Assessment

Assessment will consist of:

- Assignment 1 16%
- Assignment 2 24%
- Written Exam 60%

In order to pass the course, you must score

- at least 16/40 for the assignments
- at least 24/60 for the written exam
- a combined mark of at least 50/100

McCulloch & Pitts Model of a Single Neuron

\[
\begin{align*}
    x_1 & \xrightarrow{w_1} \Sigma \\
    x_2 & \xrightarrow{w_2} \\
    w_0 & = \text{th}
\end{align*}
\]

\[
s = w_1 x_1 + w_2 x_2 - \text{th} = w_1 x_1 + w_2 x_2 + w_0
\]

\[
g(s) \xrightarrow{g} g(s)
\]

- \(x_1, x_2\) are inputs
- \(w_1, w_2\) are synaptic weights
- \(\text{th}\) is a threshold
- \(w_0\) is a bias weight
- \(g\) is transfer function
Perceptron Learning Rule

Adjust the weights as each input is presented.
recall: \( s = w_1 x_1 + w_2 x_2 + w_0 \)

if \( g(s) = 0 \) but should be 1, \( w_k \leftarrow w_k + \eta x_k \)
if \( g(s) = 1 \) but should be 0, \( w_k \leftarrow w_k - \eta x_k \)

\( w_0 \leftarrow w_0 + \eta \)
\( w_0 \leftarrow w_0 - \eta \)

so \( s \leftarrow s + \eta (1 + \sum_k x_k^2) \)
so \( s \leftarrow s - \eta (1 + \sum_k x_k^2) \)

otherwise, weights are unchanged. (\( \eta > 0 \) is called the learning rate)

**Theorem:** This will eventually learn to classify the data correctly, as long as they are linearly separable.

Limitations of Perceptrons

Problem: many useful functions are not linearly separable (e.g. XOR)

Possible solution:
\( x_1 \text{ XOR } x_2 \) can be written as: \( (x_1 \text{ AND } x_2) \text{ NOR } (x_1 \text{ NOR } x_2) \)

Recall that AND, OR and NOR can be implemented by perceptrons.

Types of Learning

- **Supervised Learning**
  - agent is presented with examples of inputs and their target outputs

- **Reinforcement Learning**
  - agent is not presented with target outputs, but is given a reward signal, which it aims to maximize

- **Unsupervised Learning**
  - agent is only presented with the inputs themselves, and aims to find structure in these inputs
Ockham’s Razor

“The most likely hypothesis is the simplest one consistent with the data.”

Since there can be noise in the measurements, in practice need to make a tradeoff between simplicity of the hypothesis and how well it fits the data.

Local Search in Weight Space

Problem: because of the step function, the landscape will not be smooth but will instead consist almost entirely of flat local regions and “shoulders”, with occasional discontinuous jumps.

Two-Layer Neural Network

Normally, the numbers of input and output units are fixed, but we can choose the number of hidden units.

Key Idea

Replace the (discontinuous) step function with a differentiable function, such as the sigmoid:

\[ g(s) = \frac{1}{1 + e^{-s}} \]

or hyperbolic tangent

\[ g(s) = \tanh(s) = \frac{e^s - e^{-s}}{e^s + e^{-s}} = 2\left(\frac{1}{1 + e^{-2s}}\right) - 1 \]
### Gradient Descent (4.3)

Recall that the **error function** $E$ is (half) the sum over all input patterns of the square of the difference between actual output and desired output

$$E = \frac{1}{2} \sum (z - t)^2$$

The aim is to find a set of weights for which $E$ is very low.

If the functions involved are smooth, we can use multi-variable calculus to adjust the weights in such a way as to take us in the steepest downhill direction.

$$w' \leftarrow w - \eta \frac{\partial E}{\partial w}$$

Parameter $\eta$ is called the **learning rate**.

---

### Cross Entropy

For classification tasks, target $t$ is either 0 or 1, so better to use

$$E = -t \log(z) - (1 - t) \log(1 - z)$$

This can be justified mathematically, and works well in practice – especially when negative examples vastly outweigh positive ones.

