CS246: Mining Massive Data Sets **Winter 2017**

Problem Set 2

Due 11:59pm February 9, 2017

Only one late period is allowed for this homework (11:59pm 2/14).

General Instructions

Submission instructions: These questions require thought but do not require long answers. Please be as concise as possible. You should submit your answers as a writeup in PDF format via GradeScope and code via the Snap submission site.

Submitting writeup: Prepare answers to the homework questions into a single PDF file and submit it via http://gradescope.com. Make sure that the answer to each question is on a *separate page*. On top of each page write the number of the question you are answering. Please find the cover sheet and the recommended templates located here:
http://web.stanford.edu/class/cs246/homeworks/hw2/hw2_template.tex
http://web.stanford.edu/class/cs246/homeworks/hw2/hw2_template.pdf

Not including the cover sheet in your submission will result in a 2-point penalty. It is also important to tag your answers correctly on Gradescope. We will deduct 5/N points for each incorrectly tagged subproblem (where N is the number of subproblems). This means you can lose up to 5 points for incorrect tagging.

Submitting code: Upload your code at http://snap.stanford.edu/submit. Put all the code for a single question into a single file and upload it.

Questions

1 Dead ends in PageRank computations (20 points) [Sachin, Naveen, Yixin W]

Let the *matrix of the Web* M be an n-by-n matrix, where n is the number of Web pages. The entry m_{ij} in row i and column j is 0, unless there is an arc from node (page) j to node i. In that case, the value of m_{ij} is $1/k$, where k is the number of arcs (links) out of node j. Notice that if node j has $k > 0$ arcs out, then column j has k values of $1/k$ and the rest 0's. If node j is a *dead end* (i.e., it has zero arcs out), then column j is all 0's.

Let $\mathbf{r} = [r_1, r_2, \ldots, r_n]^T$ be (an estimate of) the PageRank vector; that is, r_i is the estimate of the PageRank of node i. Define $w(\mathbf{r})$ to be the sum of the components of $\mathbf{r}$; that is

$w(\mathbf{r}) = \sum_{i=1}^{n} r_i$.

In one iteration of the PageRank algorithm, we compute the next estimate $\mathbf{r}'$ of the PageRank as: $\mathbf{r}' = M\mathbf{r}$. Specifically, for each i we compute $r'_i = \sum_{j=1}^{n} M_{ij} r_j$.

(a) [6pts]

Suppose the Web has no dead ends. Prove that $w(\mathbf{r}') = w(\mathbf{r})$.

(b) [7pts]

Suppose there are still no dead ends, but we use a teleportation probability of $1 - \beta$, where $0 < \beta < 1$. The expression for the next estimate of r_i becomes $r'_i = \beta \sum_{j=1}^{n} M_{ij} r_j + (1-\beta)/n$. Under what circumstances will $w(\mathbf{r}') = w(\mathbf{r})$? Prove your conclusion.

(c) [7pts]

Now, let us assume a teleportation probability of $1 - \beta$ in addition to the fact that there are one or more dead ends. Call a node "dead" if it is a dead end and "live" if not. Assume $w(\mathbf{r}) = 1$. At each iteration, we will distribute equally to each node the sum of:

1. $(1 - \beta) r_j$ if node j is live.
2. r_j if node j is dead.

Write the equation for r'_i in terms of β, M, and $\mathbf{r}$. Then, prove that $w(\mathbf{r}')$ is also 1.

What to submit

(i) Proof [1(a)]

(ii) Condition for $w(\mathbf{r}') = w(\mathbf{r})$ and Proof [1(b)]

(iii) Equation for r'_i and Proof [1(c)]

2 Singular Value Decomposition and Principal Component Analysis (25 points) [Yixin C, Junwei, Leon]

In this problem we will explore the relationship between two of the most popular dimensionality-reduction techniques, SVD and PCA at a basic conceptual level. Before we proceed with the question itself, let us briefly recap the SVD and PCA techniques and a few important observations:

- First, recall that the eigenvalue decomposition of a *real*, *symmetric*, and *square matrix* B (of size $d \times d$) can be written as the following product:

$$B = Q\Lambda Q^{\mathrm{T}}$$

where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ contains the eigenvalues of B (which are always real) along its main diagonal and Q is an orthogonal matrix containing the eigenvectors of B as its columns.

