in S^+, W^* \cdot X_k \geq \delta \quad \text{and} \quad \forall X_k \in S^-, W^* \cdot X_k \leq -\delta
\]

WLOG assume that $W^* \cdot X = 0$ passes through the origin.

Let $\forall X_k \in S^+, Z_k = X_k$,
\[
\forall X_k \in S^-, Z_k = -X_k,
\]
\[
Z = \{Z_k\}
\]
\[
(\forall X_k \in S^+, W^* \cdot X_k \geq \delta \quad \& \quad \forall X_k \in S^-, W^* \cdot X_k \leq -\delta) \iff (\forall Z_k \in Z, W^* \cdot Z_k \geq \delta)
\]

Let $E' = \{(Z_k, 1)\}$

---

Proof of Perceptron Convergence Theorem

\[
W_{t+1} = W_t + \eta (d_k - y_k) Z_k
\]
where $W_0 = [0 \; 0 \; \ldots \; 0]^T$ and $\eta > 0$

[Weight update based on example $(Z_k, 1)$]
\[
\iff [(d_k = 1) \land (y_k = -1)]
\]
\[
\therefore W^* \cdot W_{t+1} = W^* \cdot (W_t + 2 \eta Z_k)
\]
\[
= (W^* \cdot W_t) + 2 \eta (W^* \cdot Z_k)
\]

Since $\forall Z_k \in Z, (W^* \cdot Z_k \geq \delta)$,
\[
W^* \cdot W_{t+1} \geq W^* \cdot W_t + 2 \eta \delta
\]
\[
\therefore \forall t \quad W^* \cdot W_t \geq 2 t \eta \delta ................................................(a)
\]
Proof of Perceptron Convergence Theorem

\[ \|W_{t+1}\|^2 = W_{t+1} \cdot W_{t+1} \]
\[ = (W_t + 2\eta Z_k) \cdot (W_t + 2\eta Z_k) \]
\[ = (W_t \cdot W_t) + 4\eta (W_t \cdot Z_k) + 4\eta^2 (Z_k \cdot Z_k) \]

Note weight update based on \( Z_k \Leftrightarrow (W_t \cdot Z_k \leq 0) \)

\[ \therefore \|W_{t+1}\|^2 \leq \|W_t\|^2 + 4\eta^2 \|Z_k\|^2 \leq \|W_t\|^2 + 4\eta^2 L^2 \]

Hence \( \|W_t\|^2 \leq 4\eta^2 L^2 \)

\[ \therefore \forall t \|W_t\| \leq 2\eta L\sqrt{t} \]

.................................(b)

Proof of Perceptron Convergence Theorem

From (a) we have: \( \forall t \ (W^* \cdot W_t) \geq 2t\eta \delta \)

\[ \Rightarrow \forall t \ 2t\eta \delta \leq (W^* \cdot W_t) \Rightarrow \forall t \ 2t\eta \delta \leq \|W^*\| \|W_t\| \cos \theta \]

\[ \Rightarrow \forall t \ 2t\eta \delta \leq \|W^*\| \|W_t\| \] : \( \forall \theta \ \cos \theta \leq 1, \)

Substituting for an upper bound on \( \|W_t\| \) from (b),

\[ \forall t \ 2t\eta \delta \leq \|W^*\| 2\eta L\sqrt{t} \Rightarrow \forall t \ (\delta \sqrt{t} \leq \|W^*\| L) \]

\[ \Rightarrow t \leq \left( \frac{\|W^*\| L}{\delta} \right)^2 \]
Notes on the Perceptron Convergence Theorem

- The bound on the number of weight updates does not depend on the learning rate.
- The bound is not useful in determining when to stop the algorithm because it depends on the norm of the unknown weight vector and delta.
- The convergence theorem offers no guarantees when the training data set is not linearly separable.

