Assignment 1

Hyungi Kim

Department of Electrical and Computer Engineering
Seoul National University

http://data.snu.ac.kr

Assignment Objectives

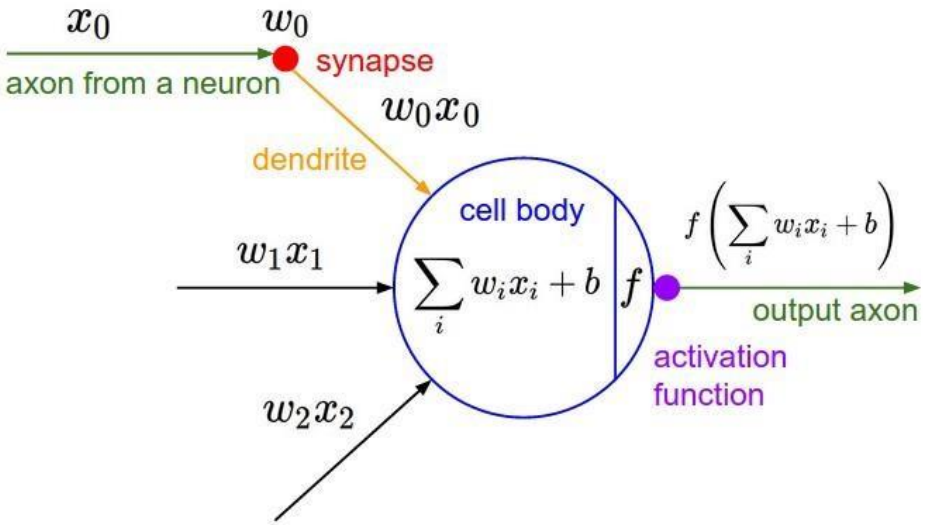
- Part 1: Data Curation
  - Practice loading and preprocessing of data using the notMNIST dataset
  - Implement a simple machine learning code using *sklearn* library
- Part 2: Implementing Neural Networks from Scratch
  - Understand the deep learning models
  - Implement a simple deep learning model
- Part 3: Neural Networks with PyTorch
  - Understand the roles of hyperparameter
  - Practice PyTorch code implementing deep learning models

notMNIST dataset

- Consist of characters rendered in a variety of fonts on a 28x28 image
- 10 classes, with letters A-J
- Training set: notMNIST_large (uncleaned, 500k instances)
- Test set: notMNIST_small (hand-cleaned, about 19k instances)

Training a Neural Network

- Artificial Neuron

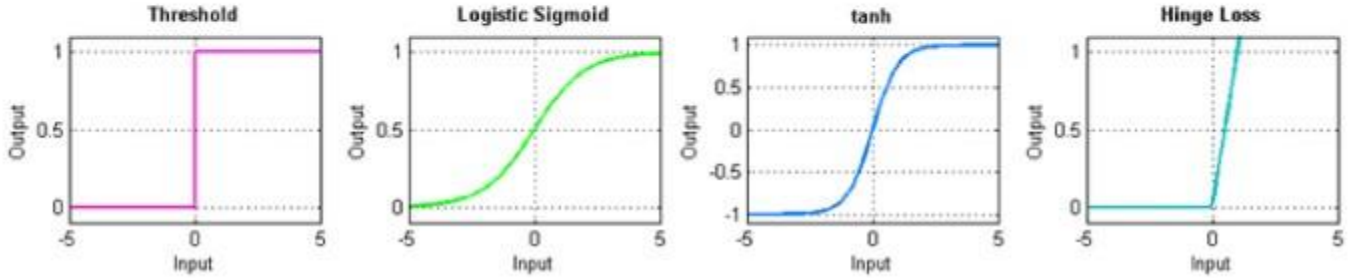


- Activation Functions

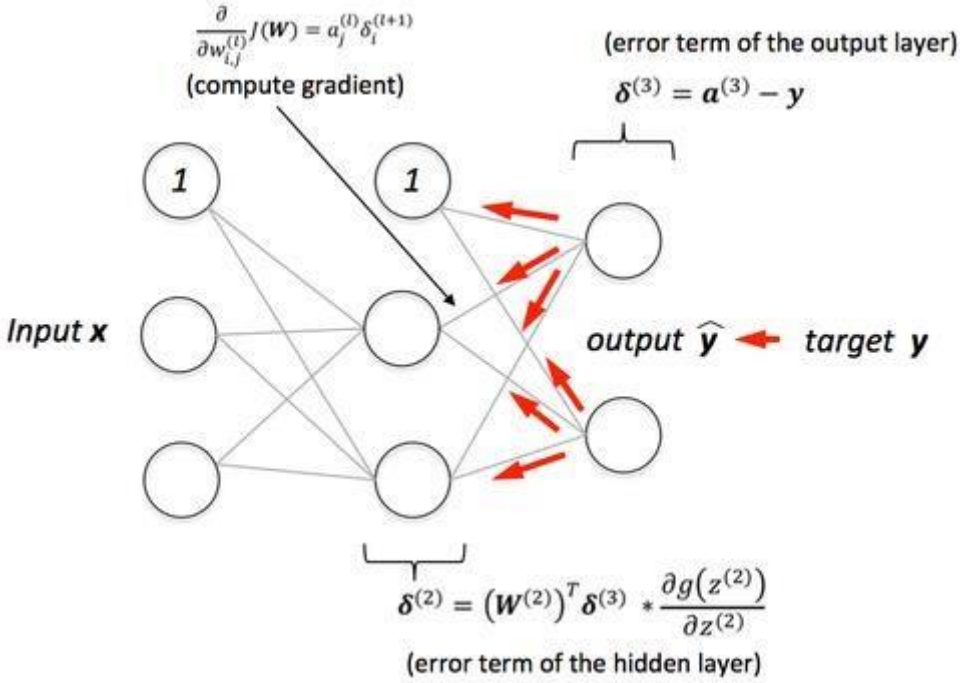
$$\text{sigmoid: } \sigma(x) = \frac{1}{1+e^{-x}}$$

