CS 224d: Assignment #1

Due date: 4/19 11:59 PM PST (You are allowed to use three (3) late days maximum for this assignment)

These questions require thought, but do not require long answers. Please be as concise as possible.

We encourage students to discuss in groups for assignments. However, each student must finish the problem set and programming assignment individually, and must turn in her/his assignment. We ask that you abide by the university Honor Code and that of the Computer Science department, and make sure that all of your submitted work is done by yourself

Please review any additional instructions posted on the assignment page at http://cs224d.stanford.edu/assignments.html. When you are ready to submit, please follow the instructions on the course website.

1 Softmax (10 points)

(a) (5 points) Prove that softmax is invariant to constant offsets in the input, that is, for any input vector $\boldsymbol{x}$ and any constant c,

$$\text{softmax}(\boldsymbol{x}) = \text{softmax}(\boldsymbol{x} + c)$$

where $\boldsymbol{x} + c$ means adding the constant c to every dimension of $\boldsymbol{x}$. Remember that

$$softmax(x)_i = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{1}$$

Note: In practice, we make use of this property and choose $c = -\max_i x_i$ when computing softmax probabilities for numerical stability (i.e. subtracting its maximum element from all elements of x).

(b) (5 points) Given an input matrix of N rows and d columns, compute the softmax prediction for each row. Write your implementation in `q1_softmax.py`. You may test by executing `python q1_softmax.py`.

Note: The provided tests are not exhaustive. Later parts of the assignment will reference this code so it is important to have a correct implementation. Your implementation should also be efficient and vectorized whenever possible. A non-vectorized implementation will not receive full credit!

2 Neural Network Basics (30 points)

(a) (3 points) Derive the gradients of the sigmoid function and show that it can be rewritten as a function of the function value (i.e. in some expression where only $\sigma(x)$, but not x, is present). Assume that the input x is a scalar for this question. Recall, the sigmoid function is

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2}$$

(b) (3 points) Derive the gradient with regard to the inputs of a softmax function when cross entropy loss is used for evaluation, i.e. find the gradients with respect to the softmax input vector $\boldsymbol{\theta}$, when the prediction is made by $\hat{\boldsymbol{y}} = \text{softmax}(\boldsymbol{\theta})$. Remember the cross entropy function is

$$CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = -\sum_i y_i \log(\hat{y}_i) \tag{3}$$

$\hat{y} = \text{Softmax}(\theta)$

$y_i = \frac{e^{\theta_i}}{\sum_j e^{\theta_j}}$

$CE(y, \hat{y}) = -\sum_i y_i \log \hat{y}_i$

$\frac{\partial(CE)}{\partial \hat{y}_i} = -\frac{y_i}{\hat{y}_i}$

$\frac{\partial(CE)}{\partial \theta_k} = \frac{\partial(CE)}{\partial \hat{y}_i} \cdot \frac{\partial \hat{y}_i}{\partial \theta_k}$

$= \frac{-y_i}{\hat{y}_i} \cdot \frac{\partial \hat{y}_i}{\partial \theta_k}$

$\hat{y}_i \rightarrow i^{th}$ element of $\hat{y}$

when $i = k$

$\frac{\partial \hat{y}_i}{\partial \theta_i} = \frac{\partial}{\partial \theta_i}\left(\frac{e^{\theta_i}}{\sum_j e^{\theta_j}}\right)$

$= \frac{e^{\theta_i}}{\sum_j e^{\theta_j}} - \frac{e^{\theta_i} \cdot e^{\theta_i}}{\left(\sum_j e^{\theta_j}\right)^2}$

$= \hat{y}_i - \hat{y}_i^2$

$= \hat{y}_i (1 - \hat{y}_i)$

when $i \neq k$

$\frac{\partial \hat{y}_i}{\partial \theta_k} = \frac{\partial}{\partial \theta_k}\left(\frac{e^{\theta_i}}{\sum_j e^{\theta_j}}\right)$