- Principal Component Analysis (PCA): Given a data matrix M (of size $p \times q$), PCA involves the computation of the eigenvectors of MM^T or M^TM. The matrix of these eigenvectors can be thought of as a rigid rotation in a high dimensional space. When you apply this transformation to the original data, the axis corresponding to the principal eigenvector is the one along which the points are most spread out, More precisely, this axis is the one along which the variance of the data is maximized. Put another way, the points can best be viewed as lying along this axis, with small deviations from this axis. Likewise, the axis corresponding to the second eigenvector (the eigenvector corresponding to the second-largest eigenvalue) is the axis along which the variance of distances from the first axis is greatest, and so on.

- Singular Value Decomposition (SVD): SVD involves the decomposition of a data matrix M (of size $p \times q$) into a product: $U\Sigma V^T$ where U (of size $p \times k$) and V (of size $q \times k$) are column-orthonormal matrices[1] and Σ (of size $k \times k$) is a diagonal matrix. The entries along the diagonal of Σ are referred to as singular values of M. The key to understanding what SVD offers is in viewing the r columns of U, Σ, and V as representing concepts that are hidden in the original matrix M.

For answering the questions below, let us define a matrix M (of size $p \times q$) and let us assume this matrix corresponds to a dataset with p data points and q dimensions.

(a) [3 points]

Are the matrices MM^T and M^TM symmetric, square and real? Explain.

(b) [5 points]

Prove that the eigenvalues of MM^T are the same as that of M^TM. Are their eigenvectors the same?

(c) [2 points]

Given that we now understand certain properties of M^TM, write an expression for M^TM in terms of Q, Q^T and Λ where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ contains the eigenvalues of M^TM along

[1] A matrix $U \in \mathbb{R}^{p \times q}$ is column-orthonormal if and only if $U^TU = I$ where I denotes the identity matrix

its main diagonal and Q is an orthogonal matrix containing the eigenvectors of M^TM as its columns?
Hint: Check the definition of eigenvalue decomposition provided in the beginning of the question to see if it is applicable.

(d) [5 points]

SVD decomposes the matrix M into the product $U\Sigma V^T$ where U and V are column-orthonormal and Σ is a diagonal matrix. Given that $M = U\Sigma V^T$, write a simplified expression for M^TM in terms of V, V^T and Σ?

(e) [10 points]

In this question, let us experimentally test if SVD decomposition of M actually provides us the eigenvectors (PCA dimensions) of M^TM. We strongly recommend students to use Python and suggested functions for this exercise.[2] Initialize matrix M as follows:

$$M = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix}$$

- Compute the SVD of M (*Use scipy.linalg.svd function in Python and set the argument* `full_matrices` *to False*). The function returns values corresponding to U, Σ and V^T. What are the values returned for U, Σ and V^T? *Note: Make sure that the first element of the returned array Σ has a greater value than the second element.*

- Compute the eigenvalue decomposition of M^TM (*Use scipy.linalg.eigh function in Python*). The function returns two parameters: a list of eigenvalues (let us call this list *Evals*) and a matrix whose columns correspond to the eigenvectors of the respective eigenvalues (let us call this matrix *Evecs*). Sort the list *Evals* in descending order such that the largest eigenvalue appears first in the list. Also, re-arrange the columns in *Evecs* such that the eigenvector corresponding to the largest eigenvalue appears in the first column of *Evecs*. What are the values of *Evals* and *Evecs* (after the sorting and re-arranging process)?

- Based on the experiment and your derivations in part (c) and (d), do you see any correspondence between V produced by SVD and the matrix of eigenvectors *Evals* (after the sorting and re-arranging process) produced by eigenvalue decomposition? If so, what is it?
Note: The function scipy.linalg.svd returns V^T (not V).

