CS246: Mining Massive Data Sets **Spring 2021**

Problem Set 4

Please read the homework submission policies at http://cs246.stanford.edu.

1 Implementation of Feed-Forward Neural Network (30 points)

Here, you will implement a simple one-layer feed-forward neural network (perceptron) from scratch to solve a classification task. To recap, a fully connected layer can be represented as a function parameterized by weight $\mathbf{W}$ and bias b, such that for row vector $\mathbf{x}$:

$$f_{\mathbf{W},b}(\mathbf{x}) = \mathbf{x}\mathbf{W} + b \tag{1}$$

For classification tasks, one of the most common practice is to use a softmax function followed by a cross entropy loss. As shown in section 13.2.5 of the textbook, for a vector $\mathbf{x} = [x_1, x_2, \ldots, x_n]$, the softmax function is defined as:

$$\mu(\mathbf{x})_i = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{2}$$

$$\mu(\mathbf{x}) = \left[\frac{e^{x_1}}{\sum_j e^{x_j}}, \ldots, \frac{e^{x_n}}{\sum_j e^{x_j}}\right] \tag{3}$$

As shown in section 13.2.9 of the textbook, for target probability distribution $\mathbf{p} = [p_1, p_2, \ldots, p_n]$ and a different probability distribution $\mathbf{q} = [q_1, q_2, \ldots, q_n]$, the cross entropy $L(\mathbf{p}, \mathbf{q})$ is defined as:

$$L(\mathbf{p}, \mathbf{q}) = -\sum_{i=1}^{n} p_i \log q_i \tag{4}$$

Note that in supervised learning, the target probability distribution $\mathbf{p}$ is usually represented as a one-hot vector, where $p_i = 1$ shows that the corresponding input belongs to the i-th class with probability 1.

We can now represent our one-layer feed-forward neural network as a function f_N, such that for input vector $\mathbf{x}$ and its one-hot label vector $\mathbf{y}$:

$$f_N(\mathbf{x}) = \mu(f_{\mathbf{W},b}(\mathbf{x})) \tag{5}$$

The loss for this input is thus $L(\mathbf{y}, f_N(\mathbf{x}))$. For a dataset $\mathcal{D}$ of size N, the loss is:

$$L(\mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N} L(\mathbf{y_i}, f_N(\mathbf{x_i})) \tag{6}$$

In order to minimize the loss function, we need to obtain the gradient with respect to weight $\mathbf{W}$ and bias b. In the next questions you will derive these gradient expressions using the chain rule.

(a) [5 Points]

Task: What is $\frac{\partial L(\mathbf{y},\mathbf{q})}{\partial q_i}$, the gradient of cross entropy loss $L(\mathbf{y}, \mathbf{q})$ with respect to q_i?

(b) [5 Points]

Task: What is $\frac{\partial \mu(\mathbf{x})_j}{\partial x_i}$, the gradient of softmax function $\mu(\mathbf{x})_j$ with respect to x_i? You should consider both $i = j$ and $i \neq j$.

(c) [5 Points]

Task: What is $\frac{\partial f_{\mathbf{W},b}(\mathbf{x})}{\partial W_{ij}}$ and $\frac{\partial f_{\mathbf{W},b}(\mathbf{x})}{\partial b}$, the gradient of fully connected layer $f_{\mathbf{W},b}(\mathbf{x})$ with respect to the ij-th entry of weight $\mathbf{W}$ and bias b?

Now, you should be able to assemble the full gradient of weight and bias $\frac{\partial L(\mathcal{D})}{\partial W_{ij}}$, $\frac{\partial L(\mathcal{D})}{\partial b}$ using chain rule and the gradients you derived above.

