Exercise 4.1: Convolutional Neural Networks
(source: Geoffrey Hinton, Coursera)

A convolutional neural network is a type of MLP well suited to image recognition. It combines weight sharing with an optimized network topology, that can exploit the 2d-structure of the input data.

(a) We have a convolutional neural network for images of 5 by 5 pixels. In this network, each hidden unit is connected to a different $4 \times 4$ region of the input image:

- The first hidden unit, $h_1$, is connected to the upper left $4 \times 4$ portion of the input image (as shown below).
- The second hidden unit, $h_2$, is connected to the upper right $4 \times 4$ portion of the input image (as shown below).
- The third hidden unit, $h_3$, is connected to the lower left $4 \times 4$ portion of the input image (not shown in the diagram).
- The fourth hidden unit, $h_4$, is connected to the lower right $4 \times 4$ portion of the input image (not shown in the diagram).
Because it’s a convolutional network, the weights (connection strengths) are the same for all hidden units: the only difference between the hidden units is that each of them connects to a different part of the input image. In the second diagram, we show the array of weights, which are the same for each of the four hidden units.

For \( h_1 \), weight \( w_{11} \) is connected to the top-left pixel, i.e. \( x_{11} \), while for hidden unit \( h_2 \), weight \( w_{11} \) connects to the pixel that is one to the right of the top left pixel, i.e. \( x_{12} \). Imagine that for some training case, we have an input image where each of the black pixels in the top diagram has value 1, and each of the white ones has value 0. Notice that the image shows a ”3” in pixels.
The network has no biases. The hidden units are linear. The weights of the network are given as follows:

\[
\begin{align*}
    w_{11} &= 1 & w_{12} &= 1 & w_{13} &= 1 & w_{14} &= 0 \\
    w_{21} &= 0 & w_{22} &= 0 & w_{23} &= 1 & w_{24} &= 0 \\
    w_{31} &= 1 & w_{32} &= 1 & w_{33} &= 1 & w_{34} &= 0 \\
    w_{41} &= 0 & w_{42} &= 0 & w_{43} &= 1 & w_{44} &= 0
\end{align*}
\]

For the training case with that "3" input image, what is the output \( y_1, y_2, y_3, y_4 \) of each of the four hidden units?

\[
\begin{align*}
    y_1 &= w_{11} \cdot x_{11} + w_{21} \cdot x_{21} + \cdots + w_{44} \cdot x_{44} = 4 \\
    y_2 &= w_{11} \cdot x_{12} + w_{21} \cdot x_{22} + \cdots + w_{44} \cdot x_{45} = 8 \\
    y_3 &= w_{11} \cdot x_{21} + w_{21} \cdot x_{31} + \cdots + w_{44} \cdot x_{54} = 2 \\
    y_4 &= w_{11} \cdot x_{22} + w_{21} \cdot x_{32} + \cdots + w_{44} \cdot x_{55} = 4
\end{align*}
\]
(b) Let’s look at what weight sharing does to gradients, computed using the backpropagation algorithm. For this question, our network has three input units, $x_1, x_2, x_3$, two logistic hidden units $h_1, h_2$, four input to hidden weights $w_1, w_2, w_3, w_4$, and two hidden to output weights $u_1, u_2$. The output neuron $y$ is a linear neuron, and we are using the squared error cost function.

Suppose we now decide to use weight sharing for $w_2$ and $w_3$. What is the derivative of the error $E$ with respect to the shared weight $w_{\text{shared}}$?
Hint: Compute the partial derivatives $\frac{\partial E}{\partial w_2}$ and $\frac{\partial E}{\partial w_3}$ first.

