Comments on Assignment 1

Jian Yao

CSC 321 : Introduction to Neural Networks and Machine Learning
Department of Computer Science
University of Toronto

February 6, 2014

Non-technical suggestions

- Please staple your homework;
- Please follow the instructions (e.g. no more than 2 pages, try the required experiments)

Model selection

- Criterion for choosing the best model
    - Having more hidden units is not always better.
    - It happens to help in this case because we are doing a decent job of controlling overfitting.

Model selection

- Criterion for choosing the best model
  - Having more hidden units is not always better.
  - It happens to help in this case because we are doing a decent job of controlling overfitting.
  - Final Cross Entropy on Test or Validation set?
- Justification for choosing the criterion
  - Test Error Measures performance on unseen data, test of generalization
  - But it is cheating in some sense to use it for model selection

What is the net expected to do?

What kind of words would be *expected* to be close ?

- Why would this property arise?
  - The network is trying to predict the 4-th word in a 4-gram.
  - Think : What kind of word representation would make it easy for the net to do that?
  - What is the backpropagated signal telling the net to do?

What is the net expected to do?

What kind of words would be *expected* to be close ?

- Why would this property arise?
  - The network is trying to predict the 4-th word in a 4-gram.
  - Think : What kind of word representation would make it easy for the net to do that?
  - What is the backpropagated signal telling the net to do?
- Words that occur next (or close) to each other in the sentence? NO
- Phonetically, morphologically close ? NO
- Words that have same semantic meaning? not always
- Words which belong to the same Part-of-speech? not always

What is the net expected to do?

What kind of words would be *expected* to be close ?

- Why would this property arise?
    - The network is trying to predict the 4-th word in a 4-gram.
    - Think : What kind of word representation would make it easy for the net to do that?
    - What is the backpropagated signal telling the net to do?
- Words that occur next (or close) to each other in the sentence? NO
- Phonetically, morphologically close ? NO
- Words that have same semantic meaning? not always
- Words which belong to the same Part-of-speech? not always
- Words that can be substituted for each other and still make up a sensible sentence/phrase. YES. This includes-
    - `'could'`, `'should'`, `'might'` - modal verbs
    - `'two'`, `'three'`, `'five'`
    - `'house'`, `'home'`, `'school'`
- Or more generally, words that are strongly correlated with another word appearing within the next 3.

Playing with `wordDistance`

- Cases where things worked/didn't work out as expected
- Understanding what the numbers really mean and how to compare them.
- It only makes sense to compare relative distances between words.
  - $d(A, B)$ and $d(A, C)$
  - $d(A, B)$ and $\langle d(A, w) \rangle$, $\langle d(B, w) \rangle$
  - NOT $d(A, B)$ and $d(C, D)$

Playing with `wordDistance`

- Cases where things worked/didn't work out as expected
- Understanding what the numbers really mean and how to compare them.
- It only makes sense to compare relative distances between words.
  - $d(A, B)$ and $d(A, C)$
  - $d(A, B)$ and $\langle d(A, w) \rangle$, $\langle d(B, w) \rangle$
  - NOT $d(A, B)$ and $d(C, D)$
- Rare words will cluster near the origin because their weights started off as small random numbers and were updated few times only. It should not be surprising if they are close to each other.

Playing with `wordDistance`

- Cases where things worked/didn't work out as expected
- Understanding what the numbers really mean and how to compare them.
- It only makes sense to compare relative distances between words.
  - $d(A, B)$ and $d(A, C)$
  - $d(A, B)$ and $\langle d(A, w) \rangle$, $\langle d(B, w) \rangle$
  - NOT $d(A, B)$ and $d(C, D)$
- Rare words will cluster near the origin because their weights started off as small random numbers and were updated few times only. It should not be surprising if they are close to each other.
- It is not correct to make a direct comparison of distances (or their differences, or ratios) across dimensions
- Δd of 0.1 means different things in $\mathbb{R}^8$ and $\mathbb{R}^{32}$
- In general, distances in high dimensional spaces are bigger and small differences mean a lot

Comments on learning in this scenario

Overfitting or underfitting ? Did it generalize well ?

- Early stopping does a good job of avoiding over-fitting
- Some examples of cases which you think can be considered overfitting
- Difference in training and validation cross entropies

Thanks!