(d) [15 Points]

Next, let's go ahead and implement the one-layer neural network. Implement both the forward pass $L(\mathcal{D})$ and the backward gradient $\frac{\partial L(\mathcal{D})}{\partial W_{ij}}$ and $\frac{\partial L(\mathcal{D})}{\partial b}$. For optimization, we will implement Minibatch gradient descent and compare different batch sizes. **For this problem, you are allowed to keep the dataset in memory, and you do not need to use Spark. You are also allowed to use any external library, but we encourage you to implement gradients from scratch to deepen your understanding.**

Mini batch gradient descent: Go through the dataset in batches of predetermined size and update the parameters as follows:

```
while convergence criteria not reached do
    Randomly pick $n = batch\_size$ random samples $\mathbf{x_k}$ from the training data
    for $k = 1, ..., n$ do
```

```
        Update $W_{ij} \leftarrow W_{ij} - \eta \frac{\partial L(\mathcal{D})}{\partial W_{ij}}$ for all i, j in $\mathbf{W}$
        Update $b \leftarrow b - \eta \frac{\partial L(\mathcal{D})}{\partial b}$
    end for
end while
```

where η is the learning rate and *batch_size* is the number of training samples considered in each batch.

For **convergence criteria**, you should stop learning when the cross entropy loss on **all data samples** is smaller than 0.4. You are encouraged to vectorize your loss function and gradient calculation, but it's not required.

You will compare the performance of 3 different batch sizes.

1. *batch_size* $= 1$. This is equivalent to stochastic gradient descent. Please use $\eta = 0.1$.
2. *batch_size* $= 20$. Please use $\eta = 0.1$.
3. *batch_size* $= N$, where N is the total number of data samples. This is equivalent to full batch gradient descent. Please use $\eta = 0.25$.

Run your implementation on the data set in **q1/data**. The data set contains the following files :

1. `features.txt` : Each line contains features (comma-separated values) for a single datapoint. It has 6414 datapoints (rows) and 122 features (columns).
2. `targets.txt` : Each line contains the target variable (y = 0 or 1) for the corresponding row in `features.txt`.

Task: Plot the value of cross entropy loss on **all data samples** $L(\mathcal{D})$ vs. the number of iterations (k) for all three batch sizes until **convergence**. Report the total wall-clock runtime (second) taken for convergence by each of the batch size. Comment on the plots and the time for convergence. What can you infer from them?

The diagram should have graphs from all the three batch sizes on the same plot.

Important Note

- You should initialize your weight $\mathbf{W}$ following a normal distribution with mean 0 and standard deviation 1. You should initialize your bias b to zeros. To make sure we have a fair comparison, please use the same initialization of weight and bias for all three runs.
- When computing the loss, remember to divide the total summed loss and gradient by the number of data points (i.e., batch size). This has been shown in Equation 6.

- You should calculate your cross entropy loss on **all data samples** for plotting and identifying convergence, not on the mini-batch of data samples.

As a sanity check, using batch size N should converge in 10-400 iterations, batch size 20 between 400-1000 iterations and batch size 1 between 1000-2500 iterations. However, the number of iterations may vary greatly due to the high randomness of this problem. If your implementation consistently takes longer iterations though, you may have a bug.

What to submit

(i) Equation for part (a)

(ii) Equation for part (b)

(iii) Two equations for part (c)

(iv) Plots for the cross entropy loss $L(\mathcal{D})$ vs. the number of iterations (k) for all three batch sizes 1, 20 and N. Total time taken for convergence by each of the batch size. Interpretation of plot and convergence times. [part (d)]

(v) Submit the code on Gradescope submission website. [part (d)]

2 Decision Tree Learning (20 points)

In this problem, we want to construct a decision tree to find out if a person will enjoy beer.

Definitions. Let there be k binary-valued attributes in the data.

We pick an attribute that maximizes the gain at each node:

$$G = I(D) - (I(D_L) + I(D_R)); \tag{7}$$

where D is the given dataset, and D_L and D_R are the sets on left and right hand-side branches after division. Ties may be broken arbitrarily.

There are three commonly used impurity measures used in binary decision trees: Entropy, Gini index, and Classification Error. In this problem, we use Gini index and define $I(D)$ as follows[1]:

$$I(D) = |D| \times \left(1 - \sum_i p_i^2\right),$$

where:

[1] As an example, if D has 10 items, with 4 positive items (*i.e.* 4 people who enjoy beer), and 6 negative items (*i.e.* 6 who do not), we have $I(D) = 10 \times (1 - (0.16 + 0.36))$.

