Math 189R problem set 8

Adam Guo 2020-04-13

1. **(K-means)** In this problem, we will implement the k-means algorithm and separate 5,000 2D data points into different number of clusters.

   Let $X = x_1, x_2 \ldots, x_m$ be the data points, and let k be the number of clusters. The k-means algorithm is summarized as following:

   (a) Randomly initialize k cluster centers, $\mu_1, \mu_2 \ldots \mu_k$, in the feature space.

   (b) Calculate the distance between each data points and the cluster centers.

   (c) Assign each data point to the cluster center c whose distance between this data point is the minimum of all the cluster centers, namely,

   $$c_i = \arg\min_j ||x_i - \mu_j||^2$$

   (d) Update each cluster center to be

   $$\mu_j = \frac{\sum_{i=1}^m 1\{c_i = j\} x_i}{\sum_{i=1}^m 1\{c_i = j\}}$$

   (e) Repeat step 2 - 4 until convergence or exhausted.

   The objective (cost) function is defined as

   $$J(c, \mu) = \sum_{i=1}^m ||x_i - \mu_{c_i}||^2$$

   In this assignment, you will first implement the k-means cost function and the algorithm. Then, for $k = 1, 2 \ldots, 20$, find the number of clusters with the optimal cost and produce a plot of the relationship between the cost and the number of clusters. Then, visualize the data points and the cluster centers on the optimal number of clusters.

   Notice that the k-means algorithm might yield different results based on the randomness of the initialization of cluster centers.

Solution:

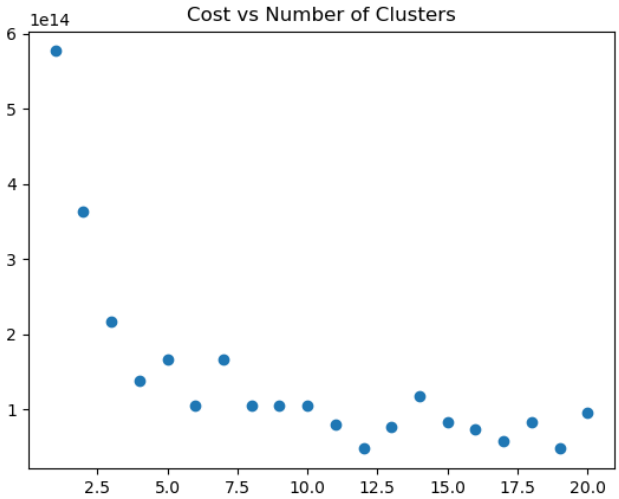

Figure 1: Plot of cost vs. number of clusters

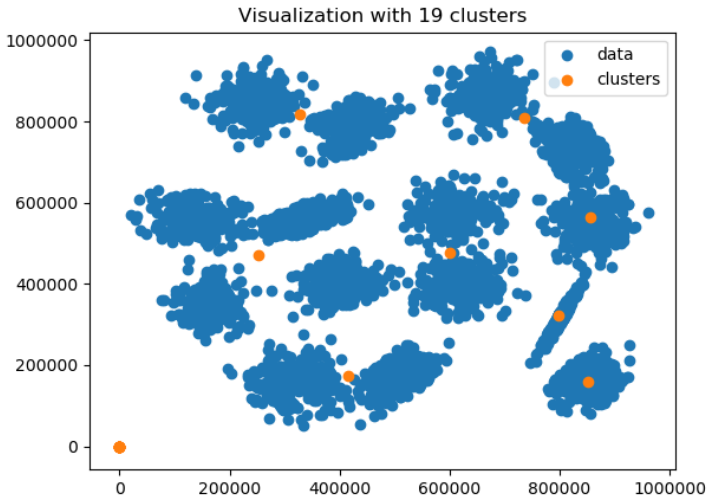

Figure 2: Plot of points and clusters

2. **Extra Credit (Non-Negative Matrix Factorization)**

In this problem, we will use the reuters dataset in nltk library. Please run nltk.download() in python shell to download the dataset. In the starter code, we have already parsed the data for you.

Choosing an appropriate objective function and algorithm from Lee and Seung 2001 [1] implement Non-Negative Matrix Factorization for topic modelling (choose an appropriate number of topics/latent features) and assert that the convergence properties proved in the paper hold. Display the 20 most relevant words for each of the topics you discover.

