Outline

• History
• Problem Setting
• How Backpropagation Works
• Examples
• Comments
History

- Backpropagation = Backward Propagation of Errors.


- Had a huge impact in the field of neural networks.
Problem Setting

Given a (multilayer) neural network
Each neuron performs a weighted sum over inputs Followed by a differentiable activation function (typically logistic)

\[ f(y) = \frac{1}{1 + \exp(-y)} \]
Problem Setting

Given a set of input/output training pairs.
We want to adjust the weights to minimize the mean square error.

\[ E = \frac{1}{2} \sum_c \sum_j (y_{j,c} - d_{j,c})^2 \]

Backpropagation is a systematic way to compute gradients:

\[ \frac{\partial E}{\partial w_{ij}} \]
How Backpropagation Works

• Notation:
  – $y_i$: Output from node $i$
  – $w_{ji}$: Weight from node $i$ to $j$
  – $x_j$: Total input to node $j$ ($x_j = \sum_i w_{ji}y_i$)
  – $f$: Activation function ($y_i = f(x_i)$)

• Knowing $\frac{\partial E}{\partial y}$ for all nodes in layer $i$ enables us to compute $\frac{\partial E}{\partial y}$ in layer $i - 1$

$$\frac{\partial E}{\partial y_i} = \sum_j \frac{\partial E}{\partial y_j} \cdot \frac{\partial y_j}{\partial x_j} \cdot \frac{\partial x_j}{\partial y_i} = \sum_j \frac{\partial E}{\partial y_j} \cdot \frac{\partial f(x_j)}{\partial x_j} \cdot w_{ji}$$
How Backpropagation Works

• For output layer:
  \[
  \frac{\partial E}{\partial y_i} = y_i - d_i \quad \text{(single example)}
  \]

• Gradient w.r.t weights:
  \[
  \frac{\partial E}{\partial w_{ji}} = \frac{\partial E}{\partial y_j} \cdot \frac{\partial y_j}{\partial x_j} \cdot \frac{\partial x_j}{\partial w_{ji}} = \frac{\partial E}{\partial y_j} \cdot \frac{\partial f(x_j)}{\partial x_j} \cdot y_i
  \]
How Backpropagation Works

• Update Frequency
  – Batch: Compute the gradient over the whole dataset then update weights.
  – Online: Update weights for each example.

• Step Size
\[ \Delta w_{ij} = -\epsilon \frac{\partial E}{\partial w_{ij}} \]

• Acceleration
\[ \Delta w_{ij}(t) = -\epsilon \frac{\partial E}{\partial w_{ij}} + \alpha \Delta w_{ij}(t - 1) \]
Example: Detecting Mirror Similarity
Example: Learning Relations

Christopher = Penelope
Margaret = Arthur
Victoria = James
Jennifer = Charles
Colin
Charlotte

Roberto = Maria
Gina = Emilio
Lucia = Marco
Angela = Tomaso
Alfonso
Sophia
Example: Iterative Nets
Comments

• Simple, easy to implement and highly parallelizable

• Applies (at least in principle) to any number of nodes and hidden layers.

• Requires a differentiable activation function, not applicable to hard thresholding

• Local method, prone to local optima

• First order method, slow convergence

• Not a plausible model for learning in brains
Is Backpropagation Dead?

Backpropagation 1963-200?
Thanks