[2] Other implementations of SVD and PCA might give slightly different results. Besides, you will just need fewer than five python commands to answer this entire question

- Based on the experiment and the expressions obtained in part (c) and part (d) for M^TM, what is the relationship (if any) between the eigenvalues of M^TM and the singular values of M? Explain.
  Note: The entries along the diagonal of Σ (part (e)) are referred to as singular values of M. The eigenvalues of M^TM are captured by the diagonal elements in Λ (part (d))

What to submit:

(i) Written solutions to questions 2(a) to 2(e) with explanations wherever required

(ii) Upload the code via Snap submission site

3 Implementing PageRank and HITS (25 points) [Nihit, Luda, Jessica]

In this problem, you will learn how to implement the PageRank and HITS algorithms. You will be experimenting with a small randomly generated graph (assume graph has no dead-ends) provided at http://snap.stanford.edu/class/cs246-data/graph.txt.

It has $n = 100$ nodes (numbered $1, 2, \ldots, 100$), and $m = 1024$ edges, 100 of which form a directed cycle (through all the nodes) which ensures that the graph is connected. It is easy to see that the existence of such a cycle ensures that there are no dead ends in the graph. There may be multiple edges between a pair of nodes, your program should handle these instead of ignoring them. The first column in `graph.txt` refers to the source node, and the second column refers to the destination node.

(a) PageRank Implementation [12 points]

Assume the directed graph $G = (V, E)$ has n nodes (numbered $1, 2, \ldots, n$) and m edges, all nodes have positive out-degree, and $M = [M_{ji}]_{n\times n}$ is a an $n \times n$ matrix as defined in class such that for any $i, j \in [\![1, n]\!]$:

$$M_{ji} = \begin{cases} \frac{1}{\deg(i)} & \text{if } (i \to j) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

Here, $\deg(i)$ is the number of outgoing edges of node i in G. By the definition of PageRank, assuming $1 - \beta$ to be the teleport probability, and denoting the PageRank vector by the column vector r, we have the following equation:

$$r = \frac{1-\beta}{n}\mathbf{1} + \beta M r, \tag{1}$$

where $\mathbf{1}$ is the $n \times 1$ vector with all entries equal to 1.

Based on this equation, the iterative procedure to compute PageRank works as follows:

1. Initialize: $r^{(0)} = \frac{1}{n}\mathbf{1}$
2. For i from 1 to k, iterate: $r^{(i)} = \frac{1-\beta}{n}\mathbf{1} + \beta M r^{(i-1)}$

Run the aforementioned iterative process for 40 iterations (assuming $\beta = 0.8$) and obtain the PageRank vector r. Compute the following:

- List the top 5 node ids with the highest PageRank scores.
- List the bottom 5 node ids with the lowest PageRank scores.

(b) HITS Implementation [13 points]

Assume the directed graph $G = (V, E)$ has n nodes (numbered $1, 2, \ldots, n$) and m edges, all nodes have non-negative out-degree, and $L = [L_{ij}]_{n\times n}$ is a an $n \times n$ matrix referred to as the *link matrix* such that for any $i, j \in [\![1, n]\!]$:

$$L_{ij} = \begin{cases} 1 & \text{if } (i \to j) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

Given the link matrix L and some scaling factors λ, μ, the hubbiness vector h and the authority vector a can be expressed using the equations:

$$h = \lambda L a, a = \mu L^T h \tag{2}$$

where $\mathbf{1}$ is the $n \times 1$ vector with all entries equal to 1.

Based on this equation, the iterative method to compute h and a is as follows:

1. Initialize h with a column vector (of size $n \times 1$) of all 1's.
2. Compute $a = L^T h$ and scale so that the largest value in the vector a has value 1.
3. Compute $h = La$ and scale so that the largest value in the vector h has value 1.
4. Go to step 2.

