Deep Learning Lab Course Assignment 1

Assignment 1: Neural Network

Hendrik Vloet

November 9, 2017

1 Brief Summary of MLPs in General

A Deep Feedforward Network or Multilayer Perceptron is a function approximator. This approximation can be used for various tasks within Machine Learning, e.g. classification. This means we have a Dataset $\mathcal{D}$, consisting of some kind of measurement data $\mathbf{X}$ which has the dimensions $N \times D$ and in case of supervised learning some label vector $\mathbf{Y}$ which is D-dimensional.
The classifier MLP now maps (with the approximation function f and some parameters $\mathbf{\Theta}$) one input data vector $\mathbf{x}$ (which is of $1 \times D$ dimensions large) to his corresponding label y. Formalized it is as in the following:

$$\mathbf{y} = f(\mathbf{x}; \mathbf{\Theta})$$

In case of a network, this corresponds to the nesting of several functions within each other that finally result in the best approximation of the classification. The nested functions are chained together and organized as so called layers. For example, we can use three layers, all using different functions which get chained up:

$$f(\mathbf{x}) = f^{(3)}(f^{(2)}(f^{(1)}(\mathbf{x})))$$

1.1 Activation Functions and Derivatives

The above mentioned functions are also called activation functions and can be of various forms, depending on the task at hand. The are used to keep up certain properties of the data or even emphasize properties like non-linearities. The use of activation during the forward pass and during the backpropagation shall be illustrated with a short example of a 2-Layer MLP.

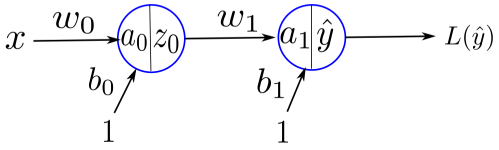


Figure 1.1: Example of a 2-layer MLP, taken for Machine Learning Lecture in Summer term 2017

1.2 Examples of Activation Functions and Derivatives

As mentioned above, activation functions can be of various forms. Below is a small list of commonly used functions and their derivatives:

$$
\begin{aligned}
\textbf{Linear:}\quad & f(\mathbf{x}) = \mathbf{x} \\
& \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \mathbf{1} \\
\textbf{ReLU:}\quad & f(\mathbf{x}) = \max(0, x) \\
& \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{cases} \mathbf{0} & \text{for } x < 0 \\ \mathbf{1} & \text{for } x \geq 0 \end{cases} \\
\textbf{Sigmoid:}\quad & f(\mathbf{x}) = \text{sigmoid}(\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{x})} \\
& \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \text{sigmoid}(\mathbf{x}) \cdot (1 - \text{sigmoid}(\mathbf{x})) \\
\textbf{Tanh:}\quad & f(\mathbf{x}) = \tanh(\mathbf{x}) = \frac{2}{1 + \exp(-2\mathbf{x})} - 1 \\
& \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = 1 - \tanh(\mathbf{x})^2
\end{aligned}
$$

1.3 Forward Pass

The forward pass goes from input to the output of the network and therefore computes the prediction of some input point x, combines this data with some weights w and bias b to get a (hopefully) correct prediction $\hat{y}$. Here the output of a layer is called z and the activation function is $h(\cdot)$ Formalized:

$$\begin{aligned} \textbf{Layer 0:}\quad & a_0 = x \cdot w_0 + b_0 \\ & z_0 = h_0(a_0) \\ \textbf{Layer 1:}\quad & a_1 = z_0 \cdot w_1 + b_1 \\ & \hat{y} = h_1(a_1) \end{aligned}$$

1.4 Backward Pass

Similar to the forward pass, the backward pass uses the information from the output to propagate back information to the input. The goal of this is to compute the gradient of the network by applying the chain rule relentlessly.
First one calculates the gradient of the used loss function with respect to the prediction and then the gradient of the loss function with respect to the activation of the last layer and so on and so forth. According to the given example, this would look like this:

$$\begin{aligned} \textbf{Layer 1:}\quad & \frac{\partial L}{\partial \hat{y}} & & \frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_1} \frac{\partial a_1}{w_1} \\ & \frac{\partial L}{\partial a_1} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_1} & & \frac{\partial L}{\partial b_1} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_1} \frac{\partial a_1}{b_1} \\ \textbf{Layer 0:}\quad & \frac{\partial L}{\partial z_0} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_1} \frac{\partial a_1}{\partial z_0} & & \frac{\partial L}{\partial w_0} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_1} \frac{\partial a_1}{\partial z_0} \frac{\partial z_0}{\partial a_0} \frac{\partial a_0}{\partial w_0} \\ & \frac{\partial L}{\partial a_0} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_1} \frac{\partial a_1}{\partial z_0} \frac{\partial z_0}{\partial a_0} & & \frac{\partial L}{\partial b_0} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_1} \frac{\partial a_1}{\partial z_0} \frac{\partial z_0}{\partial a_0} \frac{\partial a_0}{\partial b_0} \end{aligned}$$