Forward pass

$$net_1 = w_1x_1 + w_2x_2$$
$$net_2 = w_3x_2 + w_4x_3$$
$$h_1 = sig(net_1)$$
$$h_2 = sig(net_2)$$
$$y = u_1h_1 + u_2h_2$$
$$e = \frac{1}{2}(y - d)^2$$
Backward pass

\[
\frac{\partial e}{\partial y} = y - d \\
\frac{\partial e}{\partial h_1} = \frac{\partial e}{\partial y} \frac{\partial y}{\partial h_1} = (y - d)u_1 \\
\frac{\partial e}{\partial h_2} = \frac{\partial e}{\partial y} \frac{\partial y}{\partial h_2} = (y - d)u_2 \\
\frac{\partial e}{\partial \text{net}_1} = \frac{\partial e}{\partial h_1} \frac{\partial h_1}{\partial \text{net}_1} = (y - d)u_1(1 - h_1)h_1 \\
\frac{\partial e}{\partial \text{net}_2} = \frac{\partial e}{\partial h_2} \frac{\partial h_2}{\partial \text{net}_2} = (y - d)u_2(1 - h_2)h_2 \\
\frac{\partial e}{\partial w_2} = \frac{\partial e}{\partial \text{net}_1} \frac{\partial \text{net}_1}{\partial w_2} = (y - d)u_1(1 - h_1)h_1x_2 \\
\frac{\partial e}{\partial w_3} = \frac{\partial e}{\partial \text{net}_2} \frac{\partial \text{net}_2}{\partial w_3} = (y - d)u_2(1 - h_2)h_2x_2
\]
Rule of sums:

\[
\frac{\partial e}{\partial w_{\text{tied}}} = \frac{\partial e}{\partial w_2} + \frac{\partial e}{\partial w_3} = (y - d)x_2(u_1(1 - h_1)h_1 + u_2(1 - h_2)h_2)
\]
Exercise 4.2: Boosting with Decision Stumps

Apply the AdaBoost algorithm to train a classifier for the dataset specified in Table 1. Consider four decision stumps $S_N$, $S_C$, $S_R$, and $S_F$ – one for each attribute – that classify different instances as positive and negative. So, for example, the decision stump belonging to the “coughing” attribute classifies the patterns as

$$S_C(d_i) = true \text{ for } i \in \{1, 2, 6\} \text{ and } S_C(d_i) = false \text{ for } i \in \{3, 4, 5\}.$$

(a) Apply $T = 4$ iterations of the AdaBoost algorithm to the training patterns provided. Select in each iteration that decision stump that yields the lowest error on the reweighted pattern distribution.
Exercise 5.2: Boosting with Decision Stumps

Apply the AdaBoost algorithm to train a classifier for the dataset specified in Table 2. Consider four decision stumps $S_N$, $S_C$, $S_R$, and $S_F$—one for each attribute—classify different instances as positive and negative. So, for example, the decision stump belonging to the "coughing" attribute classifies the patterns as $S_C(d_i) = \text{true}$ for $i \in \{1, 2, 6\}$ and $S_C(d_i) = \text{false}$ for $i \in \{3, 4, 5\}$.

(a) Apply $T = 4$ iterations of the AdaBoost algorithm to the training patterns provided. Select in each iteration that decision stump that yields the lowest error on the reweighted pattern distribution.

(b) Verify whether your final classifier $H_{\text{final}}$ correctly classifies all training patterns.

Table 1: List of training instances.

<table>
<thead>
<tr>
<th>Training Example</th>
<th>N</th>
<th>C</th>
<th>R</th>
<th>F</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(running nose)</td>
<td>(coughing)</td>
<td>(reddened skin)</td>
<td>(fever)</td>
<td></td>
</tr>
<tr>
<td>$d_1$</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>–</td>
<td>positive (ill)</td>
</tr>
<tr>
<td>$d_2$</td>
<td>+</td>
<td>+</td>
<td>–</td>
<td>–</td>
<td>positive (ill)</td>
</tr>
<tr>
<td>$d_3$</td>
<td>–</td>
<td>–</td>
<td>+</td>
<td>+</td>
<td>positive (ill)</td>
</tr>
<tr>
<td>$d_4$</td>
<td>+</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>negative (healthy)</td>
</tr>
<tr>
<td>$d_5$</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>negative (healthy)</td>
</tr>
<tr>
<td>$d_6$</td>
<td>–</td>
<td>+</td>
<td>+</td>
<td>–</td>
<td>negative (healthy)</td>
</tr>
</tbody>
</table>

