CIS 520, Machine Learning, Fall 2020
Homework 4
Due: Monday, October 12th, 11:59pm
Submit to Gradescope

Your name here

1 Neural Networks: Backpropagation

Your task is to now compute the derivatives of the loss function given in 1 with respect to W_1, W_2, b_1 and b_2 by hand, i.e., $\frac{\partial Loss}{\partial W_1}$, $\frac{\partial Loss}{\partial b_1}$, and $\frac{\partial Loss}{\partial b_2}$.

$$Loss = y * ln(\sigma(z_2)) + (1 - y) * ln(1 - \sigma(z_2)) \tag{1}$$

Show all the intermediate derivative computation steps. You might benefit from making a rough schematic of the backpropagation process. Also recall the derivatives of the softmax function and the tanh function:

$$\frac{d\sigma(z_2)}{dz} = \sigma(z_2) * (1 - \sigma(z_2)) \tag{2}$$

$$\frac{\partial tanh(z_1)}{\partial z_1} = 1 - tanh^2(z_1) \tag{3}$$

$$\frac{\partial Loss}{\partial W_1} = \dots \tag{4}$$

$$\frac{\partial Loss}{\partial W_2} = \dots \tag{5}$$

$$\frac{\partial Loss}{\partial b_1} = \dots \tag{6}$$

$$\frac{\partial Loss}{\partial b_2} = \dots \tag{7}$$

2 Programming

2.1 Random Forests

Tabulate the prediction results on the test set in Table 1

n_estimators	**Accuracy (%)**
1	73 %
5	92 %
10	95 %
50	96 %
100	97 %
500	97 %

Table 1: Accuracy for the Random Forests classification problem on the test set

2.2 Kernel SVM

Tabulate the prediction results on the test set in Table 2.

kernel	Accuracy (%)
Linear	98 %
Poly	99 %
RBF	98 %

Table 2: Accuracy for the kernel SVM classification problem on the test set

2.3 Multi Layer Perceptron

Tabulate the prediction results on the test set in Table 3.

Network Architecture	Accuracy (%)
(3)	85 %
(10)	94 %
(10,10,10)	92 %
(20,50,20)	96 %

Table 3: Accuracy for the MLP classification problem on the test set

2.4 AdaBoost

Tabulate the prediction results on the test set in Table 4.

n_estimators	Accuracy (%)
1	43 %
5	82 %
10	83 %
50	89 %
100	92 %
150	92 %

Table 4: Accuracy for the AdaBoost classification problem on the test set

2.5 Short Answer

For the random forest, AdaBoost and the Multi Layer Perceptron, we see that accuracy improves from the 70-80 % range to the low to high 90%.

For the Kernel SVM, linear, polynomial and RBF worked well, but the polynomial kernel performed the best. This indicates that the correct decision boundary is best described by a polynomial function, and the RBF and Linear are underfitting or overfitting the model respectively.

3 Convolutional Neural Networks

3.1 Theory

1. The output volume is:

...

2. Including bias parameters, this hidden layer has the following number of parameters:

...

3. The output from the simple convolutional layer for the given filter and input is:

Row	Column	Filter	Value
1	1	1	—
1	1	2	—
1	2	1	—
2	1	1	—

Table 5: Output from convolution

4. The total number of trainable parameters in this network is:

...

5. State two advantages of convolutional neural networks over fully connected networks.

   ...

3.2 Programming

For this question, refer to the Jupyter Notebook. You will be using PyTorch to implement a convolutional neural network – the notebook will have detailed instructions. We will be using the fashion MNIST dataset for a classification task.

3.2.1 Convolutional Neural Network

Add the accuracy and the loss curve from tensorboard in this report:

