CSC321 Tutorial 5 part A: Assignment 1 review

Yue Li
Email: yueli@cs.toronto.edu

Wed 11-12 Feb 12
Fri 10-11 Feb 14

Marking Scheme

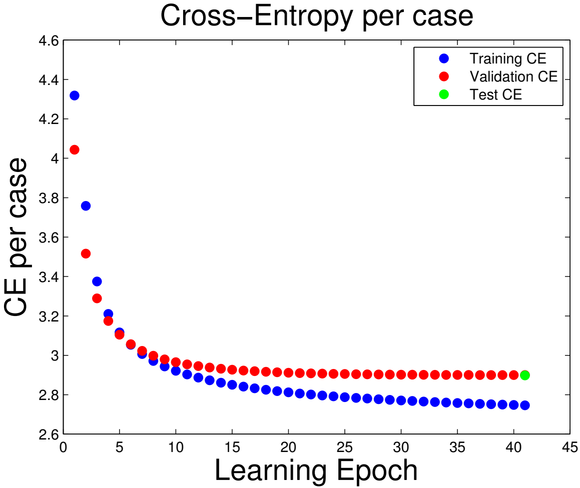
1. Train the model four times, trying all possible combinations of d=10, d=40 and numHid=50,numHid=200 (**4 marks**)

d	numHid	epoch	trainCE	validCE	testCE
10	50	38	2.73	2.89	2.89
10	200	21	2.54	2.81	2.81
40	50	23	2.60	2.82	2.82
40	200	**17**	**2.43**	**2.75**	**2.75**

Based on the CE, the best network configuration is `d` = 40, `numHid` = 200.

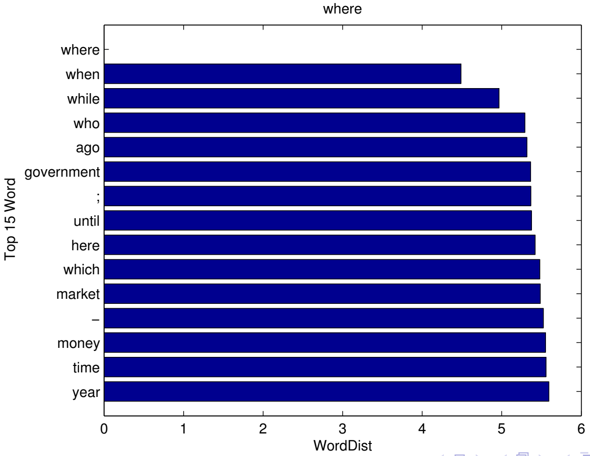
Marking Scheme

2. Show at least for one run the training, testing, and validation errors on the same plot as illustrated in Tutorial 3 (**1 mark**).



Marking Scheme

3. Compare word distances between words and make some intuitive observation (**1 mark**).
   - Two words are placed near each other if they are used in the same or similar contexts in the training data.