Table 2: List of training instances.

\[ t = 1: \text{expected error} \ E_N = E_C = E_R = E_F = \frac{4}{3} \quad \rightarrow \text{choose attribute } N \]
\[ \text{hypothesis weight} \ \alpha_1 = 0.3466 \]
\[ \text{normalized pattern distribution} \ D^{(1)} = \left( \frac{4}{8}, \frac{4}{8}, \frac{2}{8}, \frac{4}{8}, \frac{4}{8} \right) \]
\[ t = 2: \ E_N = \frac{4}{2}, \ E_C = \frac{3}{8}, \ E_R = \frac{4}{4}, \ E_F = \frac{4}{4} \quad \rightarrow \text{choose attribute } R \]
\[ \alpha_2 = 0.5493, \ D^{(2)} = \left( \frac{4}{4}, \frac{4}{4}, \frac{4}{4}, \frac{4}{4}, \frac{4}{4} \right) \]
\[ t = 3: \ E_N = \frac{4}{3}, \ E_C = \frac{5}{8}, \ E_R = \frac{4}{4}, \ E_F = \frac{4}{4} \quad \rightarrow \text{choose } N, \ \alpha_3 = 0.3466, \ D^{(3)} = \left( \frac{4}{4}, \frac{3}{4}, \frac{4}{4}, \frac{4}{4}, \frac{4}{4} \right) \]
\[ t = 4: \ E_N = \frac{4}{2}, \ E_C = \frac{7}{16}, \ E_R = \frac{3}{8}, \ E_F = \frac{4}{4} \quad \rightarrow \text{choose } F, \ \alpha_4 = 0.5493, \ D^{(4)} = \left( \frac{4}{8}, \frac{3}{16}, \frac{4}{16}, \frac{4}{16}, \frac{4}{16} \right) \]
(b) Verify whether your final classifier $H_{final}$ correctly classifies all training patterns.

$$H(x) = \text{sign}\left(\sum_{\ell=1}^{M} \alpha_{\ell} h_{\ell}(x)\right)$$

$H(x_1) = \text{sign}(0.6931) = 1$

$H(x_2) = \text{sign}(-0.4055) = -1$

$H(x_3) = \text{sign}(0.4055) = 1$

$H(x_4) = \text{sign}(-0.4055) = -1$

$H(x_5) = \text{sign}(-1.7918) = -1$

$H(x_6) = \text{sign}(-0.6931) = -1$
Exercise 4.3: Probabilities

We consider a medical diagnosis task. We have knowledge that over the entire population of people 0.8% have cancer. There exists a (binary) laboratory test that represents an imperfect indicator of this disease. That test returns a correct positive result in 98% of the cases in which the disease is present, and a correct negative results in 97% of the cases where the disease is not present.

(a) Suppose we observe a patient for whom the laboratory test returns a positive result. Calculate the a posteriori probability that this patient truly suffers from cancer.

Pr(cancer) = 0.008
Pr(¬cancer) = 0.992
Pr(⊕|cancer) = 0.98
Pr(⊕|¬cancer) = 0.03

Pr(⊕) = Pr(⊕|cancer) Pr(cancer) + Pr(⊕|¬cancer) Pr(¬cancer) = 0.0376
Bayes rule:

\[ \Pr(cancer|\oplus) = \frac{\Pr(\oplus|cancer) \Pr(cancer)}{\Pr(\oplus)} = 0.209 \]
(b) Knowing that the lab test is an imperfect one, a second test (which is assumed to be independent of the former one) is conducted. Calculate the a posteriori probabilities for cancer and ¬cancer given that the second test has returned a positive result as well.