It also makes the backprop computations simpler

$$\frac{\partial E}{\partial z} = \frac{z - t}{z(1 - z)}$$

if

$$z = \frac{1}{1 + e^{-s}}$$

$$\frac{\partial E}{\partial s} = \frac{\partial E}{\partial z} \frac{\partial z}{\partial s} = z - t$$

---

### Variations on Backprop

#### Cross Entropy

- **problem**: least squares error function unsuitable for classification, where target = 0 or 1
- **mathematical theory**: maximum likelihood
- **solution**: replace with cross entropy error function

#### Weight Decay

- **problem**: weights “blow up”, and inhibit further learning
- **mathematical theory**: Bayes’ rule
- **solution**: add weight decay term to error function

#### Momentum

- **problem**: weights oscillate in a “rain gutter”
- **solution**: weighted average of gradient over time

---

### Bayes’ Rule (3.11)

The formula for conditional probability can be manipulated to find a relationship when the two variables are swapped:

$$P(a \land b) = P(a | b)P(b) = P(b | a)P(a)$$

$\rightarrow$ **Bayes’ rule**

$$P(a | b) = \frac{P(b | a)P(a)}{P(b)}$$

This is often useful for assessing the probability of an underlying **cause** after an **effect** has been observed:

$$P(\text{Cause}|\text{Effect}) = \frac{P(\text{Effect}|\text{Cause})P(\text{Cause})}{P(\text{Effect})}$$
Bayesian Inference

$H$ is a class of hypotheses

$P(D|h) = \text{probability of data } D \text{ being generated under hypothesis } h \in H.$

$P(h|D) = \text{probability that } h \text{ is correct, given that data } D \text{ were observed.}$

Bayes’ Theorem:

$$P(h|D)P(D) = P(D|h)P(h)$$

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

$P(h)$ is called the prior.

Weight Decay (5.2.2)

Sometimes we add a penalty term to the loss function which encourages the neural network weights $w_j$ to remain small:

$$E = \frac{1}{2} \sum_i (z_i - t_i)^2 + \frac{\lambda}{2} \sum_j w_j^2$$

This can prevent the weights from “saturating” to very high values.

It is sometimes referred to as “elastic weights” because the weights experience a force as if there were a spring pulling them back towards the origin according to Hooke’s Law.

The scaling factor $\lambda$ needs to be determined from experience, or empirically.

Momentum (8.3)

If landscape is shaped like a “rain gutter”, weights will tend to oscillate without much improvement.

Solution: add a momentum factor

$$\delta w \leftarrow \alpha \delta w - \eta \frac{\partial E}{\partial w}$$

$$w \leftarrow w + \delta w$$

Hopefully, this will dampen sideways oscillations but amplify downhill motion by $\frac{1}{1-\alpha}.$

Dropout (7.12)

Nodes are randomly chosen to not be used, with some fixed probability (usually, one half).
Vanishing / Exploding Gradients

Training by backpropagation in networks with many layers is difficult. When the weights are small, the differentials become smaller and smaller as we backpropagate through the layers, and end up having no effect. When the weights are large, the activations in the higher layers will saturate to extreme values. As a result, the gradients at those layers will become very small, and will not be propagated to the earlier layers. When the weights have intermediate values, the differentials will sometimes get multiplied many times in places where the transfer function is steep, causing them to blow up to large values.
**Weight Initialization**

In order to have healthy forward and backward propagation, each term in the product must be approximately equal to 1. Any deviation from this could cause the activations to either vanish or saturate, and the differentials to either decay or explode exponentially.

\[
\text{Var}[z] \approx \left( \prod_{i=1}^{D} G_{n_i}^{\text{in}} \text{Var}[w^{(i)}] \right) \text{Var}[x]
\]

\[
\text{Var}[\frac{\partial}{\partial x}] \approx \left( \prod_{i=1}^{D} G_{1} n_i^{\text{out}} \text{Var}[w^{(i)}] \right) \text{Var}[\frac{\partial}{\partial z}]
\]

We therefore choose the initial weights \( \{w^{(i)}_{jk}\} \) in each layer \( i \) such that

\[
G_1 n_i^{\text{out}} \text{Var}[w^{(i)}] = 1
\]

**Convolutional Network Components**

- **convolution layers**: extract shift-invariant features from the previous layer
- **subsampling or pooling layers**: combine the activations of multiple units from the previous layer into one unit
- **fully connected layers**: collect spatially diffuse information
- **output layer**: choose between classes

**Convolutional Neural Networks**

\[
Z_{j,k}^{i} = g(b^{i} + \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} K_{j,m,n}^{i} V_{j+m,k+n}^{i})
\]

The same weights are applied to the next \( M \times N \) block of inputs, to compute the next hidden unit in the convolution layer (“weight sharing”).