Repeat the iterative process for 40 iterations, assume that $\lambda = 1, \mu = 1$ and then obtain the hubbiness and authority scores of all the nodes (pages). Compute the following:

- List the 5 node ids with the highest hubbiness score.
- List the 5 node ids with the lowest hubbiness score.
- List the 5 node ids with the highest authority score.
- List the 5 node ids with the lowest authority score.

What to submit

(i) List 5 node ids with the highest and least PageRank scores [3(a)]

(ii) List 5 node ids with the highest and least hubbiness and authority scores [3(b)]

(iii) Upload all the code to the snap submission site [3(a) & 3(b)]

4 k-means on MapReduce (30 points) [Rishabh, Vinaya, Michael, Anthony]

Note: This problem requires substantial computing time. Don't start it at the last minute.

This problem will help you understand the nitty gritty details of implementing clustering algorithms on Hadoop. In addition, this problem will also help you understand the impact of using various distance metrics and initialization strategies in practice. Let us say we have a set $\mathcal{X}$ of n data points in the d-dimensional space $\mathbb{R}^d$. Given the number of clusters k and the set of k centroids $\mathcal{C}$, we now proceed to define various distance metrics and the corresponding cost functions that they minimize.

Euclidean distance Given two points A and B in d dimensional space such that $A = [a_1, a_2 \cdots a_d]$ and $B = [b_1, b_2 \cdots b_d]$, the Euclidean distance between A and B is defined as:

$$||a - b|| = \sqrt{\sum_{i=1}^{d} (a_i - b_i)^2} \tag{3}$$

The corresponding cost function ϕ that is minimized when we assign points to clusters using the Euclidean distance metric is given by:

$$\phi = \sum_{x \in \mathcal{X}} \min_{c \in \mathcal{C}} ||x - c||^2 \tag{4}$$

Manhattan distance Given two random points A and B in d dimensional space such that $A = [a_1, a_2 \cdots a_d]$ and $B = [b_1, b_2 \cdots b_d]$, the Manhattan distance between A and B is defined as:

$$|a - b| = \sum_{i=1}^{d} |a_i - b_i| \tag{5}$$

The corresponding cost function ψ that is minimized when we assign points to clusters using the Manhattan distance metric is given by:

$$\psi = \sum_{x \in \mathcal{X}} \min_{c \in \mathcal{C}} |x - c| \tag{6}$$

Iterative k-Means Algorithm: We learned the basic k-Means algorithm in class which is as follows: k centroids are initialized, each point is assigned to the nearest centroid and the centroids are recomputed based on the assignments of points to clusters. In practice, the above steps are run for several iterations. We present the resulting iterative version of k-Means in Algorithm 1.

Algorithm 1 Iterative k-Means Algorithm

```
1: procedure ITERATIVE k-MEANS
2:     Select k points as initial centroids of the k clusters.
3:     for iterations := 1 to MAX_ITER do
4:         for each point p in the dataset do
5:             Assign point p to the cluster with the closest centroid
6:         end for
7:         for each cluster c do
8:             Recompute the centroid of c as the mean of all the data points assigned to c
9:         end for
10:    end for
11: end procedure
```

Iterative k-Means clustering on Hadoop: Implement iterative k-means using MapReduce where a single step of MapReduce completes one iteration of the k-means algorithm. So, to run k-means for i iterations, you will have to run a sequence of i MapReduce jobs.

Please use the dataset at http://snap.stanford.edu/class/cs246-data/hw2-q4-kmeans.zip for this problem.

The zip has 4 files:

1. `data.txt` contains the dataset which has 4601 rows and 58 columns. Each row is a document represented as a 58 dimensional vector of features. Each component in the vector represents the importance of a word in the document.

2. `c1.txt` contains k initial cluster centroids. These centroids were chosen by selecting $k = 10$ random points from the input data.

3. `c2.txt` contains initial cluster centroids which are as far apart as possible. (You can do this by choosing 1st centroid c1 randomly, and then finding the point c2 that is farthest from c1, then selecting c3 which is farthest from c1 and c2, and so on).