2 Setup

2.1 Used Architecture

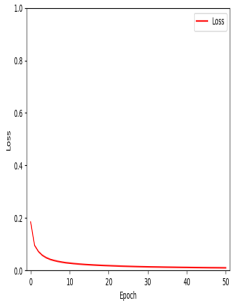
The objective dataset of this exercise is the MNIST dataset by Lecun et. al.and consists of a training set with 60000 examples and a test set with 10000 examples of handwritten digits from zero to nine (I used a reduced training set of 50000 in order to use the remaining 10000 data patterns for a validation set). The images are in the format 28x28 pixels.
Goal of this exercise is to train a feedforward neural network to correctly predict the labels of the handwritten digits.

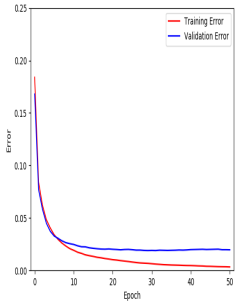
I used a stochastic gradient descent approach with a fixed batch size of 64 elements in order to achieve convergence. The network also uses a gradient checking routine that ensures a correct gradient computation in the backward propagation. My architecture also use cross validation with a validation set of 10000 examples, drawn from the original training set.
Remark: Most of the architecture was given due to the task's structure.

2.2 Used Training Setups

2.2.1 Setup 1

- Learning rate: 0.1
- Layer 1 (input)
  - units: 100
  - activation: ReLu
- Layer 2
  - units: 100
  - activation: ReLu
- Layer 3
  - units: 10 (output)
  - output: Linear Output



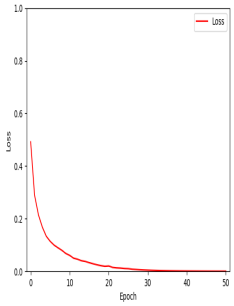


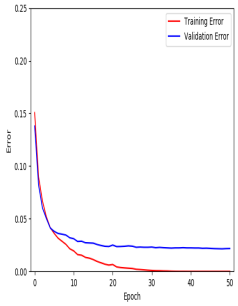
Epoch	Loss	Training Error	Validation Error	Train-Valid-Delta
0	0.1843	0.1839	0.1679	0.0160
1	0.0958	0.0844	0.0770	0.0074
2	0.0721	0.0622	0.0582	0.0040
3	0.0577	0.0484	0.0450	0.0034
4	0.0486	0.0405	0.0373	0.0032
5	0.0424	0.0340	0.0328	0.0012
6	0.0379	0.0296	0.0306	0.0010
7	0.0346	0.0260	0.0282	0.0022
8	0.0319	0.0232	0.0265	0.0033
9	0.0295	0.0208	0.0255	0.0047
10	0.0280	0.0192	0.0248	0.0056
11	0.0263	0.0174	0.0235	0.0061
12	0.0250	0.0163	0.0226	0.0063
13	0.0239	0.0149	0.0225	0.0076
14	0.0228	0.0141	0.0216	0.0075
15	0.0220	0.0134	0.0211	0.0077
16	0.0210	0.0126	0.0207	0.0081
17	0.0202	0.0121	0.0204	0.0083
18	0.0195	0.0114	0.0203	0.0089
19	0.0188	0.0109	0.0205	0.0096
20	0.0183	0.0103	0.0202	0.0099
21	0.0177	0.0099	0.0200	0.0101
22	0.0172	0.0094	0.0197	0.0103
23	0.0166	0.0090	0.0200	0.0110
24	0.0162	0.0085	0.0201	0.0116
25	0.0158	0.0080	0.0198	0.0118
26	0.0154	0.0076	0.0194	0.0118
27	0.0150	0.0072	0.0194	0.0122
28	0.0146	0.0070	0.0191	0.0121
29	0.0143	0.0068	0.0190	0.0122
30	0.0139	0.0064	0.0191	0.0127
31	0.0136	0.0061	0.0190	0.0129
32	0.0134	0.0059	0.0193	0.0134
33	0.0132	0.0056	0.0192	0.0136
34	0.0129	0.0054	0.0191	0.0137
35	0.0127	0.0052	0.0192	0.0140
36	0.0125	0.0051	0.0193	0.0142
37	0.0123	0.0050	0.0195	0.0145
38	0.0120	0.0048	0.0194	0.0146
39	0.0118	0.0047	0.0196	0.0149
40	0.0117	0.0047	0.0199	0.0152
41	0.0115	0.0045	0.0200	0.0155
42	0.0113	0.0044	0.0201	0.0157
43	0.0111	0.0043	0.0202	0.0159
44	0.0110	0.0040	0.0200	0.0160
45	0.0108	0.0040	0.0201	0.0161
46	0.0106	0.0038	0.0202	0.0164
47	0.0106	0.0037	0.0203	0.0166
48	0.0104	0.0036	0.0198	0.0162
49	0.0103	0.0035	0.0198	0.0163
50	0.0101	0.0034	0.0197	0.0163