**Batch Normalization**

We can normalize the activations \( x_k^{(i)} \) of node \( k \) in layer \( i \) relative to the mean and variance of those activations, calculated over a mini-batch of training items:

\[
x_k^{(i)} = \frac{x_k^{(i)} - \text{Mean}[x_k^{(i)}]}{\sqrt{\text{Var}[x_k^{(i)}]}}
\]

These activations can then be shifted and re-scaled to

\[
y_k^{(i)} = \beta_k^{(i)} + \gamma_k^{(i)} x_k^{(i)}
\]

\( \beta_k^{(i)}, \gamma_k^{(i)} \) are additional parameters, for each node, which are trained by backpropagation along with the other parameters (weights) in the network.

After training is complete, \( \text{Mean}[x_k^{(i)}] \) and \( \text{Var}[x_k^{(i)}] \) are either pre-computed on the entire training set, or updated using running averages.
**Residual Networks**

Idea: Take any two consecutive stacked layers in a deep network and add a “skip” connection which bypasses these layers and is added to their output.

**Stride with Zero Padding**

When combined with zero padding of width $P$,
- $j$ takes on the values $0, s, 2s, \ldots, (J + 2P - M)$
- $k$ takes on the values $0, s, 2s, \ldots, (K + 2P - N)$

The next layer is $(1 + (J + 2P - M)/s)$ by $(1 + (K + 2P - N)/s)$.

**Convolutional Filters**

First Layer  Second Layer  Third Layer

**Dense Networks**

Recently, good results have been achieved using networks with densely connected blocks, within which each layer is connected by shortcut connections to all the preceding layers.
Recurrent Networks

- Processing Temporal Sequences
- Sliding Window
- Recurrent Network Architectures
- Hidden Unit Dynamics
- Long Short Term Memory

Simple Recurrent Network (Elman, 1990)

- at each time step, hidden layer activations are copied to “context” layer
- hidden layer receives connections from input and context layers
- the inputs are fed one at a time to the network, it uses the context layer to “remember” whatever information is required for it to produce the correct output

Sliding Window

The simplest way to feed temporal input to a neural network is the “sliding window” approach, first used in the NetTalk system (Sejnowski & Rosenberg, 1987).

Back Propagation Through Time

- we can “unroll” a recurrent architecture into an equivalent feedforward architecture, with shared weights
- applying backpropagation to the unrolled architecture is referred to as “backpropagation through time”
- we can backpropagate just one timestep, or a fixed number of timesteps, or all the way back to beginning of the sequence
Oscillating Solution for $a^m b^n$

Long Range Dependencies

- Simple Recurrent Networks (SRNs) can learn medium-range dependencies but have difficulty learning long range dependencies
- Long Short Term Memory (LSTM) and Gated Recurrent Units (GRU) can learn long range dependencies better than SRN

Gated Recurrent Unit

GRU is similar to LSTM but has only two gates instead of three.

Long Short Term Memory

Gates:
- $f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f)$
- $i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$
- $g_t = \tanh(W_g x_t + U_g h_{t-1} + b_g)$
- $o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$

State:
- $c_t = c_{t-1} \odot f_t + i_t \odot g_t$

Output:
- $h_t = \tanh c_t \odot o_t$

Gates:
- $z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$
- $r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$

Candidate Activation:
- $\tilde{h}_t = \tanh(W_{\tilde{h}} x_t + U_{\tilde{h}} r_t \odot h_{t-1} + b_{\tilde{h}})$

Output:
- $h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$
Co-occurrence Matrix (2-word window)