- $|D|$ is the number of items in D;
- $1 - \sum_i p_i^2$ is the gini index;
- p_i is the probability distribution of the items in D, or in other words, p_i is the fraction of items that take value $i \in \{+, -\}$. Put differently, p_+ is the fraction of positive items and p_- is the fraction of negative items in D.

Note that this intuitively has the feel that the more evenly-distributed the numbers are, the lower the $\sum_i p_i^2$, and the larger the impurity.

(a) [10 Points]

Let $k = 3$. We have three binary attributes that we could use: "likes wine", "likes running" and "likes pizza". Suppose the following:

- There are 100 people in sample set, 40 of whom like beer and 60 who don't.
- Out of the 100 people, 50 like wine; out of those 50 people who like wine, 20 like beer.
- Out of the 100 people, 30 like running; out of those 30 people who like running, 20 like beer.
- Out of the 100 people, 80 like pizza; out of those 80 people who like pizza, 30 like beer.

Task: What are the values of G (defined in Equation 7) for wine, running and pizza attributes? Which attribute would you use to split the data at the root if you were to maximize the gain G using the gini index metric defined above?

(b) [10 Points]

Let's consider the following example:

- There are 100 attributes with binary values $a_1, a_2, a_3, \ldots, a_{100}$.
- Let there be one example corresponding to each possible assignment of 0's and 1's to the values $a_1, a_2, a_3 \ldots, a_{100}$. (Note that this gives us 2^{100} training examples.)
- Let the values taken by the target variable y depend on the values of a_1 for 99% of the datapoints. More specifically, of all the datapoints where $a_1 = 1$, let 99% of them are labeled $+$. Similarly, of all the datapoints where $a_1 = 0$, let 99% of them are labeled with $-$. (Assume that the values taken by y depend on $a_2, a_3, \ldots, a_{100}$ for fewer than 99% of the datapoints.)
- Assume that we build a complete binary decision tree (*i.e.*, we use values of all attributes).

Task: Explain what the decision tree will look like. (A one line explanation will suffice.) Also, in 2-3 sentences, identify what the desired decision tree for this situation should look like to avoid overfitting, and why.(The desired decision tree isn't necessarily a complete binary decision tree)

What to submit

(i) Values of G for wine, running and pizza attributes. [part (a)]

(ii) The attribute you would use for splitting the data at the root. [part (a)]

(iii) Explain what the decision tree looks like in the described setting. Explain how a decision tree should look like to avoid overfitting. (1-2 lines each) [part (b)]

3 Clustering Data Streams (20 points)

Introduction. In this problem, we study an approach for clustering massive data streams. We will study a framework for turning an approximate clustering algorithm into one that can work on data streams, *i.e.*, one which needs a small amount of memory and a small number of (actually, just one) passes over the data. As the instance of the clustering problem, we will focus on the k-means problem.

Definitions. Before going into further details, we need some definitions:

- The function $d : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^+$ denotes the Euclidean distance:
$$d(x, y) = ||x - y||_2.$$
- For any $x \in \mathbb{R}^p$ and $T \subset \mathbb{R}^p$, we define:
$$d(x, T) = \min_{z \in T}\{d(x, z)\}.$$
- Having subsets $S, T \subset \mathbb{R}^p$, and a weight function $w : S \to \mathbb{R}^+$, we define:
$$\text{cost}_w(S, T) = \sum_{x \in S} w(x) d(x, T)^2.$$
- Finally, if for all $x \in S$ we have $w(x) = 1$, we simply denote $\text{cost}_w(S, T)$ by $\text{cost}(S, T)$.

Reminder: k-means clustering. The k-means clustering problem is as follows: given a subset $S \subset \mathbb{R}^p$, and an integer k, find the set T (with $|T| = k$), which minimizes $\text{cost}(S, T)$. If a weight function w is also given, the k-means objective would be to minimize $\text{cost}_w(S, T)$, and we call the problem the weighted k-means problem.