Bayes rule:

\[
\Pr(\oplus, \oplus) = \Pr(\oplus \oplus | cancer) \Pr(cancer) + \Pr(\oplus \oplus | \neg cancer) \Pr(\neg cancer) \\
= \Pr(\oplus | cancer) \Pr(\oplus | cancer) \Pr(cancer) \\
+ \Pr(\oplus | \neg cancer) \Pr(\oplus | \neg cancer) \Pr(\neg cancer) \\
= 0.008576
\]

\[
\Pr(cancer | \oplus, \oplus) = \frac{\Pr(\oplus, \oplus | cancer) \Pr(cancer)}{\Pr(\oplus, \oplus)} \\
= \frac{\Pr(\oplus | cancer) \Pr(\oplus | cancer) \Pr(cancer)}{\Pr(\oplus, \oplus)} \\
= 0.896
\]
Exercise 4.4: Naïve Bayes Classifier

In the following, we consider the data set introduced in Table 2 where the task is to describe whether a person is *ill*. We use a representation based on four features per subject to describe an individual person. These features are “running nose”, “coughing”, and “reddened skin”, each of which can take the value true (‘+) or false (‘–’).

<table>
<thead>
<tr>
<th>Training Example</th>
<th>N (running nose)</th>
<th>C (coughing)</th>
<th>R (reddened skin)</th>
<th>F (fever)</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>(d_1)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>–</td>
<td>positive (ill)</td>
</tr>
<tr>
<td>(d_2)</td>
<td>+</td>
<td>+</td>
<td>–</td>
<td>–</td>
<td>positive (ill)</td>
</tr>
<tr>
<td>(d_3)</td>
<td>–</td>
<td>–</td>
<td>+</td>
<td>+</td>
<td>positive (ill)</td>
</tr>
<tr>
<td>(d_4)</td>
<td>+</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>negative (healthy)</td>
</tr>
<tr>
<td>(d_5)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>negative (healthy)</td>
</tr>
<tr>
<td>(d_6)</td>
<td>–</td>
<td>+</td>
<td>+</td>
<td>–</td>
<td>negative (healthy)</td>
</tr>
</tbody>
</table>

Table 2: Training examples
\[
\Pr(a_1, \ldots, a_n | v_j) = \prod_{i=1}^{n} \Pr(a_i | v_j)
\]
\[
v_{NB} = \arg \max_{v_j \in V} p(v_j) \prod_{i=1}^{n} \Pr(a_i | v_j)
\]
\[
\Pr(a_i | v_j) = \frac{n_c + mp}{n + m}
\]
(a) Given the data set in Table 2, determine all probabilities required to apply the naïve Bayes classifier for predicting whether a new person is ill or not. Use the \( m \)-estimate of probability with an equivalent sample size \( m = 4 \) and a uniform prior \( p \).

\[
\begin{align*}
\Pr(N|\text{ill}) &= \frac{2 + 0.5 \cdot 4}{3 + 4} = \frac{4}{7} \\
\Pr(C|\text{ill}) &= \frac{2 + 0.5 \cdot 4}{3 + 4} = \frac{4}{7} \\
\Pr(R|\text{ill}) &= \frac{2 + 0.5 \cdot 4}{3 + 4} = \frac{4}{7} \\
\Pr(F|\text{ill}) &= \frac{1 + 0.5 \cdot 4}{3 + 4} = \frac{3}{7} \\
\Pr(\text{ill}) &= \frac{3}{6} \\
\Pr(N|\text{healthy}) &= \frac{1 + 0.5 \cdot 4}{3 + 4} = \frac{3}{7} \\
\Pr(C|\text{healthy}) &= \frac{1 + 0.5 \cdot 4}{3 + 4} = \frac{3}{7} \\
\Pr(R|\text{healthy}) &= \frac{1 + 0.5 \cdot 4}{3 + 4} = \frac{3}{7} \\
\Pr(F|\text{healthy}) &= \frac{0 + 0.5 \cdot 4}{3 + 4} = \frac{2}{7} \\
\Pr(\text{healthy}) &= \frac{3}{6}
\end{align*}
\]

(b) Verify whether the naïve Bayes classifier classifies all training examples
(\(d_1, \ldots, d_6\)) correctly.