<table>
<thead>
<tr>
<th>word</th>
<th>all</th>
<th>and</th>
<th>bought</th>
<th>cat</th>
<th>caught</th>
<th>crooked</th>
<th>found</th>
<th>he</th>
<th>house</th>
<th>in</th>
<th>little</th>
<th>lived</th>
<th>man</th>
<th>mile</th>
<th>mouse</th>
<th>sixpence</th>
<th>stile</th>
<th>there</th>
<th>they</th>
<th>together</th>
<th>upon</th>
<th>walked</th>
<th>who</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>all</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>and</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>bought</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>cat</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>caught</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>crooked</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>found</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>he</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>house</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>in</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>little</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>lived</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>man</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>mile</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>mouse</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>sixpence</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>stile</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>there</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>they</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>together</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>upon</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>walked</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>who</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Singular Value Decomposition

Co-occurrence matrix X can be decomposed as \( X = USV^T \) where U, V are unitary (all columns have unit length) and S is diagonal.

\[
\begin{align*}
L & = \begin{bmatrix} r \\ u_1 \\ u_2 \\ u_r \\ s_1 \\ s_2 \\ s_r \end{bmatrix}
\end{align*}
\]

Columns 1 to \( n \) of row \( k \) of \( U \) then provide an \( n \)-dimensional vector representing the \( k^{th} \) word in the vocabulary.

SVD is computationally expensive, proportional to \( L \times M^2 \) if \( L \geq M \). Can we do something similar with less computation, and incrementally?

Continuous Bag of Words

If several context words are each used independently to predict the center word, the hidden activation becomes a sum (or average) over all the context words.

Note the difference between this and NetTalk – in word2vec (CBOW) all context words share the same input-to-hidden weights.
**Negative Sampling**

The idea of negative sampling is that we train the network to increase its estimation of the target word $j^*$ and reduce its estimate not of all the words in the vocabulary but just a subset of them $W_{neg}$, drawn from an appropriate distribution.

$$E = -\log \sigma (v_j^{\prime} \mathbf{h}) - \sum_{j \in W_{neg}} \log \sigma (-v_j^{\prime} \mathbf{h})$$

This is a simplified version of Noise Constrastive Estimation (NCE). It is not guaranteed to produce a well-defined probability distribution, but in practice it does produce high-quality word embeddings.
Reinforcement Learning Framework

- An agent interacts with its environment.
- There is a set $S$ of states and a set $A$ of actions.
- At each time step $t$, the agent is in some state $s_t$.
  It must choose an action $a_t$, whereupon it goes into state $s_{t+1} = \delta(s_t, a_t)$ and receives reward $r_t = R(s_t, a_t)$.
- Agent has a policy $\pi : S \rightarrow A$. We aim to find an optimal policy $\pi^*$ which maximizes the cumulative reward.
- In general, $\delta$, $R$, and $\pi$ can be multi-valued, with a random element, in which case we write them as probability distributions
  $$\delta(s_{t+1} = s \mid s_t, a_t) \quad R(r_t = r \mid s_t, a_t) \quad \pi(a_t = a \mid s_t)$$

Models of optimality

- Is a fast nickel worth a slow dime?
  - Finite horizon reward $\sum_{i=0}^{h-1} r_{t+i}$
  - Infinite discounted reward $\sum_{i=0}^{\infty} \gamma^i r_{t+i}, \quad 0 \leq \gamma < 1$
  - Average reward $\lim_{h \to \infty} \frac{1}{h} \sum_{i=0}^{h-1} r_{t+i}$
- Finite horizon reward is simple computationally
- Infinite discounted reward is easier for proving theorems
- Average reward is hard to deal with, because can’t sensibly choose between small reward soon and large reward very far in the future.
Let’s first assume that $R$ and $\delta$ are deterministic. Then the (true) value $V^*(s)$ of the current state $s$ should be equal to the immediate reward plus the discounted value of the next state

$$V^*(s) = R(s,a) + \gamma V^*(\delta(s,a))$$

We can turn this into an update rule for the estimated value, i.e.

$$V(s_t) \leftarrow r_t + \gamma V(s_{t+1})$$

If $R$ and $\delta$ are stochastic (multi-valued), it is not safe to simply replace $V(s)$ with the expression on the right hand side. Instead, we move its value fractionally in this direction, proportional to a learning rate $\eta$

$$V(s_t) \leftarrow V(s_t) + \eta [r_t + \gamma V(s_{t+1}) - V(s_t)]$$

Most of the time we should choose what we think is the best action. However, in order to ensure convergence to the optimal strategy, we must occasionally choose something different from our preferred action, e.g.