Solution:

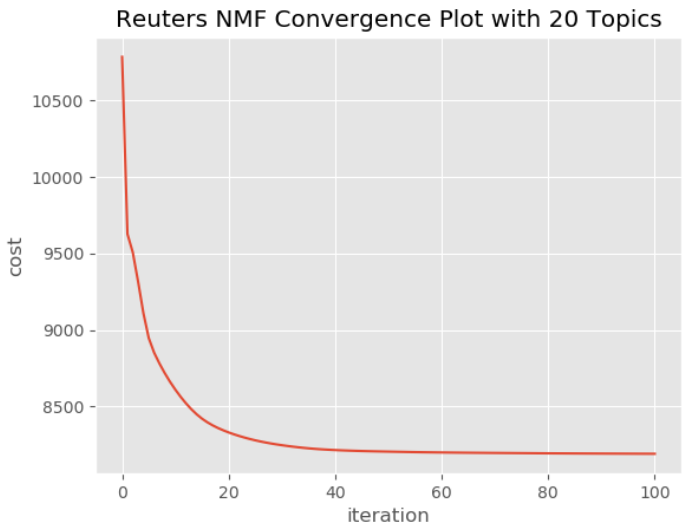

Figure 3: NMF convergence plot

```
==> Finding most frequent words for each topic...
-- topic 1: ['billion' 'surplus' 'deficit' 'francs' 'marks' 'in' 'deposits' 'reserves'
 'account' 'trade']
-- topic 2: ['japan' 'trade' 'dollar' 'yen' 'to' 'japanese' 'the' 'bank' 'dealers'
 'paris']
-- topic 3: ['loss' 'vs' 'revs' 'year' '4th' 'shr' 'includes' 'inc' 'discontinued'
 'of']
-- topic 4: ['pct' 'in' 'february' 'january' 'rose' 'year' 'rise' 'from' 'rate'
 'index']
-- topic 5: ['stg' 'bank' 'money' 'market' 'the' 'england' 'bills' 'of' 'band'
 'assistance']
-- topic 6: ['oper' 'excludes' 'cts' 'vs' 'net' 'or' 'shr' 'discontinued' 'gain'
 'note']
-- topic 7: ['vs' 'cts' 'net' 'shr' 'revs' 'qtr' 'mths' 'nine' '3rd' '1st']
-- topic 8: ['oil' 'crude' 'barrel' 'prices' 'bbl' '50' 'opec' 'raises' 'postings'
 'gas']
-- topic 9: ['000' 'vs' 'net' 'cts' 'includes' 'year' 'note' 'gain' '500' 'sales']
-- topic 10: ['profit' 'vs' 'cts' 'net' 'shr' 'revs' 'qtr' 'six' 'nine' '4th']
-- topic 11: ['the' 'to' 'and' 'of' 'that' 'said' 'in' 'he' 'on' 'would']
-- topic 12: ['mar' '1987' 'apr' '20' 'feb' '26' 'oct' '12' '16' '25']
-- topic 13: ['mln' 'vs' 'stg' '1986' 'tax' '11' '16' '12' '13' 'year']
-- topic 14: ['cts' 'qtly' 'div' 'record' 'pay' 'april' 'prior' 'dividend' 'vs' 'sets']
-- topic 15: ['fed' 'customer' 'says' 'repurchase' 'reserves' 'federal' 'agreements'
 'funds' 'reserve' 'repurchases']
-- topic 16: ['tonnes' 'wheat' 'sugar' 'corn' '87' 'export' 'for' 'grain' 'at' 'to']
-- topic 17: ['vs' 'avg' 'shrs' 'cts' 'net' 'shr' 'sales' 'qtr' 'lt' '1st']
-- topic 18: ['dlrs' 'share' 'quarter' 'of' 'and' 'earnings' 'or' 'in' 'year' 'net']
-- topic 19: ['the' 'of' 'in' 'said' 'quarter' 'to' 'first' 'it' 'company' 'for']
-- topic 20: ['it' 'shares' 'of' 'said' 'to' 'stock' 'its' 'lt' 'company' 'inc']
```

Figure 4: Most common words in each topic

[1] https://papers.nips.cc/paper/1861-algorithms-for-non-negative-matrix-factorization.pdf