Set number of iterations (`MAX_ITER`) to 20 and number of clusters k to 10 for all the experiments carried out in this question.

*Hint about **job chaining**:*

We need to run a sequence of Hadoop jobs where the output of one job will be the input for the next one. There are multiple ways to do this and you are free to use any method you are

comfortable with. One simple way to handle such a multistage job is to configure the output path of the first job to be the input path of the second and so on.

The following pseudo code demonstrates job chaining.

```
var inputDir
var outputDir
var centroidDir

for i in no-of-iterations (
     Configure job here with all params
     Set job input directory = inputDir
     Set job output directory = outputDir + i
     Run job
     centroidDir =  outputDir + i
)
```

You will also need to share the location of the centroid file with the mapper. There are many ways to do this and you can use any method you find suitable. One way is to use the Hadoop Configuration object. You can set it as a property in the Configuration object and retrieve the property value in the Mapper setup function.

For more details see :

1. http://hadoop.apache.org/docs/r1.0.4/api/org/apache/hadoop/conf/Configuration.html#set(java.lang.String,java.lang.String)
2. http://hadoop.apache.org/docs/r1.0.4/api/org/apache/hadoop/conf/Configuration.html#get(java.lang.String)
3. http://hadoop.apache.org/docs/r1.0.4/api/org/apache/hadoop/mapreduce/Mapper.html#setup(org.apache.hadoop.mapreduce.Mapper.Context)

(a) Exploring initialization strategies with Euclidean distance [15 pts]

1. **[10 pts]** Using the Euclidean distance (refer to Equation 3) as the distance measure, compute the cost function $\phi(i)$ (refer to Equation 4) for every iteration i. This means that, for your first MapReduce job iteration, you'll be computing the cost function using the initial centroids located in one of the two text files. Run the k-means on `data.txt` using `c1.txt` and `c2.txt`. Generate a graph where you plot the cost function $\phi(i)$ as a function of the number of iterations i=1..20 for `c1.txt` and also for `c2.txt`.

   (Hint: Note that you do not need to write a separate MapReduce job to compute $\phi(i)$. You can just incorporate the computation of $\phi(i)$ into the Mapper/Reducer.)

2. **[5 pts]** What is the percentage change in cost after 10 iterations of the K-Means algorithm when the cluster centroids are initialized using `c1.txt` vs. `c2.txt` and the distance metric being used is Euclidean distance? Is random initialization of k-means using `c1.txt` better than initialization using `c2.txt` in terms of cost $\phi(i)$? Explain your reasoning.

(b) Exploring initialization strategies with Manhattan distance [15 pts]

1. **[10 pts]** Using the Manhattan distance metric (refer to Equation 5) as the distance measure, compute the cost function $\psi(i)$ (refer to Equation 6) for every iteration i. This means that, for your first MapReduce job iteration, you'll be computing the cost function using the initial centroids located in one of the two text files. Run the k-means on `data.txt` using `c1.txt` and `c2.txt`. Generate a graph where you plot the cost function $\psi(i)$ as a function of the number of iterations i=1..20 for `c1.txt` and also for `c2.txt`.

   (Hint: This problem can be solved in a similar manner to that of part (a))

2. **[5 pts]** What is the percentage change in cost after 10 iterations of the K-Means algorithm when the cluster centroids are initialized using `c1.txt` vs. `c2.txt` and the distance metric being used is Manhattan distance? Is random initialization of k-means using `c1.txt` better than initialization using `c2.txt` in terms of cost $\psi(i)$? Explain your reasoning.

What to submit:

(i) Upload the code for 4(a) and 4(b) to Snap submission site

(ii) A plot of cost vs. iteration for two initialization strategies for 4(a)

(iii) Percentage improvement values and your explanation for 4(a)

(iv) A plot of cost vs. iteration for two initialization strategies for 4(b)

(v) Percentage improvement values and your explanation for 4(b)