Some arbitrarily chosen correct classification/misclassification examples:
Overall Test Error in Classification: 2.0800 %

2.2.2 Setup 2

- Learning rate: 0.1
- Layer 1 (input)
  - units: 100
  - activation: ReLu
- Layer 2
  - units: 100
  - activation: ReLu
- Layer 3
  - units: 10 (output)
  - output: Softmax



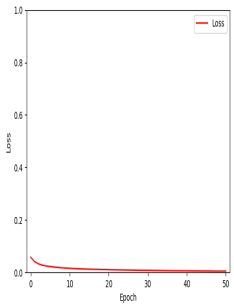


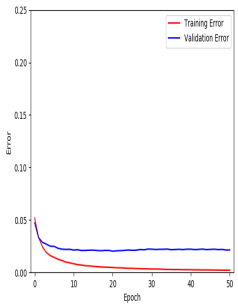
Epoch	Loss	Training Error	Validation Error	Train-Valid-Delta
0	0.4917	0.1506	0.1379	0.0127
1	0.2894	0.0898	0.0819	0.0079
2	0.2143	0.0668	0.0602	0.0066
3	0.1677	0.0518	0.0502	0.0016
4	0.1334	0.0415	0.0412	0.0003
5	0.1142	0.0361	0.0380	0.0019
6	0.0998	0.0315	0.0360	0.0045
7	0.0896	0.0286	0.0353	0.0067
8	0.0799	0.0257	0.0345	0.0088
9	0.0680	0.0214	0.0319	0.0105
10	0.0607	0.0195	0.0310	0.0115
11	0.0501	0.0159	0.0284	0.0125
12	0.0462	0.0154	0.0286	0.0132
13	0.0405	0.0133	0.0272	0.0139
14	0.0376	0.0125	0.0270	0.0145
15	0.0331	0.0112	0.0268	0.0156
16	0.0287	0.0094	0.0255	0.0161
17	0.0250	0.0081	0.0246	0.0165
18	0.0217	0.0068	0.0238	0.0170
19	0.0192	0.0060	0.0236	0.0176
20	0.0203	0.0065	0.0250	0.0185
21	0.0151	0.0042	0.0235	0.0193
22	0.0134	0.0037	0.0236	0.0199
23	0.0124	0.0033	0.0238	0.0205
24	0.0108	0.0030	0.0242	0.0212
25	0.0102	0.0027	0.0239	0.0212
26	0.0083	0.0020	0.0229	0.0209
27	0.0073	0.0017	0.0231	0.0214
28	0.0062	0.0014	0.0229	0.0215
29	0.0054	0.0011	0.0229	0.0218
30	0.0047	0.0008	0.0231	0.0223
31	0.0041	0.0007	0.0225	0.0218
32	0.0036	0.0006	0.0228	0.0222
33	0.0031	0.0004	0.0225	0.0221
34	0.0028	0.0004	0.0223	0.0219
35	0.0025	0.0002	0.0221	0.0219
36	0.0022	0.0001	0.0223	0.0222
37	0.0020	0.0000	0.0223	0.0223
38	0.0018	0.0000	0.0225	0.0225
39	0.0017	0.0000	0.0223	0.0223
40	0.0015	0.0000	0.0223	0.0223
41	0.0014	0.0000	0.0222	0.0222
42	0.0013	0.0000	0.0222	0.0222
43	0.0012	0.0000	0.0219	0.0219
44	0.0011	0.0000	0.0220	0.0220
45	0.0011	0.0000	0.0218	0.0218
46	0.0010	0.0000	0.0216	0.0216
47	0.0009	0.0000	0.0215	0.0215
48	0.0009	0.0000	0.0214	0.0214
49	0.0008	0.0000	0.0216	0.0216
50	0.0008	0.0000	0.0217	0.0217

