Your name:

Your student ID:

Midterm for CSC 321
February 28 2013, 12:10pm – 1:00pm
Closed book.

This midterm has two sections, each of which is worth a total of 6 points. Answer all 6 questions in section A, and 3 of the 6 questions in section B.

Section A. Answer all 6 of these questions. Each is worth one mark. These are short questions. When we say "briefly explain", don't start writing a whole page of text. The main idea, in one or two sentences, is enough.

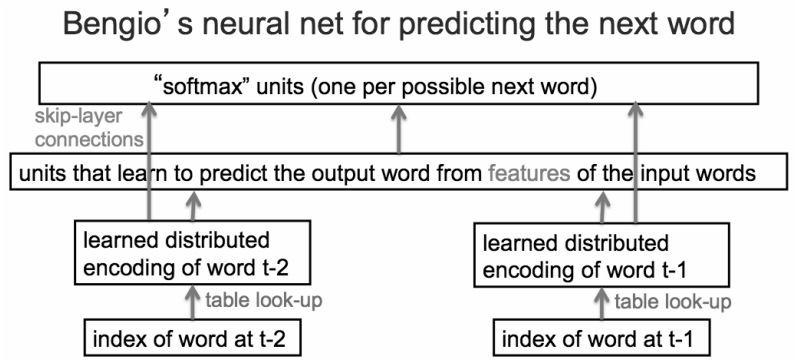
A1. Briefly explain the rprop method.

A2. Briefly explain how the rprop method can be adapted to work well with mini-batches.

A3. Give an example of a learning task where a group of 3 softmax output units is more appropriate than a group of 3 linear output units. It doesn't have to be a learning task that we've seen in class: be creative; use your imagination. Hint: you'd probably write something like "Learning to predict/estimate …, given …" (fill in the dots)

A4. Give an example of a learning task where a group of 3 linear output units is more appropriate than a group of 3 softmax output units. It doesn't have to be a learning task that we've seen in class: be creative; use your imagination. Hint: you'd probably write something like "Learning to predict/estimate …, given …" (fill in the dots)

A5. Is Bengio's language model a recurrent neural network or an autoregressive model? Briefly explain why. Below, as a reminder, is the lecture slide about that model.



A6. When, in general, is a distributed representation of data preferable over a localist representation?

Section B. Answer 3 of these 6 questions. Each of these is worth 2 marks. If you answer more than 3 of these 6 questions, your worst 3 answers will be used, so just don't do that. If you wrote something for a question and you later decide not to answer that question after all, cross out what you wrote and clearly write "don't mark this". Again, explaining the main idea briefly is better than writing something lengthy.

B1. 2 marks. If some of the input values (say a person's age in days) are typically much larger than some other input values (say the person's height in feet), neural network learning can suffer.

a) (1 mark) Briefly explain what the problem is.

b) (1 mark) Briefly explain how to prevent it.

B2. 2 marks.

a) (1 mark) Explain the "gradient plateau" problem that can arise with logistic hidden units. Be sure to describe not just the phenomenon, but also what makes it a problem.

b) (1 mark) Explain how rmsprop attempts to overcome the gradient plateau problem.

B3. 2 marks.

Which direction to change the parameters (weights) in, in neural network training, is not a trivial question. In fact, even the question "which direction would be the very best direction to change the parameters in?" is ambiguous, because there are different reasonable interpretations of what "best" could mean.

One direction that we could change the parameters in is the negative gradient of the loss function (negative because we want to reduce the loss).

a) (1 mark) Give one argument for calling that the "best" direction.

b) (1 mark) Give one argument for why that is often not the "best" direction.

B4. 2 marks. The perceptron learning procedure guarantees improvement, although there is a required condition for that guarantee.

a) (1 mark) What does this improvement consist of, i.e. what exactly gets better?

b) (1 mark) What is the required condition?

B5. 2 marks. Training a multilayer neural network with gradient descent, without momentum, in full-batch mode, also guarantees improvement. Here, too, there's a required condition for the guarantee.

a) (1 mark) What does this improvement consist of, i.e. what exactly gets better?

b) (1 mark) What is the required condition?

B6. 2 marks.

A large vocabulary can cause problems in language models. One such problem is that if we have a vocabulary of 100,000 words, we might need a softmax over those 100,000 output alternatives, and a 100,000-way softmax is not very practical. We studied a variety of solutions to this problem. One of them involves placing the output alternatives in a binary tree as the leaves, and then declaring the probability of a certain output to be the product of probabilities of choosing, at every level, the tree branch that leads to that particular output leaf. As usual, we train to optimize the log probability of the right output, and at test time we simply output the word with the greatest probability.

a) (1 mark) Briefly explain how this drastically reduces the required computation time at training time.

b) (1 mark) Does this also drastically reduce the required computation time at test time? Briefly explain your reasoning.