- choose a random action 5% of the time, or
- use Softmax (Boltzmann distribution) to choose the next action:

$$P(a) = \frac{e^{R(a)/T}}{\sum_{b \in A} e^{R(b)/T}}$$

For a deterministic environment, $\pi^*, Q^*$ and $V^*$ are related by

$$\pi^*(s) = \arg\max_a Q^*(s,a)$$

$$Q^*(s,a) = R(s,a) + \gamma V^*(\delta(s,a))$$

$$V^*(s) = \max_b Q^*(s,b)$$

So

$$Q^*(s,a) = R(s,a) + \gamma \max_b Q^*(\delta(s,a),b)$$

This allows us to iteratively approximate $Q$ by

$$Q(s_t,a_t) \leftarrow r_t + \gamma \max_b Q(s_{t+1},b)$$

If the environment is stochastic, we instead write

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \eta [r_t + \gamma \max_b Q(s_{t+1},b) - Q(s_t,a_t)]$$
**Policy Gradients**

If $r_{\text{total}} = +1$ for a win and $-1$ for a loss, we can simply multiply the log probability by $r_{\text{total}}$. Differentials can be calculated using the gradient

$$\nabla_\theta r_{\text{total}} \sum_{t=1}^{m} \log \pi_\theta (a_t | s_t) = r_{\text{total}} \sum_{t=1}^{m} \nabla_\theta \log \pi_\theta (a_t | s_t)$$

The gradient of the log probability can be calculated nicely using Softmax. If $r_{\text{total}}$ takes some other range of values, we can replace it with $(r_{\text{total}} - b)$ where $b$ is a fixed value, called the **baseline**.

**REINFORCE Algorithm**

We then get the following REINFORCE algorithm:

```
for each trial
    run trial and collect states $s_t$, actions $a_t$, and reward $r_{\text{total}}$
    for $t = 1$ to length(trial)
        $\theta \leftarrow \theta + \eta (r_{\text{total}} - b) \nabla_\theta \log \pi_\theta (a_t | s_t)$
    end
end
```

This algorithm has successfully been applied, for example, to learn to play the game of Pong from raw image pixels.

**Deep Q-Network**

![Deep Q-Network Diagram]

- choose actions using current Q function ($\varepsilon$-greedy)
- build a database of experiences $(s_t, a_t, r_t, s_{t+1})$
- sample asynchronously from database and apply update, to minimize
  $$\left[ r_t + \gamma \max_b Q_w (s_{t+1}, b) - Q_w (s_t, a_t) \right]^2$$
- removes temporal correlations by sampling from variety of game situations in random order
- makes it easier to parallelize the algorithm on multiple GPUs

**Deep Q-Learning with Experience Replay**

- build a database of experiences $(s_t, a_t, r_t, s_{t+1})$
- sample asynchronously from database and apply update, to minimize
  $$\left[ r_t + \gamma \max_b Q_w (s_{t+1}, b) - Q_w (s_t, a_t) \right]^2$$
- removes temporal correlations by sampling from variety of game situations in random order
- makes it easier to parallelize the algorithm on multiple GPUs
Double Q-Learning

- if the same weights $w$ are used to select actions and evaluate actions, this can lead to a kind of confirmation bias
- could maintain two sets of weights $w$ and $\overline{w}$, with one used for selection and the other for evaluation (then swap their roles)
- in the context of Deep Q-Learning, a simpler approach is to use the current “online” version of $w$ for selection, and an older “target” version $\overline{w}$ for evaluation; we therefore minimize
\[
[r_t + \gamma Q_w(s_{t+1}, \text{argmax}_b Q_w(s_{t+1}, b)) - Q_w(s_t, a_t)]^2
\]
- a new version of $\overline{w}$ is periodically calculated from the distributed values of $w$, and this $\overline{w}$ is broadcast to all processors.