Overall Test Error in Classification: 2.1600 %

2.2.3 Setup 3

- Learning rate: 0.5
- Layer 1 (input)
  - units: 100
  - activation: ReLu
- Layer 2
  - units: 100
  - activation: ReLu
- Layer 3
  - units: 10 (output)
  - output: Linear



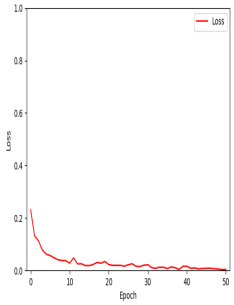


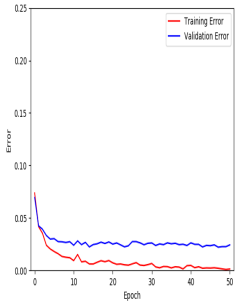
Epoch	Loss	Training Error	Validation Error	Train-Valid-Delta
0	0.0583	0.0519	0.0472	0.0047
1	0.0413	0.0336	0.0332	0.0004
2	0.0329	0.0244	0.0287	0.0043
3	0.0278	0.0191	0.0270	0.0079
4	0.0250	0.0161	0.0250	0.0089
5	0.0229	0.0144	0.0249	0.0105
6	0.0207	0.0126	0.0230	0.0104
7	0.0194	0.0114	0.0222	0.0108
8	0.0178	0.0099	0.0220	0.0121
9	0.0170	0.0092	0.0221	0.0129
10	0.0160	0.0083	0.0213	0.0130
11	0.0151	0.0075	0.0217	0.0142
12	0.0144	0.0071	0.0209	0.0138
13	0.0137	0.0064	0.0210	0.0146
14	0.0133	0.0063	0.0212	0.0149
15	0.0126	0.0058	0.0213	0.0155
16	0.0123	0.0056	0.0209	0.0153
17	0.0119	0.0053	0.0207	0.0154
18	0.0114	0.0051	0.0210	0.0159
19	0.0113	0.0049	0.0210	0.0161
20	0.0109	0.0047	0.0203	0.0156
21	0.0104	0.0045	0.0206	0.0161
22	0.0102	0.0045	0.0208	0.0163
23	0.0100	0.0042	0.0210	0.0168
24	0.0096	0.0040	0.0214	0.0174
25	0.0093	0.0040	0.0211	0.0171
26	0.0091	0.0038	0.0212	0.0174
27	0.0088	0.0037	0.0219	0.0182
28	0.0084	0.0035	0.0216	0.0181
29	0.0083	0.0034	0.0223	0.0189
30	0.0082	0.0033	0.0222	0.0189
31	0.0080	0.0033	0.0219	0.0186
32	0.0078	0.0032	0.0221	0.0189
33	0.0076	0.0030	0.0221	0.0191
34	0.0073	0.0029	0.0223	0.0194
35	0.0071	0.0029	0.0217	0.0188
36	0.0069	0.0028	0.0219	0.0191
37	0.0068	0.0028	0.0221	0.0193
38	0.0066	0.0027	0.0218	0.0191
39	0.0065	0.0026	0.0222	0.0196
40	0.0065	0.0026	0.0222	0.0196
41	0.0061	0.0025	0.0218	0.0193
42	0.0060	0.0025	0.0220	0.0195
43	0.0059	0.0024	0.0223	0.0199
44	0.0057	0.0025	0.0218	0.0193
45	0.0056	0.0024	0.0220	0.0196
46	0.0055	0.0023	0.0222	0.0199
47	0.0054	0.0023	0.0218	0.0195
48	0.0053	0.0022	0.0220	0.0198
49	0.0052	0.0022	0.0214	0.0192
50	0.0052	0.0021	0.0215	0.0194

Some arbitrarily chosen correct classification/misclassification examples:
Overall Test Error in Classification: 2.4100 %

2.2.4 Setup 4

- Learning rate: 0.5
- Layer 1 (input)
  - units: 100
  - activation: ReLu
- Layer 2
  - units: 100
  - activation: ReLu
- Layer 3
  - units: 10 (output)
  - output: Softmax