\[
d_1: \Pr(N, C, R, \overline{F}|ill) = \frac{4}{7} \cdot \frac{4}{7} \cdot \frac{4}{7} \cdot \frac{1}{7} = 0.107
\]

\[
\Pr(N, C, R, \overline{F}|healthy) = \frac{3}{7} \cdot \frac{3}{7} \cdot \frac{3}{7} \cdot \frac{3}{7} = 0.056
\]

\[
v_{NB} = \arg \max_{v_j \in V} \Pr(v_j) \Pr(N, C, R, \overline{F}|v_j) = \text{ill}
\]

\[
d_2: \Pr(N, C, \overline{R}, \overline{F}|ill) = \frac{4}{7} \cdot \frac{4}{7} \cdot \frac{3}{7} \cdot \frac{4}{7} = 0.080
\]

\[
\Pr(N, C, \overline{R}, \overline{F}|healthy) = \frac{3}{7} \cdot \frac{3}{7} \cdot \frac{4}{7} \cdot \frac{3}{7} = 0.075
\]

\[
v_{NB} = \text{ill}
\]

\[
d_3: \Pr(\overline{N}, \overline{C}, R, F|ill) = \frac{3}{7} \cdot \frac{3}{7} \cdot \frac{4}{7} \cdot \frac{3}{7} = 0.045
\]

\[
\Pr(\overline{N}, \overline{C}, R, F|healthy) = 0.040
\]

\[
v_{NB} = \text{ill}
\]

\[
d_4: v_{NB} = \text{healthy}
\]

\[
d_5: v_{NB} = \text{healthy}
\]

\[
d_6: v_{NB} = \text{ill} \quad \Rightarrow \text{wrongly classified}
\]
(c) Apply your naïve Bayes classifier to the test patterns corresponding to the following subjects: a person who is coughing and has fever, a person whose nose is running and who suffers from fever, and a person with a running nose and reddened skin ($d_7 = (\overline{N}, C, \overline{R}, F)$, $d_8 = (N, \overline{C}, \overline{R}, F)$, and $d_9 = (N, C, R, \overline{F})$).

\[
d_7: \Pr(\overline{N}, C, \overline{R}, F|\text{ill}) = \frac{3}{7} \cdot \frac{4}{7} \cdot \frac{3}{7} \cdot \frac{3}{7} = 0.045
\]

\[
\Pr(\overline{N}, C, \overline{R}, F|\text{healthy}) = \frac{4}{7} \cdot \frac{3}{7} \cdot \frac{4}{7} \cdot \frac{2}{7} = 0.040
\]

$v_{NB} = \text{ill}$

$d_8 : v_{NB} = \text{ill}$

$d_9 : v_{NB} = \text{ill}$
(d) Now, we no longer distinguish between positive and negative training examples, but each instance is assigned one out of $k$ classes. The corresponding training data is provided in Table 3. Calculate all probabilities required for the application of a naïve Bayes classifier, i.e. $P(v)$ and $P((a_F, a_V, a_D, a_{Sh})|v)$ for $v \in \{H, I, S, B\}$ and $a_F \in \{no, average, high\}$ and $a_V, a_D, a_{Sh} \in \{yes, no\}$. Again, use the $m$-estimate of probability method with $m = 6$ and $p$ uniform.

<table>
<thead>
<tr>
<th>Training</th>
<th>Fever</th>
<th>Vomiting</th>
<th>Diarrhea</th>
<th>Shivering</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_1$</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>healthy (H)</td>
</tr>
<tr>
<td>$d_2$</td>
<td>average</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>influenza (I)</td>
</tr>
<tr>
<td>$d_3$</td>
<td>high</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>influenza (I)</td>
</tr>
<tr>
<td>$d_4$</td>
<td>high</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>salmonella poisoning (S)</td>
</tr>
<tr>
<td>$d_5$</td>
<td>average</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>salmonella poisoning (S)</td>
</tr>
<tr>
<td>$d_6$</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>bowel inflammation (B)</td>
</tr>
<tr>
<td>$d_7$</td>
<td>average</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>bowel inflammation (B)</td>
</tr>
</tbody>
</table>