$= \frac{-e^{\theta_i} \cdot e^{\theta_j}}{\left(\sum e^{\theta_j}\right)^2} = -\hat{y}_i \hat{y}_k$

$\frac{\partial(CE)}{\partial \theta_k} = -y_i (1 - \hat{y}_k) \quad i = k$

$= y_i \hat{y}_k \quad i \neq k$

$i = k$ & $y_i = 1$

$\frac{\partial(CE)}{\partial \theta_j} = (\hat{y}_j - 1) \quad i = j$

$= \hat{y}_j$

$\frac{\partial(CE)}{\partial \theta} = \hat{y} - y$

$\delta_1 = \frac{\partial(CE)}{\partial \theta}$

$= \frac{\partial(CE)}{\partial z_2} = \hat{y} - y$

$\delta_2 = \frac{\partial(CE)}{\partial h} \cdot \frac{\partial z_2}{\partial h}$

$= \delta_1 \cdot W_2^T$

$\delta_3 = \frac{\partial(CE)}{\partial z_1} \cdot \frac{\partial h}{\partial h}$

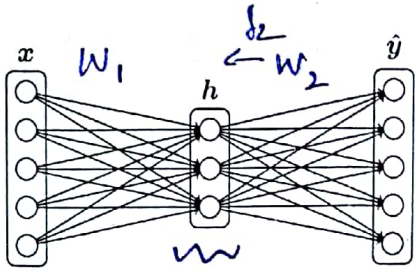
$= \delta_2 \circ \sigma'(z_1)$

$\frac{\partial(CE)}{\partial x} = \frac{\partial(CE)}{\partial z_1} \frac{\partial z_1}{\partial x}$

$= \delta_3 W_1^T$

where $\boldsymbol{y}$ is the one-hot label vector, and $\hat{\boldsymbol{y}}$ is the predicted probability vector for all classes. (*Hint: you might want to consider the fact many elements of $\boldsymbol{y}$ are zeros, and assume that only the k-th dimension of $\boldsymbol{y}$ is one.*)

(c) (6 points) Derive the gradients with respect to the *inputs* $\boldsymbol{x}$ to an one-hidden-layer neural network (that is, find $\frac{\partial J}{\partial \boldsymbol{x}}$ where J is the cost function for the neural network). The neural network employs sigmoid activation function for the hidden layer, and softmax for the output layer. Assume the one-hot label vector is $\boldsymbol{y}$, and cross entropy cost is used. (feel free to use $\sigma'(x)$ as the shorthand for sigmoid gradient, and feel free to define any variables whenever you see fit)



Recall that the forward propagation is as follows

$$h = \text{sigmoid}(\boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b}_1)$$

$$\hat{\boldsymbol{y}} = \text{softmax}(\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{b}_2)$$

Note that here we're assuming that the input vector (thus the hidden variables and output probabilities) is a row vector to be consistent with the programming assignment. When we apply the sigmoid function to a vector, we are applying it to each of the elements of that vector. $\boldsymbol{W}_i$ and $\boldsymbol{b}_i$ ($i = 1, 2$) are the weights and biases, respectively, of the two layers.

(d) (2 points) How many parameters are there in this neural network, assuming the input is D_x-dimensional, the output is D_y-dimensional, and there are H hidden units?

(e) (4 points) Fill in the implementation for the sigmoid activation function and its gradient in `q2_sigmoid.py`. Test your implementation using `python q2_sigmoid.py`. *Again, thoroughly test your code as the provided tests may not be exhaustive.*

(f) (4 points) To make debugging easier, we will now implement a gradient checker. Fill in the implementation for `gradcheck_naive` in `q2_gradcheck.py`. Test your code using `python q2_gradcheck.py`.

(g) (8 points) Now, implement the forward and backward passes for a neural network with one sigmoid hidden layer. Fill in your implementation in `q2_neural.py`. Sanity check your implementation with `q2_neural.py`.