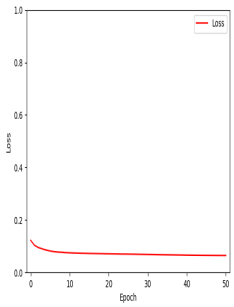


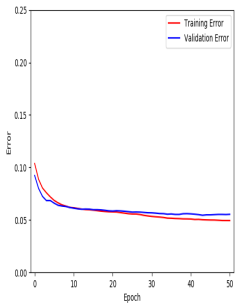
Epoch	Loss	Training Error	Validation Error	Train-Valid-Delta
0	0.2312	0.0737	0.0694	0.0043
1	0.1324	0.0418	0.0426	0.0008
2	0.1130	0.0355	0.0393	0.0038
3	0.0776	0.0240	0.0333	0.0093
4	0.0624	0.0205	0.0299	0.0094
5	0.0566	0.0181	0.0303	0.0122
6	0.0484	0.0160	0.0275	0.0115
7	0.0411	0.0134	0.0273	0.0139
8	0.0381	0.0126	0.0267	0.0141
9	0.0371	0.0121	0.0274	0.0153
10	0.0280	0.0094	0.0240	0.0146
11	0.0482	0.0152	0.0282	0.0130
12	0.0252	0.0082	0.0246	0.0164
13	0.0257	0.0088	0.0268	0.0180
14	0.0186	0.0062	0.0224	0.0162
15	0.0185	0.0062	0.0247	0.0185
16	0.0230	0.0078	0.0254	0.0176
17	0.0305	0.0093	0.0270	0.0177
18	0.0277	0.0084	0.0257	0.0173
19	0.0344	0.0094	0.0272	0.0178
20	0.0226	0.0073	0.0251	0.0178
21	0.0196	0.0059	0.0262	0.0203
22	0.0195	0.0062	0.0244	0.0182
23	0.0195	0.0054	0.0225	0.0171
24	0.0163	0.0051	0.0235	0.0184
25	0.0218	0.0063	0.0275	0.0212
26	0.0255	0.0075	0.0275	0.0200
27	0.0159	0.0051	0.0263	0.0212
28	0.0154	0.0047	0.0245	0.0198
29	0.0203	0.0056	0.0259	0.0203
30	0.0220	0.0067	0.0262	0.0195
31	0.0108	0.0034	0.0238	0.0204
32	0.0081	0.0026	0.0253	0.0227
33	0.0124	0.0038	0.0246	0.0208
34	0.0123	0.0037	0.0263	0.0226
35	0.0070	0.0025	0.0254	0.0229
36	0.0139	0.0035	0.0259	0.0224
37	0.0103	0.0032	0.0246	0.0214
38	0.0041	0.0015	0.0250	0.0235
39	0.0159	0.0046	0.0238	0.0192
40	0.0161	0.0048	0.0263	0.0215
41	0.0081	0.0027	0.0249	0.0222
42	0.0100	0.0036	0.0249	0.0213
43	0.0061	0.0021	0.0224	0.0203
44	0.0076	0.0024	0.0239	0.0215
45	0.0080	0.0023	0.0236	0.0213
46	0.0087	0.0026	0.0244	0.0218
47	0.0066	0.0021	0.0222	0.0201
48	0.0056	0.0015	0.0227	0.0212
49	0.0035	0.0010	0.0227	0.0217
50	0.0043	0.0014	0.0244	0.0230

Overall Test Error in Classification: 2.2000 %

2.2.5 Setup 5

- Learning rate: 0.1
- Layer 1 (input)
  - units: 20
  - activation: ReLu
- Layer 2
  - units: 10 (output)
  - output: Linear



