CSE 546 Homework 3A

Jake Gonzales

November 23, 2023

Collaborators: n/a

1 Conceptual Questions

A1.

(a) **False.** Training deep neural networks requires minimizing a non-convex loss function and therefore gradient descent isn't the best method to use because it may not necessarily reach a global minimum.

(b) **False.** It is not good practice to initialize all weights to zero when training a DNN because then all the weights may learn the same features.

(c) **True**. We use non-linear activation functions in hidden layers to learn non-linear boundaries otherwise we would only be able to learn functions that are a linear combination of the inputs.

(d) **False**. If we look at the governing equations then we can see that the time complexity of the backward pass step and forward pass are the same.

(e) **False**. Neural networks are not the best model (no model is the best model; it depends on the problem). If we had a limited amount of data or a simple linear problem then there is a different method that would be better like regression.

2 Kernels

A2.

(a) Suppose we have inputs x that are one-dimensional and our feature map is an infinite-dimensional $\phi(x)$ is a vector whose i-th component is:

$$\frac{1}{\sqrt{i!}} e^{\frac{-x^2}{2}} x^i$$

for all nonnegative integers i. We can write out the feature map as

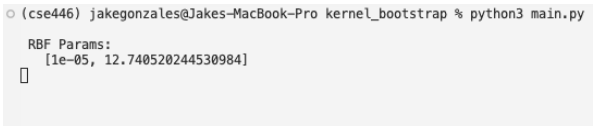
$$\begin{aligned}
\phi(x) \cdot \phi(x') &= (\frac{1}{\sqrt{i!}} e^{\frac{-x^2}{2}} x^i)(\frac{1}{\sqrt{i!}} e^{\frac{-x'^2}{2}} x'^i) \\
&= \sum_{i=1}^{\infty} \frac{1}{i!} x^i x'^i e^{\frac{-x^2-x'^2}{2}} \\
&= e^{\frac{-x^2-x'^2}{2}} \sum_{i=1}^{\infty} \frac{x^i x'^i}{i!} \\
&= e^{\frac{-x^2-x'^2}{2}} \\
&= e^{-\frac{(x-x')^2}{2}}
\end{aligned}$$

thus we have shown that $e^{-\frac{(x-x')^2}{2}}$ is a kernel function for the feature map.

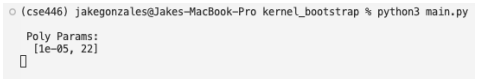
A3.

(a) The values are listed below in text and in the image from the output:

- RBF Kernel: $\gamma = 12.7$ and $\lambda = 1e-5$
- Polynomial Kernel: $d = 22$ and $\lambda = 1e-5$



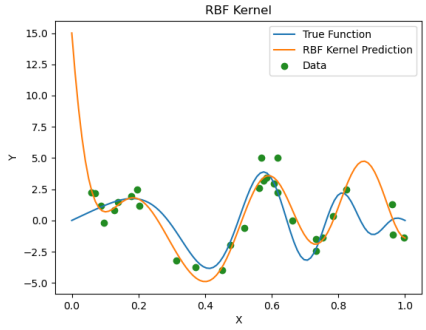
(a) RBF Params



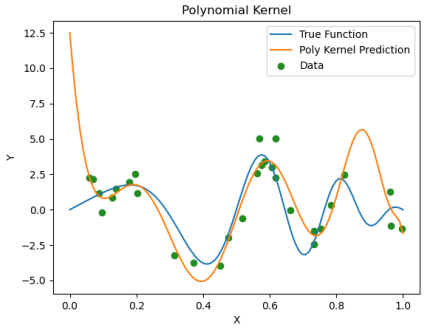
(b) Polynomial Params

Figure 1: Reported values for d, γ, λ

(b) Two plots for RBF kernel and polynomial kernel are below:



(a) RBF Kernel



(b) Polynomial Kernel

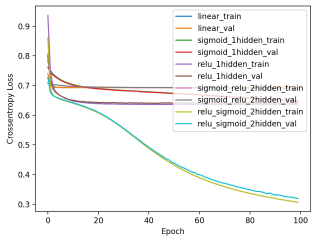
Figure 2: Plots for part B

3 Introduction to PyTorch

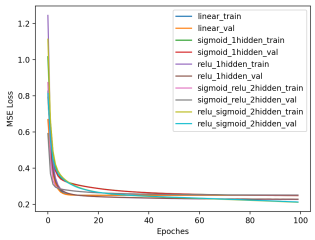
A4.