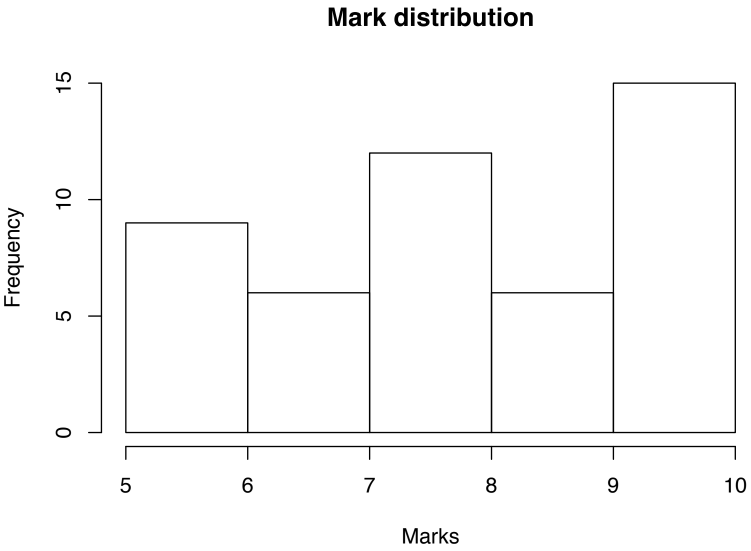


Marking Scheme

4. I found that the test error decreased as the number of weights increased **(1 mark)**.
   - This suggests that there is enough training data to adequately train the largest model (d=40 and NumHid=200).
   - To confirm this, the variable `trainData` shows that there are 372,550 training cases in `4grams.mat` ("by default", `loadData(100, 1000)` loads 100,000 cases).
   - The total number of weights in the largest model is 84,450 (this is the total size of the arrays `wordRepsFinal`, `repToHidFinal`, `hidToOutFinal`, `hidBiasFinal`, `outBiasFinal`).
   - This is about 4.4 (or 1.2) training cases per weight, which is a bit low but could be enough to learn the weights accurately.

Marking Scheme

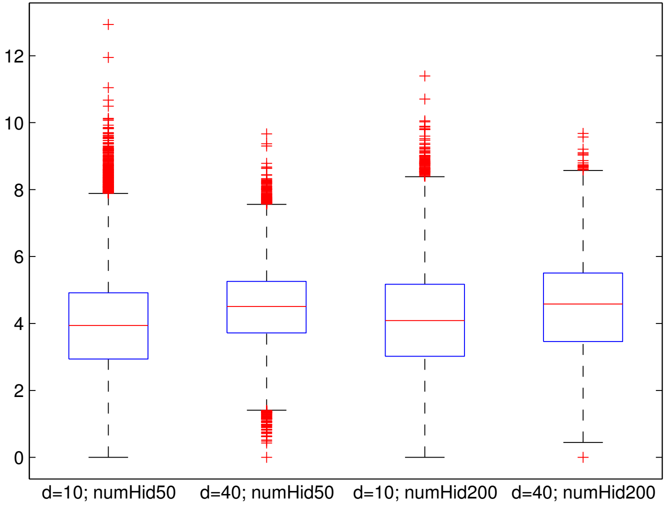
5. Any three related observations below (**3 marks**)
    - Models with more weights have lower training error, as one would expect. Testing and validation errors are bigger than training errors, as one would expect.
    - Testing errors are very close to validation errors, and usually they are bigger, as one would expect.
    - A training epoch takes more time in models with more weights. This makes sense since models with more weights have to do more work during back propagation.
    - I found that models with more weights took fewer epochs to terminate. This makes sense since with more weights it is easier to overfit the data, so the test error should rise sooner.
    - This shows that generalization is possible (and overfitting can be avoided) in a neural net with a large number of weights (on the order of 100,000 weights) if there is enough training data.
    - Running each model several times gives similar results and shows that the training error, validation error and test error are all accurate to two or three significant digits.



Median=8; Mean=7.9; Max=10; Min=5

Common issues:

1. Quantitative evaluation (testCE) vs qualitative evaluation (wordDistance); the former is most rigorous and should be performed first.
2. All of the training data batch are used in each epoch. Overfitting occurs not because of adding new input data but rather because of over-updating weights
3. We should not compare the absolute distance of the same word pair between different embedding schemes because the higher the number of the embedding dimension, the higher the distance value due to Euclidean distance formula. An unbiased comparison between the four settings is to compare either the rankings of words or the distance b/w two words with the averaged distance of each word with all of the other words (see boxplot in next slide).
4. The word distance indicates whether the two words are interchangeable rather than adjacent to each other.
5. Smaller number of epochs does not necessarily imply higher efficiency since each epoch may take longer time when network is large (i.e., high d and numHid).


12
10
8
6
4
2
0
d=10; numHid50
d=40; numHid50
d=10; numHid200
d=40; numHid200