Strategy for clustering data streams. We assume we have an algorithm ALG which is an α-approximate weighted k-means clustering algorithm (for some $\alpha > 1$). In other words, given any $S \subset \mathbb{R}^p$, $k \in \mathbb{N}$, and a weight function w, ALG returns a set $T \subset \mathbb{R}^p$, $|T| = k$, such that:

$$\text{cost}_w(S, T) \leq \alpha \min_{|T'|=k} \{\text{cost}_w(S, T')\}.$$

We will see how we can use ALG as a building block to make an algorithm for the k-means problem on data streams.

The basic idea here is that of divide and conquer: if S is a huge set that does not fit into main memory, we can read a portion of it that does fit into memory, solve the problem on this subset (*i.e.*, do a clustering on this subset), record the result (*i.e.*, the cluster centers and some corresponding weights, as we will see), and then read a next portion of S which is again small enough to fit into memory, solve the problem on this part, record the result, etc. At the end, we will have to combine the results of the partial problems to construct a solution for the main big problem (*i.e.*, clustering S).

To formalize this idea, we consider the following algorithm, which we denote as ALGSTR:

- Partition S into ℓ parts $S_1, \ldots, S_\ell$.
- For each $i = 1$ to ℓ, run ALG on S_i to get a set of k centers $T_i = \{t_{i1}, t_{i2}, \ldots, t_{ik}\}$, and assume $\{S_{i1}, S_{i2}, \ldots, S_{ik}\}$ is the corresponding clustering of S_i (*i.e.*, $S_{ij} = \{x \in S_i | \, d(x, t_{ij}) < d(x, t_{ij'}) \; \forall j' \neq j, 1 \leq j' \leq k\}$).
- Let $\widehat{S} = \bigcup_{i=1}^{\ell} T_i$, and define weights $w(t_{ij}) = |S_{ij}|$.
- Run ALG on $\widehat{S}$ with weights w, to get k centers T.
- Return T.

Now, we analyze this algorithm. Assuming $T^* = \{t_1^*, \ldots, t_k^*\}$ to be the optimal k-means solution for S (that is, $T^* = \text{argmin}_{|T'|=k}\{\text{cost}(S, T')\}$), we would like to compare $\text{cost}(S, T)$ (where T is returned by ALGSTR) with $\text{cost}(S, T^*)$.

A small fact might be useful in the analysis below: for any $(a, b) \in \mathbb{R}^+$ we have:

$$(a + b)^2 \leq 2a^2 + 2b^2.$$

(a) [5pts]

First, we show that the cost of the final clustering can be bounded in terms of the total cost of the intermediate clusterings:

Task: Prove that:

$$\text{cost}(S, T) \leq 2 \cdot \text{cost}_w(\widehat{S}, T) + 2 \sum_{i=1}^{\ell} \text{cost}(S_i, T_i).$$

Hint: You might want to use Triangle Inequality for Euclidean distance d.

(b) [5pts]

So, to bound the cost of the final clustering, we can bound the terms on the right hand side of the inequality in part (a). Intuitively speaking, we expect the second term to be small compared to $\text{cost}(S, T^*)$, because T^* only uses k centers to represent the data set (S), while the T_i's, in total, use $k\ell$ centers to represent the same data set (and $k\ell$ is potentially much bigger than k). We show this formally:

Task: Prove that:

$$\sum_{i=1}^{\ell} \text{cost}(S_i, T_i) \leq \alpha \cdot \text{cost}(S, T^*).$$

(c) [10pt]

Prove that ALGSTR is a $(4\alpha^2 + 6\alpha)$-approximation algorithm for the k-means problem.