$$\tanh: \sigma(x) = \frac{e^x - e^x}{e^x + e^x}$$

$$\text{ReLU: } \sigma(x) = max(0, x)$$



Training a Neural Network



- BackPropagation
  - Correct parameter weights by calculating the derivatives of the cost function w.r.t. each parameter of the NN.
  - Optimized with (mini-batch, stochastic) gradient descent

Hyperparameters

- Hyperparameters
  - Learning rate
  - Mini-batch size
  - Number of training iterations
  - Weight initialization
  - …
- Choosing a set of optimal hyperparameters
  - Is challenging and heuristic

How to install assignment files

- Assignment files
  - Utils/ (image file included for explanation)
  - data/ (empty)
  - model_checkpoints/ (empty)
  - Assignment1-1_Data_Curation.ipynb
  - Assignment1-2_NN_from_scratch.ipynb
  - Assignment1-3_NN_with_PyTorch.ipynb
  - CollectSubmission.sh

- Install assignment files
  - $ tar zxvf Assignment1.tar.gz (decompress tar gz file)
  - $ chmod 755 CollectSubmission.sh (get permission of script file)

- Open the notebooks on your browser and get started!

Output Examples

- Assignment1-1_Data_Curation

```
100 samples accuracy:  0.05
500 samples accuracy:  0.21
2500 samples accuracy:  0.37
10000 samples accuracy:  0.76
```

- Assignment1-2_NN_from_scratch

```
Loss after iteration 0: 2.253004
prediction accuracy :  21.99 %
Loss after iteration 1000: 0.121848
Loss after iteration 2000: 0.014129
Loss after iteration 3000: 0.012137
Loss after iteration 4000: 0.009871
Loss after iteration 5000: 0.009233
```

- Assignment1-3_NN_with_PyTorch

```
epoch: 50 [1000 / 3125]  train_loss: 0.256       train_accuracy: 90.6
epoch: 50 [2000 / 3125]  train_loss: 0.150       train_accuracy: 98.4
epoch: 50 [3000 / 3125]  train_loss: 0.226       train_accuracy: 93.8
-------- validation --------
accuracy: 90.1

-------- test ---------
accuracy: 95.5

my_model saved
```

Important Notes

- Due: 10/7 23:59
- PLEASE read the notes on the notebooks carefully
- Google first before mailing TAs

- Submitting your work
  - DO NOT clear the final outputs
  - After you are done **all three parts**
    - ✓ $./CollectSubmission.sh 2000-00000(학번)
    - ✓ Upload the 2000-00000.tar.gz on ETL

- TA email: deeplearning.snu@gmail.com

FAQ :1-1

Q: 1-1에서 exercise는 점수에 포함되나요?

A: 포함되지 않으며, reference code를 참조해보며 공부를 돕기 위한 문제들입니다.

Q: exercise 5에서 hash function이 1대1 함수가 아니기 때문에 hash 값이 같은 key 값을 보장하지 않아서 정확한 overlap을 구할 수 없는 것 같은데 이걸 사용하는게 맞나요?

A: hashing을 이용하면 정확한 overlap을 보장할 수 없는 게 맞으나 exercise5를 통해 의도한 바는 train, validation, test data set이 서로 disjoint한 dataset인지 확인하고, overla되는 data가 있다면 '대략적으로 ' 어느정도 되는지(전체 dataset의 몇 % 정도)를 추정하는 것이기 때문에 힌트에 나와 있는 hash와 set 함수를 이용하는 것을 추천드립니다.

FAQ :1-1

Q: problem에서 [100, 500, 2500, 10000] 크기의 train set으로 train 했을 경우 test set도 [100, 500, 2500, 10000]으로 해야 하나요?
A: test set에 대해서 train set의 크기를 변경했을 때 어떠한 변화가 발생하는 지를 확인하는 것이 목적이기 때문에 test set은 그대로 유지 (전체 test set을 사용)하여 테스트 하시면 됩니다.

Q: problem에서 logistic regression 함수를 사용하는 이유는 무엇인가요.
A: 해당 library의 해당 함수가 multi class classification task를 하게끔 하는 기능이 있습니다.

FAQ :1-2, 1-3

1-2

Q: loss 값과 정확도 값의 통과 기준은 무엇인가요?

A: training이 잘 진행되는 지 관찰하는게 목적이기 때문에, loss가 감소하고 test accuracy가 적절한 값이 나오는 지만 확인하면 됩니다.

1-3

Q: 제시한 모델의 개선 방법 중 아무것도 수행하지 않았는데 요구한 성능에 도달했는데 괜찮은가요?

A: 개선 방법을 제시해드린 것은 hyperparameter tuning에 익숙해질 수 있게끔 여러가지 방법을 추천해드린 것이기 때문에 이미 요구조건을 만족시키셨다면 굳이 하지 않으셔도 상관 없습니다만 이후에 있을 과제에 대비하기 위해 한 번 해보는 것을 추천드립니다.


Thank You!