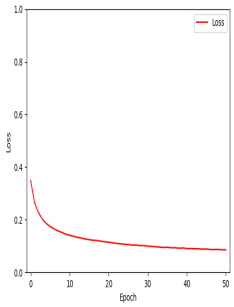


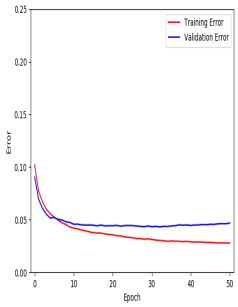
Epoch	Loss	Training Error	Validation Error	Train-Valid-Delta
0	0.1225	0.1037	0.0923	0.0114
1	0.1032	0.0888	0.0801	0.0087
2	0.0948	0.0805	0.0725	0.0080
3	0.0898	0.0760	0.0685	0.0075
4	0.0854	0.0720	0.0685	0.0035
5	0.0818	0.0686	0.0660	0.0026
6	0.0794	0.0663	0.0640	0.0023
7	0.0778	0.0644	0.0633	0.0011
8	0.0769	0.0632	0.0628	0.0004
9	0.0756	0.0620	0.0619	0.0001
10	0.0748	0.0617	0.0612	0.0005
11	0.0743	0.0610	0.0606	0.0004
12	0.0737	0.0604	0.0602	0.0002
13	0.0733	0.0599	0.0604	0.0005
14	0.0729	0.0597	0.0603	0.0006
15	0.0725	0.0594	0.0598	0.0004
16	0.0722	0.0589	0.0598	0.0009
17	0.0720	0.0584	0.0596	0.0012
18	0.0717	0.0580	0.0592	0.0012
19	0.0714	0.0578	0.0587	0.0009
20	0.0712	0.0577	0.0585	0.0008
21	0.0709	0.0576	0.0588	0.0012
22	0.0706	0.0570	0.0586	0.0016
23	0.0704	0.0565	0.0583	0.0018
24	0.0703	0.0560	0.0579	0.0019
25	0.0701	0.0557	0.0575	0.0018
26	0.0698	0.0557	0.0576	0.0019
27	0.0696	0.0552	0.0575	0.0023
28	0.0692	0.0544	0.0572	0.0028
29	0.0690	0.0539	0.0569	0.0030
30	0.0687	0.0534	0.0568	0.0034
31	0.0684	0.0530	0.0565	0.0035
32	0.0681	0.0528	0.0561	0.0033
33	0.0678	0.0524	0.0560	0.0036
34	0.0675	0.0518	0.0555	0.0037
35	0.0673	0.0516	0.0557	0.0041
36	0.0671	0.0514	0.0553	0.0039
37	0.0668	0.0513	0.0553	0.0040
38	0.0666	0.0510	0.0559	0.0049
39	0.0664	0.0510	0.0560	0.0050
40	0.0660	0.0509	0.0558	0.0049
41	0.0657	0.0505	0.0555	0.0050
42	0.0656	0.0506	0.0551	0.0045
43	0.0654	0.0502	0.0545	0.0043
44	0.0653	0.0501	0.0549	0.0048
45	0.0652	0.0500	0.0549	0.0049
46	0.0651	0.0499	0.0551	0.0052
47	0.0649	0.0497	0.0553	0.0056
48	0.0648	0.0495	0.0553	0.0058
49	0.0647	0.0495	0.0552	0.0057
50	0.0646	0.0495	0.0554	0.0059

Overall Test Error in Classification: 5.7300 %

2.2.6 Setup 6

- Learning rate: 0.1
- Layer 1 (input)
  - units: 20
  - activation: ReLu
- Layer 2
  - units: 10 (output)
  - output: Softmax