Task: Prove that:

$$\text{cost}(S, T) \leq (4\alpha^2 + 6\alpha) \cdot \text{cost}(S, T^*).$$

Hint: You might want to first prove two useful facts, which help bound the first term on the right hand side of the inequality in part (a):

$$\text{cost}_w(\widehat{S}, T) \leq \alpha \cdot \text{cost}_w(\widehat{S}, T^*).$$

$$\text{cost}_w(\widehat{S}, T^*) \leq 2 \sum_{i=1}^{\ell} \text{cost}(S_i, T_i) + 2 \cdot \text{cost}(S, T^*).$$

Additional notes: We have shown above that ALGSTR is a $(4\alpha^2 + 6\alpha)$-approximation algorithm for the k-means problem. Clearly, $4\alpha^2 + 6\alpha > \alpha$, so ALGSTR has a somewhat worse approximation guarantee than ALG (with which we started). However, ALGSTR is better suited for the streaming application, as not only it takes just one pass over the data, but also it needs a much smaller amount of memory.

Assuming that ALG needs $\Theta(n)$ memory to work on an input set S of size n (note that just representing S in memory will need $\Omega(n)$ space), if we partitioning S into $\sqrt{n/k}$ equal parts, ALGSTR can work with only $O(\sqrt{nk})$ memory. (Like in the rest of the problem, k represents the number of clusters per partition.)

Note that for typical values of n and k, assuming $k \ll n$, we have $\sqrt{nk} \ll n$. For instance, with $n = 10^6$, and $k = 100$, we have $\sqrt{nk} = 10^4$, which is 100 times smaller than n.

What to submit

(a) Proof that $\text{cost}(S, T) \leq 2 \cdot \text{cost}_w(\widehat{S}, T) + 2\sum_{i=1}^{\ell} \text{cost}(S_i, T_i)$.

(b) Proof that $\sum_{i=1}^{\ell} \text{cost}(S_i, T_i) \leq \alpha \cdot \text{cost}(S, T^*)$.

(c) Proof that $\text{cost}(S, T) \leq (4\alpha^2 + 6\alpha) \cdot \text{cost}(S, T^*)$.

4 Data Streams (30 points)

In this problem, we study an approach to approximating the frequency of occurrences of different items in a data stream. Assume $S = \langle a_1, a_2, \ldots, a_t \rangle$ is a data stream of items from the set $\{1, 2, \ldots, n\}$. Assume for any $1 \leq i \leq n$, $F[i]$ is the number of times i has appeared in S. We would like to have good approximations of the values $F[i]$ $(1 \leq i \leq n)$ at all times.

A simple way to do this is to just keep the counts for each item $1 \leq i \leq n$ separately. However, this will require $\mathcal{O}(n)$ space, and in many applications (e.g., think online advertising and counts of user's clicks on ads) this can be prohibitively large. We see in this problem that it is possible to approximate these counts using a much smaller amount of space. To do so, we consider the algorithm explained below.

Strategy. The algorithm has two parameters $\delta, \epsilon > 0$. It picks $\lceil \log \frac{1}{\delta} \rceil$ independent hash functions:

$$\forall j \in \left[\!\left[1; \left\lceil \log \frac{1}{\delta} \right\rceil \right]\!\right], \quad h_j : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, \left\lceil \frac{e}{\epsilon} \right\rceil\},$$

where log denotes natural logarithm. Also, it associates a count $c_{j,x}$ to any $1 \leq j \leq \lceil \log \frac{1}{\delta} \rceil$ and $1 \leq x \leq \lceil \frac{e}{\epsilon} \rceil$. In the beginning of the stream, all these counts are initialized to 0. Then, upon arrival of each a_k $(1 \leq k \leq t)$, each of the counts $c_{j,h_j(a_k)}$ $(1 \leq j \leq \lceil \log \frac{1}{\delta} \rceil)$ is incremented by 1.

For any $1 \leq i \leq n$, we define $\tilde{F}[i] = \min_j \{c_{j,h_j(i)}\}$. We will show that $\tilde{F}[i]$ provides a good approximation to $F[i]$.

Memory cost. Note that this algorithm only uses $\mathcal{O}\left(\frac{1}{\epsilon} \log \frac{1}{\delta}\right)$ space.