3 word2vec (40 points + 5 bonus)

(a) (3 points) Assume you are given a predicted word vector $\boldsymbol{v}_c$ corresponding to the center word $\boldsymbol{c}$ for skipgram, and word prediction is made with the softmax function found in `word2vec` models

$$\hat{\boldsymbol{y}}_o = p(\boldsymbol{o} \mid \boldsymbol{c}) = \frac{\exp(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}{\sum_{w=1}^{W} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)} \tag{4}$$

where $\boldsymbol{w}$ denotes the w-th word and $\boldsymbol{u}_w$ ($w = 1, \ldots, W$) are the "output" word vectors for all words in the vocabulary. Assume cross entropy cost is applied to this prediction and word $\boldsymbol{o}$ is the expected word (the $\boldsymbol{o}$-th element of the one-hot label vector is one), derive the gradients with respect to $\boldsymbol{v}_c$.

$$J_i = -y_i \log \hat{y}_i \quad ; \quad \hat{y}_o = \frac{e^{u_o^T v_c}}{\sum_{w=1}^{W} e^{u_w^T v_c}}$$

$$\frac{\partial J_i}{\partial v_c} = \frac{\partial}{\partial v_c} \left(-u_o^T v_c + \log \sum_{w=1}^{W} e^{u_w^T v_c} \right)$$

$$= -u_i + \frac{\sum_{w=1}^{W} e^{u_w^T v_c} \cdot u_w}{\sum_{w=1}^{W} e^{u_w^T v_c}}$$

$$= -u_i + \sum_{w=1}^{W} \hat{y}_w u_w$$

$$\boxed{U(\hat{y} - y)}$$

Vectorized ??

$$\frac{\partial J_i}{\partial u_w} = \frac{\partial}{\partial u_w} \left(-u_i^T v_c + \log \sum_{w=1}^{W} e^{u_w^T v_c} \right)$$

w is not in context of c

$$= \frac{\partial}{\partial u_w} \left(\log \sum_{w=1}^{W} e^{u_w^T v_c} \right) = \frac{1}{\sum_{w=1}^{W} e^{u_w^T v_c}} \cdot e^{u_w^T v_c} \cdot v_c = \hat{y}_w v_c$$

w is in context of c

$$= \frac{\partial}{\partial u_w} \left(-u_i^T v_c + \log \sum_{w=1}^{W} e^{u_w^T v_c} \right)$$

$$= -v_c + \frac{1}{\sum_{w=1}^{W} e^{u_w^T v_c}} e^{u_w^T v_c} \cdot v_c = -v_c + \hat{y}_w v_c$$

$$= (\hat{y}_w - 1) v_c$$

$$\boxed{v_c (\hat{y} - y)^T}$$

Hint: It will be helpful to use notation from question 2. For instance, letting $\hat{\boldsymbol{y}}$ be the vector of softmax predictions for every word, $\boldsymbol{y}$ as the expected word vector, and the loss function

$$J_{softmax-CE}(\boldsymbol{o}, \boldsymbol{v}_c, \boldsymbol{U}) = CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) \tag{5}$$

where $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_W]$ is the matrix of all the output vectors. *Make sure you state the orientation of your vectors and matrices.*

(b) (3 points) As in the previous part, derive gradients for the "output" word vectors $\boldsymbol{u}_w$'s (including $\boldsymbol{u}_o$).

(c) (6 points) Repeat part (a) and (b) assuming we are using the negative sampling loss for the predicted vector $\boldsymbol{v}_c$, and the expected output word is $\boldsymbol{o}$. Assume that K negative samples (words) are drawn, and they are $\boldsymbol{1}, \cdots, \boldsymbol{K}$, respectively for simplicity of notation ($o \notin \{1, \ldots, K\}$). Again, for a given word, $\boldsymbol{o}$, denote its output vector as $\boldsymbol{u}_o$. The negative sampling loss function in this case is

$$J_{neg-sample}(\boldsymbol{o}, \boldsymbol{v}_c, \boldsymbol{U}) = -\log(\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)) - \sum_{k=1}^{K} \log(\sigma(-\boldsymbol{u}_k^\top \boldsymbol{v}_c)) \tag{6}$$

where $\sigma(\cdot)$ is the sigmoid function.

