Artificial Neural Networks & Deep Learning

HW #2

Breast Cancer Wisconsin Dataset

- Classification problem
  - 10 input variables
  - 1 binary output variable (benign or malignant)
- Originally hosted by UCI
- 569 data samples
  - Use the first 100 samples as test set
  - Use the next 100 samples as validation set
  - Use the others as training set

- https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

Data Preparation

- Download **breast-cancer-wisconsin.data**
- Remove the rows with missing values "?"
  - With any text editor
- Load it in the python
- Drop the first column:
  - The first column is ID, which does not carry any information about the tissue.
- Normalize the input variables.
- Set the output variable
  - Set Malignant: 1, benign: 0
- Data split: train, test, & validation set

Basic model

- Model Structure
  - 9 inputs
  - 10 hidden neurons with ReLu activation functions
  - 1 output neuron with sigmoid activation function.
- Compile and learning condition
  - Optimizer=rmsprop,
  - Loss function=binary crossentropy
  - Epochs=200
  - Batch_size=10
  - EarlyStopping with patience=2

Q1: (1 point)

- Show your code.

- 0.5 points for data preparation.
  - 1. data load
  - 2. data normalization
  - 3. output coding
  - 4. data split

- 0.5 points for the model definition and learning.
  - 1. model definition
  - 2. model setup (loss, optimizer)
  - 3. correct fitting procedure
  - 4. correct evaluation procedure

Q2: (1 point)

- Repeat training of the model 5 times, and collect their losses and accuracies using the table below.
  - 0.5 points for collecting the data.
- Are they all consistent over trials? If not, why?
  - 0.5 points for the answer of this question

	Trial #1	Trial #2	Trial #3	Trial #4	Trial #5
Training loss					
Training accuracy					
Test Loss					
Test accuracy					

Q3: (1 point)

- Investigate whether the activation function of the hidden layer affects the accuracy.
  - Still the same model (the # of hidden neurons: 10)
  - 4 different cases: None, Relu, sigmoid, tanh
  - For each case, repeat training 10 times and report the mean and standard deviation of loss and accuracy in the training and test data set.
  - Use the similar table in the problem #2.
  - 0.5 points for collecting data
- Which one is the best? Why? (0.5 points)

Q4: (2 points)

- Let’s investigate how the number of hidden neurons affects the performance.
  - Set the activation function of the hidden layer to Relu.
- Change # of hidden neurons systematically, and then re-training the model.
  - Collect the data and construct the table for the following # of hidden neurons: 2, 5, 10, 20, 50, 100, 200, 500, 1000, 2000.
  - For each case, repeat training 5 times and report the mean and standard deviation of loss and accuracy in the training and test data set.
  - 1 points for collecting data.
- What is the best case? Why did you select it? (i.e. which one did you use among 4 metrics you collected?)
  - 1 points for the answer.

Q E1(0.5 points)

- Generally, after performing the rough search we did in the Q4, we performed the more fine-tuned search for the optimal # of hidden neurons.
  - As an example, if we found that the best performance was achieved near 20~50, we performed another experiment varying # of hidden neurons: 25, 30, 35, 40, 45, and select the case with the best performance.
- The question is “why didn’t we try all cases at once?”
  - As an example, we can try for all cases: 2, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60 ….
  - But we don’t. Why? (0.5 points)

Q E2 (0.5 point)

- After learning, we can analyze the learned weights.
- Construct a model without a hidden layer; all input units are directly corrected to the output.
- After learning, using the following commands, you can get the weights and bias.
  - w=model.get_weights()[0]
  - b=model.get_weights()[1]

- Please analyze the model based on the learned weights. What does the large weight mean? What does the weight near zero mean? What does the negative value mean?
    - Check breast-cancer-wisconsin.names.

Deadlines & Submission

- Total scores: 5 points + extra 1 points

- Due: Apr 12. 2021. 11:59 PM (Monday)
  - No grace period.
  - Be punctual, 1 day delay = 1 point penalty.

- How to submit
  - Use Blackboard's assignment tab.
  - 블랙보드(kulms.korea.ac.kr) -> assignments 탭에서 제출
  - No email submission. (이메일 제출 안 받습니다.)

  - Use the given template file to write a report.
  - 주어진 템플릿 파일 사용해서 리포트 작성.

  - Also submit the codes you used. The compressed file name should be **hw2_codes.zip**. If you do not submit the codes, the score will not given.
  - 사용하신 코드를 압축해서 제출해주세요. 압축화일 이름은 **hw2_codes.zip**으로 해주세요. 코드를 제출 안 하시면 숙제 점수는 0점입니다.