Properties. A few important properties of the algorithm presented above:

- For any $1 \leq i \leq n$:
$$\tilde{F}[i] \geq F[i].$$
- For any $1 \leq i \leq n$ and $1 \leq j \leq \lceil \log(\frac{1}{\delta}) \rceil$:
$$\mathsf{E}\left[c_{j,h_j(i)}\right] \leq F[i] + \frac{\epsilon}{e}(t - F[i]).$$

(a) [10 Points]

Prove that:

$$\mathsf{Pr}\left[\tilde{F}[i] \leq F[i] + \epsilon t\right] \geq 1 - \delta.$$

Hint: Use Markov inequality and the property of independence of hash functions.

Based on the proof in part (a) and the properties presented earlier, it can be inferred that $\tilde{F}[i]$ is a good approximation of $F[i]$ for any item i such that $F[i]$ is not very small (compared to t). In many applications (*e.g.*, when the values $F[i]$ have a heavy-tail distribution), we are indeed only interested in approximating the frequencies for items which are not too infrequent. We next consider one such application.

(b) [20 Points]

Warning. This implementation question requires substantial computation time Python implementation reported to take 15min - 1 hour. Therefore, we advise you to start early.

Dataset. The dataset in **q4/data** contains the following files:

1. `words_stream.txt` Each line of this file is a number, corresponding to the ID of a word in the stream.
2. `counts.txt` Each line is a pair of numbers separated by a tab. The first number is an ID of a word and the second number is its associated exact frequency count in the stream.
3. `words_stream_tiny.txt` and `counts_tiny.txt` are smaller versions of the dataset above that you can use for debugging your implementation.
4. `hash_params.txt` Each line is a pair of numbers separated by a tab, corresponding to parameters a and b which you may use to define your own hash functions (See explanation below).

Instructions. Implement the algorithm and run it on the dataset with parameters $\delta = e^{-5}, \epsilon = e \times 10^{-4}$. (Note: with this choice of δ you will be using 5 hash functions - the 5 pairs (a, b) that you'll need for the hash functions are in `hash_params.txt`). Then for each distinct word i in the dataset, compute the relative error $E_r[i] = \frac{\tilde{F}[i]-F[i]}{F[i]}$ and plot these values as a function of the exact word frequency $\frac{F[i]}{t}$. (**You do not have to implement the algorithm in Spark.**)

The plot should use a logarithm scale both for the x and the y axes, and there should be ticks to allow reading the powers of 10 (e.g. 10^{-1}, 10^{0}, 10^{1} etc...). The plot should have a title, as well as the x and y axes. The exact frequencies $F[i]$ should be read from the counts file. Note that words of low frequency can have a very large relative error. That is not a bug in your implementation, but just a consequence of the bound we proved in question (a).

Answer the following question by reading values from your plot: What is an approximate condition on a word frequency in the document to have a relative error below $1 = 10^0$?

Hash functions. You may use the following hash function (see example pseudo-code), with $p = 123457$, a and b values provided in the hash params file and `n_buckets` (which is equivalent to $\lceil \frac{e}{\epsilon} \rceil$) chosen according to the specification of the algorithm. In the provided file, each line gives you a, b values to create one hash function.

```
# Returns hash(x) for hash function given by parameters a, b, p and n_buckets
def hash_fun(a, b, p, n_buckets, x)
{
y = x [modulo] p
hash_val = (a*y + b) [modulo] p
return hash_val [modulo] n_buckets
}
```

Note: This hash function implementation produces outputs of value from 0 to (`n_buckets` − 1), which is different from our specification in the **Strategy** part. You can either keep the range as $\{0, ..., \texttt{n\_buckets} - 1\}$, or add 1 to the hash result so the value range becomes $\{1, ..., \texttt{n\_buckets}\}$, as long as you stay consistent within your implementation.

What to submit

(i) Proof that $\mathsf{Pr}\left[\tilde{F}[i] \leq F[i] + \epsilon t\right] \geq 1 - \delta$. [part (a)]

(ii) Log-log plot of the relative error as a function of the frequency. Answer for which word frequencies is the relative error below 1. [part (b)]

(iii) Submit the code on Gradescope submission site. [part (b)]