After you've done this, describe with one sentence why this cost function is much more efficient to compute than the softmax-CE loss (you could provide a speed-up ratio, i.e. the runtime of the softmax-CE loss divided by the runtime of the negative sampling loss).

Note: the cost function here is the negative of what Mikolov et al had in their original paper, because we are doing a minimization instead of maximization in our code.

(d) (8 points) Derive gradients for all of the word vectors for skip-gram and CBOW given the previous parts and given a set of context words $[\text{word}_{c-m}, \ldots, \text{word}_{c-1}, \text{word}_c, \text{word}_{c+1}, \ldots, \text{word}_{c+m}]$, where m is the context size. Denote the "input" and "output" word vectors for word_k as $\boldsymbol{v}_k$ and $\boldsymbol{u}_k$ respectively.

Hint: feel free to use $F(\boldsymbol{o}, \boldsymbol{v}_c)$ (where $\boldsymbol{o}$ is the expected word) as a placeholder for the $J_{softmax-CE}(\boldsymbol{o}, \boldsymbol{v}_c, \ldots)$ or $J_{neg-sample}(\boldsymbol{o}, \boldsymbol{v}_c, \ldots)$ cost functions in this part — you'll see that this is a useful abstraction for the coding part. That is, your solution may contain terms of the form $\frac{\partial F(o, v_c)}{\partial \ldots}$.

Recall that for skip-gram, the cost for a context centered around c is

$$J_{\text{skip-gram}}(\text{word}_{c-m \ldots c+m}) = \sum_{-m \le j \le m, j \neq 0} F(\boldsymbol{w}_{c+j}, \boldsymbol{v}_c) \tag{7}$$

where $\boldsymbol{w}_{c+j}$ refers to the word at the j-th index from the center.

CBOW is slightly different. Instead of using $\boldsymbol{v}_c$ as the predicted vector, we use $\hat{\boldsymbol{v}}$ defined below. For (a simpler variant of) CBOW, we sum up the input word vectors in the context

$$\hat{\boldsymbol{v}} = \sum_{-m \le j \le m, j \neq 0} \boldsymbol{v}_{c+j} \tag{8}$$

then the CBOW cost is

$$J_{\text{CBOW}}(\text{word}_{c-m \ldots c+m}) = F(\boldsymbol{w}_c, \hat{\boldsymbol{v}}) \tag{9}$$

Note: To be consistent with the $\hat{\boldsymbol{v}}$ notation such as for the code portion, for skip-gram $\hat{\boldsymbol{v}} = \boldsymbol{v}_c$.

(e) (12 points) In this part you will implement the word2vec models and train your own word vectors with stochastic gradient descent (SGD). First, write a helper function to normalize r ws of a matrix in `q3_word2vec.py`. In the same file, fill in the implementation for the softmax and negative sampling cost and gradient functions. Then, fill in the implementation of the cost and gradient functions for the skip-gram model. When you are done, test your implementation by running `python q3_word2vec.py`. *Note: If you choose not to implement CBOW (part h), simply remove the NotImplementedError so that your tests will complete.*

$$J = -\log \sigma(u_o^T v_c) - \sum_{k=1}^{K} \log \sigma(-u_k^T v_c)$$

$$\frac{\partial J}{\partial v_c} = -\frac{1 \cdot \sigma'(u_o^T v_c) \cdot u_o}{\sigma(u_o^T v_c)} - \sum_{k=1}^{K} \frac{1 \cdot \sigma'(-u_k^T v_c) u_k}{\sigma(-u_k^T v_c)}$$

$$\sigma'(x) = \sigma(x)(1 - \sigma(x))$$

$$= (\sigma(u_o^T v_c) - 1) u_o - \sum_{k=1}^{K} (\sigma(-u_k^T v_c) - 1) u_k$$

$u_o \rightarrow$ where o is in the o/p word

$$\frac{\partial J}{\partial u_o} = \sigma(u_o^T v_c - 1) v_c$$

where k is other sampling word

$$\frac{\partial J}{\partial u_k} = -(\sigma(-u_k^T v_c) - 1) v_c \qquad k = 1, 2, 3, \ldots K$$