Exercise: Prove that the perceptron algorithm is robust with respect to fluctuations in the learning rate:

$$0 < \eta_{\text{min}} \leq \eta_t \leq \eta_{\text{max}} < \infty$$

Multiple classes

- $K - 1$ binary classifiers
- $\frac{K(K-1)}{2}$ binary classifiers

One-versus-rest

One-versus-one

Problem: Green region has ambiguous class membership.
Multi-category classifiers

Define $K$ linear functions of the form:

$$ y_k (X) = W_k^T X + w_{k0} $$

$$ h(X) = \arg \max_k y_k (X) = \arg \max_k (W_k^T X + w_{k0}) $$

Decision surface between class $C_k$ and $C_j$

$$ (W_k - W_j)^T X + (w_{k0} - w_{j0}) = 0 $$

Linear separator for $K$ classes

- Decision regions defined by

  $$ (W_k - W_j)^T X + (w_{k0} - w_{j0}) = 0 $$

  are singly connected and convex

For any points $X_A, X_B \in R_k$,

any $\hat{X}$ that lies on the line connecting $X_A$ and $X_B$

$$ \hat{X} = \lambda X_A + (1 - \lambda) X_B \text{ where } 0 \leq \lambda \leq 1 $$

also lies in $R_k$
Winner-Take-All Networks

\[ y_{ip} = 1 \text{ iff } W_i \cdot X_p > W_j \cdot X_p \quad \forall j \neq i \]

\[ y_{ip} = 0 \text{ otherwise} \quad \text{Note: } W_j \text{ are augmented weight vectors} \]

\[ W_i = [1 \ -1 \ -1]^T, \ W_2 = [1 \ 1 \ 1]^T, \ W_3 = [2 \ 0 \ 0]^T \]

\[
\begin{array}{|c|c|c|c|c|}
\hline
 & W_1.X_p & W_2.X_p & W_3.X_p & y_1 & y_2 & y_3 \\
\hline
1 & -1 & -1 & 3 & -1 & 2 & 1 & 0 & 0 \\
1 & -1 & +1 & 1 & 1 & 2 & 0 & 0 & 1 \\
1 & +1 & -1 & 1 & 1 & 2 & 0 & 0 & 1 \\
1 & +1 & +1 & -1 & 3 & 2 & 0 & 1 & 0 \\
\hline
\end{array}
\]

What does neuron 3 compute?

Linear separability of multiple classes

Let \( S_1, S_2, S_3, \ldots S_M \) be multisets of instances

Let \( C_1, C_2, C_3, \ldots C_M \) be disjoint classes

\[ \forall i \quad S_i \subseteq C_i \]

\[ \forall i \neq j \quad C_i \cap C_j = \emptyset \]

We say that the sets \( S_1, S_2, S_3, \ldots S_M \) are linearly separable iff \( \exists \) weight vectors \( W_1^*, W_2^*, \ldots W_M^* \) such that

\[ \forall i \quad \left( \forall X_p \in S_i, \left( W_i^* \cdot X_p > W_j^* \cdot X_p \right) \forall j \neq i \right) \]

Copyright Vasant Honavar, 2006.
Training WTA Classifiers

\[ d_{kp} = 1 \text{ iff } X_p \in C_k; \quad d_{kp} = 0 \text{ otherwise} \]
\[ y_{kp} = 1 \text{ iff } W_k \cdot X_p > W_j \cdot X_p \quad \forall k \neq j \]

Suppose \( d_{kp} = 1, y_{jp} = 1 \) and \( y_{kp} = 0 \)

\[ W_k \leftarrow W_k + \eta X_p; \quad W_j \leftarrow W_j - \eta X_p; \]

All other weights are left unchanged.

Suppose \( d_{kp} = 1, y_{jp} = 0 \) and \( y_{kp} = 1 \).

The weights are unchanged.

Suppose \( d_{kp} = 1, \forall j \, y_{jp} = 0 \) (there was a tie)

\[ W_k \leftarrow W_k + \eta X_p \]

All other weights are left unchanged.

WTA Convergence Theorem

Given a linearly separable training set, the WTA learning algorithm is guaranteed to converge to a solution within a finite number of weight updates.