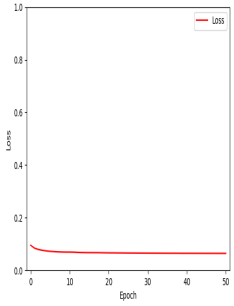


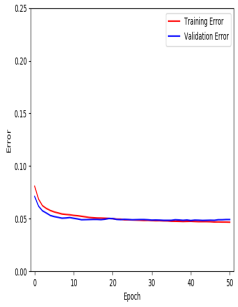
Epoch	Loss	Training Error	Validation Error	Train-Valid-Delta
0	0.3484	0.1019	0.0905	0.0114
1	0.2644	0.0773	0.0695	0.0078
2	0.2268	0.0670	0.0608	0.0062
3	0.2019	0.0599	0.0552	0.0047
4	0.1854	0.0557	0.0516	0.0041
5	0.1742	0.0524	0.0521	0.0003
6	0.1648	0.0496	0.0504	0.0008
7	0.1572	0.0471	0.0497	0.0026
8	0.1515	0.0454	0.0481	0.0027
9	0.1448	0.0431	0.0474	0.0043
10	0.1406	0.0421	0.0457	0.0036
11	0.1369	0.0413	0.0458	0.0045
12	0.1330	0.0403	0.0451	0.0048
13	0.1301	0.0395	0.0449	0.0054
14	0.1266	0.0385	0.0450	0.0065
15	0.1242	0.0377	0.0448	0.0071
16	0.1221	0.0374	0.0442	0.0068
17	0.1207	0.0373	0.0449	0.0076
18	0.1184	0.0365	0.0442	0.0077
19	0.1164	0.0360	0.0443	0.0083
20	0.1145	0.0357	0.0443	0.0086
21	0.1120	0.0349	0.0446	0.0097
22	0.1101	0.0346	0.0439	0.0093
23	0.1086	0.0338	0.0444	0.0106
24	0.1066	0.0333	0.0445	0.0112
25	0.1055	0.0329	0.0444	0.0115
26	0.1037	0.0323	0.0441	0.0118
27	0.1035	0.0322	0.0437	0.0115
28	0.1016	0.0314	0.0434	0.0120
29	0.1018	0.0318	0.0441	0.0123
30	0.0996	0.0313	0.0434	0.0121
31	0.0986	0.0307	0.0437	0.0130
32	0.0971	0.0303	0.0432	0.0129
33	0.0961	0.0300	0.0437	0.0137
34	0.0945	0.0295	0.0436	0.0141
35	0.0951	0.0298	0.0441	0.0143
36	0.0934	0.0296	0.0443	0.0147
37	0.0936	0.0295	0.0451	0.0156
38	0.0918	0.0292	0.0448	0.0156
39	0.0928	0.0295	0.0450	0.0155
40	0.0905	0.0290	0.0445	0.0155
41	0.0903	0.0288	0.0450	0.0162
42	0.0898	0.0289	0.0450	0.0161
43	0.0895	0.0287	0.0455	0.0168
44	0.0881	0.0284	0.0452	0.0168
45	0.0888	0.0285	0.0457	0.0172
46	0.0867	0.0282	0.0456	0.0174
47	0.0867	0.0280	0.0461	0.0181
48	0.0870	0.0280	0.0463	0.0183
49	0.0858	0.0279	0.0462	0.0183
50	0.0855	0.0278	0.0468	0.0190

Overall Test Error in Classification: 4.9100 %

2.2.7 Setup 7

- Learning rate: 0.5
- Layer 1 (input)
  - units: 20
  - activation: ReLu
- Layer 2
  - units: 10 (output)
  - output: Linear



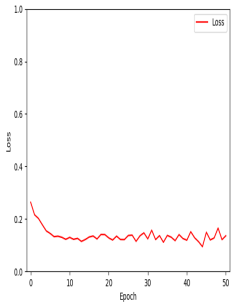


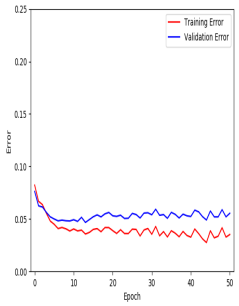
Epoch	Loss	Training Error	Validation Error	Train-Valid-Delta
0	0.0954	0.0808	0.0711	0.0097
1	0.0843	0.0686	0.0619	0.0067
2	0.0794	0.0624	0.0575	0.0049
3	0.0761	0.0597	0.0554	0.0043
4	0.0739	0.0578	0.0531	0.0047
5	0.0724	0.0564	0.0520	0.0044
6	0.0714	0.0555	0.0513	0.0042
7	0.0703	0.0545	0.0505	0.0040
8	0.0698	0.0540	0.0507	0.0033
9	0.0694	0.0537	0.0511	0.0026
10	0.0695	0.0532	0.0504	0.0028
11	0.0691	0.0528	0.0498	0.0030
12	0.0683	0.0523	0.0490	0.0033
13	0.0678	0.0517	0.0491	0.0026
14	0.0675	0.0512	0.0492	0.0020
15	0.0674	0.0509	0.0493	0.0016
16	0.0672	0.0507	0.0493	0.0014
17	0.0673	0.0506	0.0491	0.0015
18	0.0670	0.0505	0.0495	0.0010
19	0.0667	0.0503	0.0502	0.0001
20	0.0665	0.0501	0.0501	0.0000
21	0.0664	0.0496	0.0493	0.0003
22	0.0663	0.0495	0.0491	0.0004
23	0.0661	0.0489	0.0494	0.0005
24	0.0660	0.0489	0.0492	0.0003
25	0.0658	0.0488	0.0490	0.0002
26	0.0658	0.0487	0.0491	0.0004
27	0.0657	0.0486	0.0492	0.0006
28	0.0655	0.0484	0.0492	0.0008
29	0.0655	0.0484	0.0490	0.0006
30	0.0654	0.0483	0.0487	0.0004
31	0.0654	0.0481	0.0488	0.0007
32	0.0653	0.0481	0.0487	0.0006
33	0.0652	0.0480	0.0485	0.0005
34	0.0651	0.0479	0.0485	0.0006
35	0.0651	0.0476	0.0485	0.0009
36	0.0650	0.0475	0.0490	0.0015
37	0.0650	0.0475	0.0488	0.0013
38	0.0650	0.0473	0.0485	0.0012
39	0.0649	0.0474	0.0488	0.0014
40	0.0649	0.0475	0.0483	0.0008
41	0.0648	0.0473	0.0487	0.0014
42	0.0648	0.0472	0.0486	0.0014
43	0.0647	0.0472	0.0484	0.0012
44	0.0647	0.0472	0.0485	0.0013
45	0.0646	0.0471	0.0486	0.0015
46	0.0646	0.0469	0.0485	0.0016
47	0.0646	0.0469	0.0490	0.0021
48	0.0646	0.0468	0.0490	0.0022
49	0.0645	0.0468	0.0492	0.0024
50	0.0645	0.0467	0.0492	0.0025