$$\frac{\partial J_{skip\text{-}gram}(word_{c-m} \ldots c+m)}{\partial U} = \sum_{-m \le j \le m, j \neq 0} \frac{\partial F(w_{c+j}, v_c)}{\partial U}$$

$$\frac{\partial J_{skip\text{-}gram}(word_{c-m} \ldots c+m)}{\partial v_c} = \sum_{-m \le j \le m, j \neq 0} \frac{\partial F(w_{c+j}, v_c)}{\partial v_c}$$

$$\frac{\partial J_{skip\text{-}gram}(word_{c-m} \ldots c+m)}{\partial v_j} = 0 \qquad \text{for all } j \neq c$$

(f) (4 points) Complete the implementation for your SGD optimizer in `q3_sgd.py`. Test your implementation by running `python q3_sgd.py`.

(g) (4 points) Show time! Now we are going to load some real data and train word vectors with everything you just implemented! We are going to use the Stanford Sentiment Treebank (SST) dataset to train word vectors, and later apply them to a simple sentiment analysis task. There is no additional code to write for this part; just run `python q3_run.py`.

Note: The training process may take a long time depending on the efficiency of your implementation ***(an efficient implementation takes approximately an hour)****. Plan accordingly!*

When the script finishes, a visualization for your word vectors will appear. It will also be saved as `q3_word_vectors.png` in your project directory. **Include the plot in your homework write up.** Briefly explain in at most three sentences what you see in the plot.

(h) Extra credit (5 points) Implement the CBOW model in `q3_word2vec.py`. *Note: This part is optional but the gradient derivations for CBOW in part (d) are not!.*

4 Sentiment Analysis (20 points)

Now, with the word vectors you trained, we are going to perform a simple sentiment analysis. For each sentence in the Stanford Sentiment Treebank dataset, we are going to use the average of all the word vectors in that sentence as its feature, and try to predict the sentiment level of the said sentence. The sentiment level of the phrases are represented as real values in the original dataset, here we'll just use five classes:

"very negative", "negative", "neutral", "positive", "very positive"

which are represented by 0 to 4 in the code, respectively. For this part, you will learn to train a softmax regressor with SGD, and perform train/dev validation to improve generalization of your regressor.

(a) (10 points) Implement a sentence featurizer and softmax regression. Fill in the implementation in `q4_softmaxreg.py`. Sanity check your implementation with `python q4_softmaxreg.py`.

(b) (2 points) Explain in fewer than three sentences why we want to introduce regularization when doing classification (in fact, most machine learning tasks).

(c) (4 points) Fill in the hyperparameter selection code in `q4_sentiment.py` to search for the "optimal" regularization parameter. **What value did you select? Report your train, dev, and test accuracies. Justify your hyperparameter search methodology in at most one sentence.** *Note: you should be able to attain at least 30% accuracy on dev.*

(d) (4 points) Plot the classification accuracy on the train and dev set with respect to the regularization value, using a logarithmic scale on the x-axis. This should have been done automatically. **Include `q4_reg_acc.png` in your homework write up.** Briefly explain in at most three sentences what you see in the plot.

$$\frac{\partial J_{CBOW}(Word_{c-m} \ldots c_{+m})}{\partial U} = \frac{\partial F(w_c, \hat{v})}{\partial U}$$

$$\frac{\partial J_{CBOW}(Word_{c-m} \ldots c_{+m})}{\partial v_j} = \frac{\partial F(w_c, \hat{v})}{\partial \hat{v}} \quad j \in \{c-m \ldots c+m\}$$

$$\frac{\partial J_{CBOW}(Word_{c-m} \ldots c_{+m})}{\partial v_j} = 0 \quad j \notin \{c-m \ldots c+m\}$$