Table 3: Multi-class training examples.
\[
\begin{align*}
\Pr(V|H) &= \frac{0+0.5 \cdot 6}{1+6} = \frac{3}{7} \\
\Pr(V|I) &= \frac{3}{8} \\
\Pr(V|S) &= \frac{4}{8} \\
\Pr(V|B) &= \frac{5}{8} \\
\Pr(F_{\text{no}}|H) &= \frac{1+\frac{1}{3} \cdot 6}{1+6} = \frac{3}{7} \\
\Pr(F_{\text{no}}|I) &= \frac{0+\frac{1}{3} \cdot 6}{2+6} = \frac{2}{8} \\
\Pr(F_{\text{no}}|S) &= \frac{2}{8} \\
\Pr(F_{\text{no}}|B) &= \frac{3}{8} \\
\Pr(Sh|H) &= \frac{3}{7} \\
\Pr(Sh|I) &= \frac{4}{3} \\
\Pr(Sh|S) &= \frac{5}{3} \\
\Pr(Sh|B) &= \frac{6}{3} \\
\Pr(D|H) &= \frac{3}{7} \\
\Pr(D|I) &= \frac{3}{8} \\
\Pr(D|S) &= \frac{3}{8} \\
\Pr(D|B) &= \frac{3}{8} \\
\Pr(F_{\text{avg}}|H) &= \frac{2}{7} \\
\Pr(F_{\text{avg}}|I) &= \frac{3}{7} \\
\Pr(F_{\text{avg}}|S) &= \frac{3}{7} \\
\Pr(F_{\text{avg}}|B) &= \frac{3}{7} \\
\Pr(F_{\text{high}}|H) &= \frac{2}{7} \\
\Pr(F_{\text{high}}|I) &= \frac{3}{7} \\
\Pr(F_{\text{high}}|S) &= \frac{3}{7} \\
\Pr(F_{\text{high}}|B) &= \frac{3}{7}
\end{align*}
\]
(e) Apply your recently constructed naïve Bayes classifier to a person with high fever, i.e. to \((high, no, no, no)\), as well as to a person who suffers from vomiting and shivering, i.e. to \((no, yes, no, yes)\).

\[
\begin{align*}
\Pr(H) \Pr(F_{high}, V, \overline{D}, \overline{Sh}|H) &= 0.0076 \\
\Pr(I) \Pr(F_{high}, V, \overline{D}, \overline{Sh}|I) &= 0.0209 \\
\Pr(S) \Pr(F_{high}, V, \overline{D}, \overline{Sh}|S) &= 0.0126 \\
\Pr(B) \Pr(F_{high}, V, \overline{D}, \overline{Sh}|B) &= 0.0063
\end{align*}
\]

\[
\begin{align*}
\Pr(H) \Pr(F_{no}, V, \overline{D}, Sh|H) &= \frac{1}{7} \cdot \frac{3}{7} \cdot \frac{3}{7} \cdot \frac{4}{7} \cdot \frac{3}{7} = 0.0064 \\
\Pr(I) \Pr(F_{no}, V, \overline{D}, Sh|I) &= \frac{2}{7} \cdot \frac{2}{8} \cdot \frac{3}{8} \cdot \frac{5}{8} \cdot \frac{4}{8} = 0.0084 \\
\Pr(S) \Pr(F_{no}, V, \overline{D}, Sh|S) &= \frac{2}{7} \cdot \frac{2}{8} \cdot \frac{4}{8} \cdot \frac{3}{8} \cdot \frac{3}{8} = 0.0050 \\
\Pr(B) \Pr(F_{no}, V, \overline{D}, Sh|B) &= \frac{2}{7} \cdot \frac{3}{8} \cdot \frac{5}{8} \cdot \frac{3}{8} \cdot \frac{3}{8} = 0.0094
\end{align*}
\]