Advantage Actor Critic

Recall that in the REINFORCE algorithm, a baseline $b$ could be subtracted from $r_{total}$
\[
\theta \leftarrow \theta + \eta (r_{total} - b) \nabla \log \pi_\theta (a_t | s_t)
\]
In the actor-critic framework, $r_{total}$ is replaced by $Q(s_t, a_t)$
\[
\theta \leftarrow \theta + \eta Q(s_t, a_t) \nabla \log \pi_\theta (a_t | s_t)
\]
We can also subtract a baseline from $Q(s_t, a_t)$. This baseline must be independent of the action $a_t$, but it could be dependent on the state $s_t$.
A good choice of baseline is the value function $V_u(s)$, in which case the $Q$ function is replaced by the advantage function
\[
A_w(s, a) = Q(s, a) - V_u(s)
\]

Asynchronous Advantage Actor Critic

- use policy network to choose actions
- learn a parameterized Value function $V_u(s)$ by TD-Learning
- estimate Q-value by n-step sample
\[
Q(s_t, a_t) = r_{t+1} + \gamma r_{t+2} + \ldots + \gamma^{n-1} r_{t+n} + \gamma^n V_u(s_{t+n})
\]
- update policy by
\[
\theta \leftarrow \theta + \eta \theta [Q(s_t, a_t) - V_u(s_t)] \nabla \log \pi_\theta (a_t | s_t)
\]
- update Value function my minimizing
\[
[Q(s_t, a_t) - V_u(s_t)]^2
\]

Hopfield Network

\[
E(x) = -\left(\frac{1}{2} \sum_{i,j} w_{ij} x_i x_j + \sum_i b_i x_i\right)
\]
Start with an initial state $x$ and then repeatedly try to “flip” neuron activations one at a time, in order to reach a lower-energy state. If we choose to modify neuron $x_i$, its new value should be
\[
x_i \leftarrow \begin{cases} 
+1, & \text{if } \sum_j w_{ij} x_j + b_i > 0, \\
 x_i, & \text{if } \sum_j w_{ij} x_j + b_i = 0, \\
-1, & \text{if } \sum_j w_{ij} x_j + b_i < 0. 
\end{cases}
\]
This ensures that the energy $E(x)$ will never increase. It will eventually reach a local minimum.
Boltzmann Machine (20.1)

The Boltzmann Machine uses exactly the same energy function as the Hopfield network:

$$E(x) = - \left( \sum_{i < j} x_i w_{ij} x_j + \sum b_i x_i \right)$$

The Boltzmann Machine is very similar to the Hopfield Network, except that
- components (neurons) $x_i$ take on the values 0, 1 instead of $-1, +1$
- used to generate new states rather than retrieving stored states
- update is not deterministic but stochastic, using the sigmoid

Boltzmann Machine

The Boltzmann Machine operates similarly to a Hopfield Network, except that there is some randomness in the neuron updates.

In both cases, we repeatedly choose one neuron $x_i$ and decide whether or not to “flip” the value of $x_i$, thus changing from state $x$ into $x'$.

For the Hopfield Network, we do not change from $x$ to $x'$ unless $\Delta E \leq 0$, i.e. we never move to a higher energy state. For the Boltzmann machine, we instead choose $x_i = 1$ with probability

$$p = \frac{1}{1 + e^{-\Delta E / T}}$$

In other words, there is some probability of moving to a higher energy state (or remaining in a higher energy state when a lower one is available).

Restricted Boltzmann Machine (16.7)

If we allow visible-to-visible and hidden-to-hidden connections, the network takes too long to train. So we normally restrict the model by allowing only visible-to-hidden connections.

This is known as a Restricted Boltzmann Machine.
Autoencoder Networks

- output is trained to reproduce the input as closely as possible
- activations normally pass through a bottleneck, so the network is forced to compress the data in some way
- like the RBM, Autoencoders can be used to automatically extract abstract features from the input

Quick Contrastive Divergence

It was noticed in the early 2000's that the process can be sped up by taking just one additional sample instead of running for many iterations.