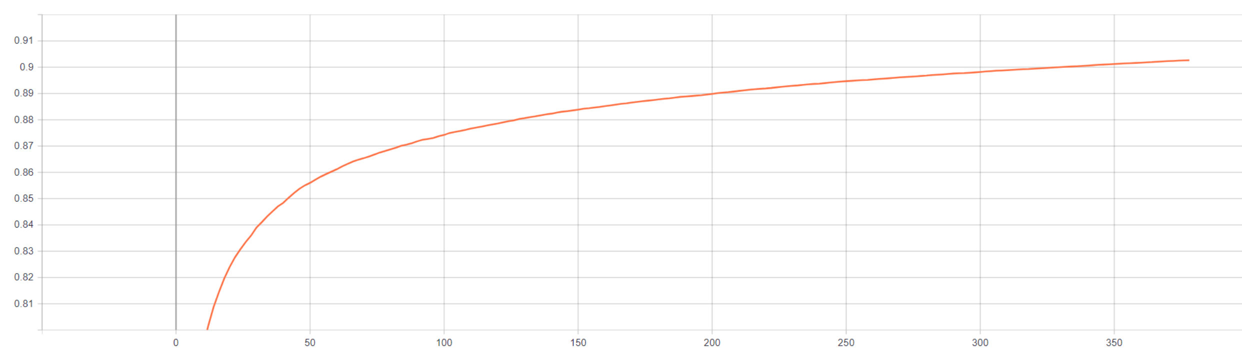


Figure 1: Accuracy curve

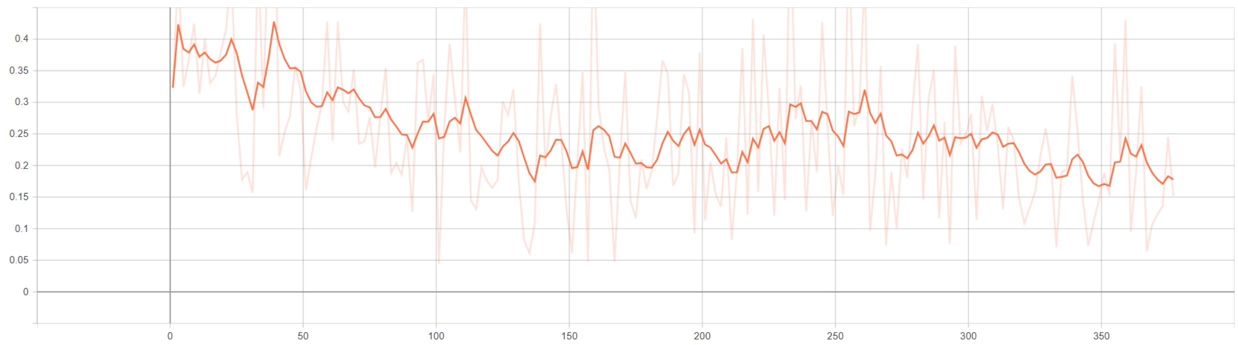


Figure 2: Loss curve

3.2.2 Accuracy

Report the overall accuracy and the per-class accuracy:

We did not choose to make any improvements to the baseline model, hence do not need to provide an explanation of our model architecture.

Overall Accuracy	89 %

Table 6: Overall Accuracy for Convolutional Neural Network

Class	**Accuracy**
T-shirt/top	88 %
Trouser	99 %
Pullover	85 %
Dress	90 %
Coat	82 %
Sandal	96 %
Shirt	74 %
Sneaker	96 %
Bag	98 %
Ankle boot	96 %

Table 7: Per Class Accuracy for Convolutional Neural Network

Identify the problematic classes and list the possible reasons as to why these classes may have significantly lower accuracy compared to other classes.

The problematic classes here were Shirt, Coat and T-Shirt / top. Some possible reasons why these classes have significantly lower accuracy compared to other classes:

- A lot of t-shirts get classified as shirts and vice versa because of the similarities in visual appearance of the garments.
- Coats can also be wrongly classified as shirts, since both have collars and long sleeves.

To rectify this issue, we can perform data augmentation for these problematic classes to improve the classifiers performance.

4 Multiclass Adaboost and Support Vector Machines

4.1 Adaboost: Theory

(a) Show that

$$D_{T+1}(i) \;=\; \frac{\frac{1}{m}\, e^{-F_{T,y_i}(x_i)}}{\prod_{t=1}^{T} Z_t}\,.$$

(b) Show that

$$\mathbf{1}(H(x_i) \neq y_i) \;\leq\; \mathbf{1}\big(F_{T,y_i}(x_i) < 0\big)\,.$$

(c) Show that

$$\mathrm{er}_S[H] \;\leq\; \frac{1}{m}\sum_{i=1}^{m} e^{-F_{T,y_i}(x_i)} \;=\; \prod_{t=1}^{T} Z_t\,.$$

Hint: For the inequality, use the result of part (b) above, and the fact that $\mathbf{1}(u<0) \leq e^{-u}$; for the equality, use the result of part (a) above.

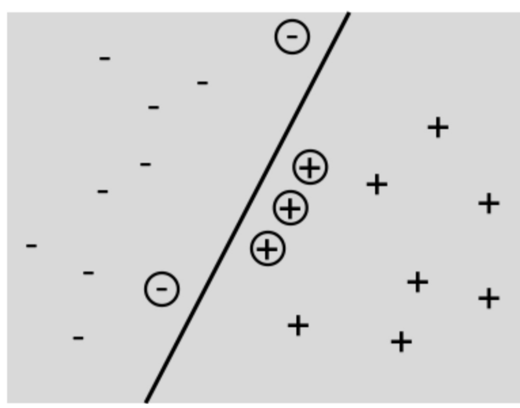
(d) Show that for the given choice of α_t, we have

$$Z_t \;=\; 2\sqrt{\mathrm{er}_t(1-\mathrm{er}_t)}\,.$$

(e) Suppose $\mathrm{er}_t \leq \frac{1}{2} - \gamma$ for all t (where $0 < \gamma \leq \frac{1}{2}$). Then show that

$$\mathrm{er}_S[H] \;\leq\; e^{-2T\gamma^2}\,.$$

4.2 Support Vector Machine



What is the largest number of data points that can be removed from the training set without changing the hard margin SVM solution? Explain your solution.

. . .