PCA Documentation

Vincent Xue

June 10, 2011

1 Steps

matrix.txt - A text file that has the matrix
matrix.seqFile - A hadoop formated sequence file.

1. Load text file into Sequence File Format.

```
hutils createVectorFromTxt -i <matrix.txt> -o <matrix.seqFile>
```

2. Make sure that the matrix.seqFile is in a new folder.

```
hadoop dfs -mkdir newFolder
hadoop dfs -mv matrix.seqFile newFolder
```

3. Run the pcaJob

```
hutils pcaJob
```

   - A - The starting matrix
   - $rows$ - The number or rows in the starting matrix
   - $cols$ - The number of columns in the starting matrix
   - A_n - Matrix A normalized per row
   - A_n^T - Matrix A normalized per row and tranposed
   - r -rank- the number of raw eigen vectors the SVD produces. This should be greater than n.
   - n - the number of eigen vectors desired from the top.
   - E_n - the top n eigen vectors.
   - E_n^T - the n eigen vectors transposed.

   What is happening?

   (a) Normalize each line (horizontally) of matrix A. (A_n).
   (b) Transpose the normalized matrix(A_n) and store it for matrix multiplication later. (A_n^T);
   (c) Run SVD on the normalized matrix(A_n) and retrieve r number of raw eigen vectors. Each eigen vector has $cols$ number of elements.
   (d) Use tail and get the top n eigen vectors. (E_n)
   (e) Transpose the tailed top n eigen vectors. (E_n^T)
   (f) Multiply (A_n^T) "transposeTimes" (E_n^T).

2 Results

The results of the pcaJob is a matrix with dimensions $cols$ by n. The eigen vectors are sorted in order, and therefore the vector with the most variability is the one furthest to the right. The order of the rows is unchanged from the original matrix.