- \( v_0, h_0 \) are used as positive sample, and \( v_1, h_1 \) as negative sample
- this can be compared to the Negative Sampling that was used with word2vec – it is not guaranteed to approximate the true gradient, but it works well in practice

Regularized Autoencoders (14.2)

- sparse autoencoders
- autoencoders with dropout at hidden layer(s)
- contractive autoencoders
- denoising autoencoders
Entropy and KL-Divergence

- The entropy of a distribution $q(\theta)$ is
  \[ H(q) = \int q(\theta)(-\log q(\theta))d\theta \]
- In Information Theory, $H(q)$ is the amount of information (bits) required to transmit a random sample from distribution $q(\theta)$.
- For a Gaussian distribution, $H(q) = \sum \log \sigma_i$.
- KL-Divergence $D_{KL}(q \parallel p) = \int q(\theta)(\log q(\theta) - \log p(\theta))d\theta$.
- $D_{KL}(q \parallel p)$ is the number of extra bits we need to transmit if we designed a code for $p(\theta)$, but the samples are drawn from $q(\theta)$ instead.
- If $p(z)$ is Standard Normal distribution, minimizing $D_{KL}(q_\phi(z) \parallel p(z))$ encourages $q_\phi(\theta)$ to center on zero and spread out to approximate $p(\theta)$.

Generative Models

- Sometimes, as well as reproducing the training items $\{x^{(i)}\}$, we also want to be able to use the decoder to generate new items which are of a similar “style” to the training items.
- In other words, we want to be able to choose latent variables $z$ from a standard Normal distribution $p(z)$, feed these values of $z$ to the decoder, and have it produce a new item $x$ which is somehow similar to the training items.
- Generative models can be:
  - explicit (Variational Autoencoders)
  - implicit (Generative Adversarial Networks)

Gaussian Distribution (3.9.3)

\[ p_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]

\[ \mu = \text{mean} \]
\[ \sigma = \text{standard deviation} \]

Multivariate Gaussian:
\[ p_{\mu,\sigma}(x) = \prod_i p_{\mu_i,\sigma_i}(x_i) \]

Variational Autoencoder (20.10.3)

Instead of producing a single $z$ for each $x^{(i)}$, the encoder (with parameters $\phi$) can be made to produce a mean $\mu_{\phi(x^{(i)})}$ and standard deviation $\sigma_{\phi(x^{(i)})}$:

This defines a conditional (Gaussian) probability distribution $q_\phi(z|x^{(i)})$.

We then train the system to maximize
\[ \mathbb{E}_{z \sim q_\phi(z|x^{(i)})}[\log p_\theta(x^{(i)}|z)] \quad - D_{KL}(q_\phi(z|x^{(i)}) \parallel p(z)) \]

- the first term enforces that any sample $z$ drawn from the conditional distribution $q_\phi(z|x^{(i)})$ should, when fed to the decoder, produce something approximating $x^{(i)}$.
- the second term encourages $q_\phi(z|x^{(i)})$ to approximate $p(z)$.
- in practice, the distributions $q_\phi(z|x^{(i)})$ for various $x^{(i)}$ will occupy complementary regions within the overall distribution $p(z)$.
Generative Adversarial Networks

Alternate between:

Gradient ascent on Discriminator:
\[ \max_\psi \left( \mathbb{E}_{x \sim p_{\text{data}}} \left[ \log D_\psi(x) \right] + \mathbb{E}_{z \sim p_{\text{model}}} \left[ \log \left(1 - D_\psi(G_\theta(z))\right) \right] \right) \]

Gradient descent on Generator, using:
\[ \min_\theta \mathbb{E}_{z \sim p_{\text{model}}} \left[ \log \left(1 - D_\psi(G_\theta(z))\right) \right] \]

This formula puts too much emphasis on images that are correctly classified. Better to do gradient ascent on Generator, using:
\[ \max_\theta \mathbb{E}_{z \sim p_{\text{model}}} \left[ \log \left(D_\psi(G_\theta(z))\right)\right] \]

This puts more emphasis on the images that are wrongly classified.
Sample 1-mark Question

Which of these architectures would have the best chance of learning long range dependencies?

(a) Feedforward network with sliding window
(b) Simple Recurrent Network
(c) Elman Network
(d) Long Short Term Memory

Sample 2-mark Question

One bag contains 2 red balls and 3 white balls. Another bag contains 3 red balls and 2 green balls. One of these bags is chosen at random, and two balls are drawn randomly from that bag, without replacement. Both of the balls turn out to be red. What is the probability that the first bag is the one that was chosen?

(a) 1/4
(b) 1/3
(c) 1/2
(d) 2/3
UNSW myExperience Survey

Please remember to fill in the UNSW myExperience Survey.

Neural Networks and Deep Learning

GOOD LUCK!