(a) Code submitted.

(b) Two plots (one per loss function) for cross entropy and mean squared error (MSE) showing both training and validation loss.



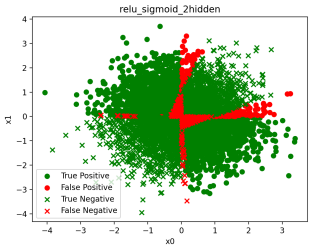
(a) Training Curves Crossentropy



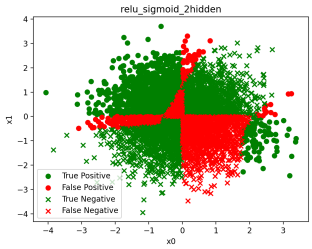
(b) Training Curves MSE

Figure 3: Part B plots

(c) Names of the best performing models for both loss functions was the ReLu then Sigmoid with 2 hidden layers. The model accuracy for each is below in Figure 5. Also, the scatter plots (one per loss function) with predictions of the best performing models (relu then sigmoid with 2 hidden) is below in Figure 4.

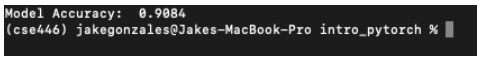


(a) Crossentropy Scatterplot

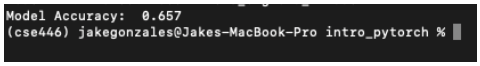


(b) MSE scatterplot

Figure 4: Part C scatterplots



(a) Crossentropy Model Accuracy



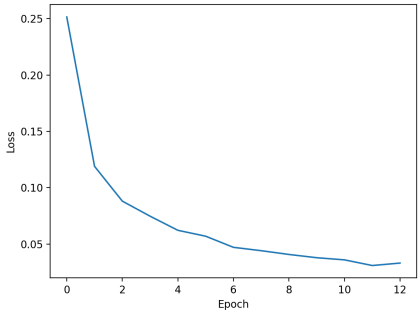
(b) MSE Model Accuracy

Figure 5: Part C Reported Model Accuracy

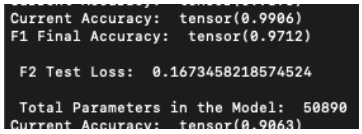
4 Neural Networks for MNIST

A5.

(a) The final accuracy was 97.12 and the loss on test data was 0.167. The reported values and epoch v. loss function is below.



(a) F1 Training Loss

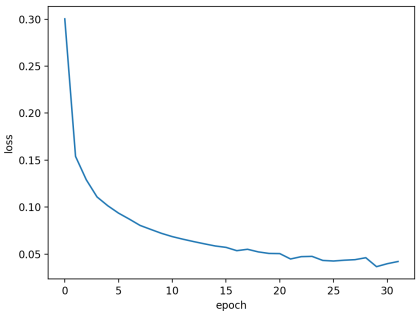


(b) F1 Accuracy and Loss

Figure 6: Part A Training Loss and Accuracy

(b) The final accuracy was 0.9639 and the loss on the test data was 0.1913.

(c) The two-layer neural network model had 50,890 parameters and the three-layer had 26,506 parameters (both listed above in the figures). Both of the models performed similarly, the training time for the three-layer model was less extensive because there are nearly half the parameters in the model.



(a) F2 Training Loss

```
Current Accuracy:  tensor(0.9845)
F2 Final Accuracy:  tensor(0.9639)

 F2 Test Loss:  0.1913653016090393

 Total Parameters in the Model:  26506
(cse446) jakegonzales@Jakes-MacBook-Pro neural_network_mnist %
```

(b) F2 Accuracy and Loss

Figure 7: Part B Training Loss and Accuracy

5 Administrative

A6.

(a) About 25 hours.