Proof Sketch: Transform the WTA training problem to the problem of training a single perceptron using a suitably transformed training set. Then the proof of WTA learning algorithm reduces to the proof of perceptron learning algorithm.
WTA Convergence Theorem

Let $\mathbf{W}^T = [\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_M]^T$ be a concatenation of the weight vectors associated with the $M$ neurons in the WTA group. Consider a multi-category training set $E = \{(\mathbf{X}_p, f(\mathbf{X}_p))\}$ where $\forall \mathbf{X}_p$, $f(\mathbf{X}_p) \in \{C_1, \ldots, C_M\}$.

Let $\mathbf{X}_p \in C_i$. Generate $(M-1)$ training examples using $\mathbf{X}_p$ for an $M(n+1)$-input perceptron:

- $\mathbf{X}_{p12} = [\mathbf{X}_p - \mathbf{X}_p\phi \phi \ldots \phi]$
- $\mathbf{X}_{p13} = [\mathbf{X}_p \phi - \mathbf{X}_p \phi \phi \ldots \phi]$
- ...
- $\mathbf{X}_{p1M} = [\mathbf{X}_p \phi \phi \ldots \phi - \mathbf{X}_p]$  

where $\phi$ is an all zero vector with the same dimension as $\mathbf{X}_p$ and set the desired output of the corresponding perceptron to be 1 in each case.

Similarly, from each training example for an $(n+1)$-input WTA, we can generate $(M-1)$ examples for an $M(n+1)$ input single neuron. Let the union of the resulting $|E|(M-1)$ examples be $E'$.

By construction, there is a one-to-one correspondence between the weight vector $\mathbf{W}^T = [\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_M]^T$ that results from training an $M$-neuron WTA on the multi-category set of examples $E$ and the result of training an $M(n+1)$ input perceptron on the transformed training set $E'$. Hence the convergence proof of WTA learning algorithm follows from the perceptron convergence theorem.
Weight space representation

Pattern space representation
- Coordinates of space correspond to attributes (features)
- A point in the space represents an instance
- Weight vector $W_v$ defines a hyperplane $W_v \cdot X = 0$

Weight space (dual) representation
- Coordinates define a weight space
- A point in the space represents a choice of weights $W_v$
- An instance $X_p$ defines a hyperplane $W$. $X_p = 0$
Weight space representation

\[ W_{t+1} \leftarrow W_t + \eta X_p \]

**Fractional correction rule**

\[ W_{t+1} \leftarrow W_t + \lambda \left( \frac{|W_t \cdot X_p| + \varepsilon}{X_p \cdot X_p + \varepsilon} \right) (d_p - y_p) X_p \]

- \( \lambda > 0 \) is a constant (to handle the case when the dot product \( W_t \cdot X_p \), or \( X_p \cdot X_p \) (or both) approach zero.
- \( 0 < \lambda < 1; \lambda \neq 0.5 \) when \( d_p, y_p \in \{-1,1\} \)

The Perceptron Algorithm Revisited

The perceptron works by adding misclassified positive or subtracting misclassified negative examples to an arbitrary weight vector, which (without loss of generality) we assumed to be the zero vector. So the final weight vector is a linear combination of training points

\[ w = \sum_{i=1}^{l} \alpha_i y_i x_i, \]

where, since the sign of the coefficient of \( x_i \) is given by label \( y_i \), the \( \alpha_i \) are positive values, proportional to the number of times, misclassification of \( x_i \) has caused the weight to be updated. It is called the embedding strength of the pattern \( x_i \).
Dual Representation

The decision function can be rewritten as:

\[
h(x) = \text{sgn}(\langle w, x \rangle + b) = \text{sgn}\left(\sum_{j=1}^{l} \alpha_j y_j \langle x_j, x \rangle + b\right)
\]

\[
= \text{sgn}\left(\sum_{j=1}^{l} \alpha_j y_j \langle x_j, x \rangle + b\right)
\]

on training example \((x_i, y_i)\)

The update rule is:

if \(y_i \left(\sum_{j=1}^{l} \alpha_j y_j \langle x_j, x_i \rangle + b\right) \leq 0,

then \(\alpha_i \leftarrow \alpha_i + \eta\)

WLOG, we can take \(\eta = 1\).