Overall Test Error in Classification: 5.7600 %

2.2.8 Setup 8

- Learning rate: 0.5
- Layer 1 (input)
  - units: 20
  - activation: ReLu
- Layer 2
  - units: 10 (output)
  - output: Softmax





Epoch	Loss	Training Error	Validation Error	Train-Valid-Delta
0	0.2636	0.0822	0.0762	0.0060
1	0.2160	0.0668	0.0625	0.0043
2	0.2020	0.0637	0.0614	0.0023
3	0.1778	0.0554	0.0563	0.0009
4	0.1546	0.0480	0.0520	0.0040
5	0.1449	0.0450	0.0500	0.0050
6	0.1327	0.0409	0.0482	0.0073
7	0.1345	0.0419	0.0489	0.0070
8	0.1300	0.0407	0.0483	0.0076
9	0.1225	0.0385	0.0481	0.0096
10	0.1300	0.0405	0.0493	0.0088
11	0.1224	0.0386	0.0477	0.0091
12	0.1263	0.0394	0.0516	0.0122
13	0.1146	0.0357	0.0467	0.0110
14	0.1218	0.0372	0.0494	0.0122
15	0.1314	0.0400	0.0521	0.0121
16	0.1352	0.0408	0.0539	0.0131
17	0.1238	0.0379	0.0520	0.0141
18	0.1408	0.0418	0.0549	0.0131
19	0.1406	0.0418	0.0562	0.0144
20	0.1279	0.0389	0.0530	0.0141
21	0.1197	0.0362	0.0525	0.0163
22	0.1347	0.0398	0.0536	0.0138
23	0.1218	0.0362	0.0505	0.0143
24	0.1214	0.0362	0.0507	0.0145
25	0.1371	0.0403	0.0553	0.0150
26	0.1385	0.0400	0.0542	0.0142
27	0.1147	0.0338	0.0509	0.0171
28	0.1358	0.0395	0.0556	0.0161
29	0.1472	0.0410	0.0559	0.0149
30	0.1242	0.0354	0.0540	0.0186
31	0.1573	0.0429	0.0593	0.0164
32	0.1214	0.0339	0.0535	0.0196
33	0.1367	0.0380	0.0542	0.0162
34	0.1113	0.0329	0.0506	0.0177
35	0.1374	0.0388	0.0563	0.0175
36	0.1306	0.0365	0.0544	0.0179
37	0.1173	0.0331	0.0509	0.0178
38	0.1406	0.0381	0.0546	0.0165
39	0.1255	0.0343	0.0530	0.0187
40	0.1191	0.0327	0.0524	0.0197
41	0.1516	0.0405	0.0584	0.0179
42	0.1283	0.0361	0.0568	0.0207
43	0.1135	0.0310	0.0522	0.0212
44	0.0943	0.0275	0.0490	0.0215
45	0.1495	0.0388	0.0576	0.0188
46	0.1201	0.0320	0.0520	0.0200
47	0.1280	0.0336	0.0520	0.0184
48	0.1655	0.0417	0.0589	0.0172
49	0.1211	0.0328	0.0520	0.0192
50	0.1363	0.0352	0.0555	0.0203

Overall Test Error in Classification: 5.8400 %

2.3 Misclassification/Correct Classification Examples

Here are some examples of handwritten digits that the network misclassified as well as some it get correctly.