Math189R SP19
Homework 8
Thursday, Apr 15, 2019

Feel free to work with other students, but make sure you write up the homework and code on your own (no copying homework *or* code; no pair programming). Feel free to ask students or instructors for help debugging code or whatever else, though.

The starter files for problem 1 and extra credit can be found under the Resource tab on course website. Please print out all the graphs generated by your own code and submit them together with the written part, and make sure you upload the code to your Github repository.

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 $X = x_1, x_2 \dots, x_m$ be the data points, and let k be the number of clusters. The k-means algorithm is summarized as following:
1. Randomly initialize k cluster centers, $\mu_1, \mu_2 \dots \mu_k$, in the feature space.
2. Calculate the distance between each data points and the cluster centers.
3. Assign each data point to the cluster center c whose distance between this data point is the minimum of all the cluster centers, namely,

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

4. Update each cluster center to be

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

5. 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 $k = 1, 2 \dots, 20$, 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.

Coding Question ■

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[a] 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.

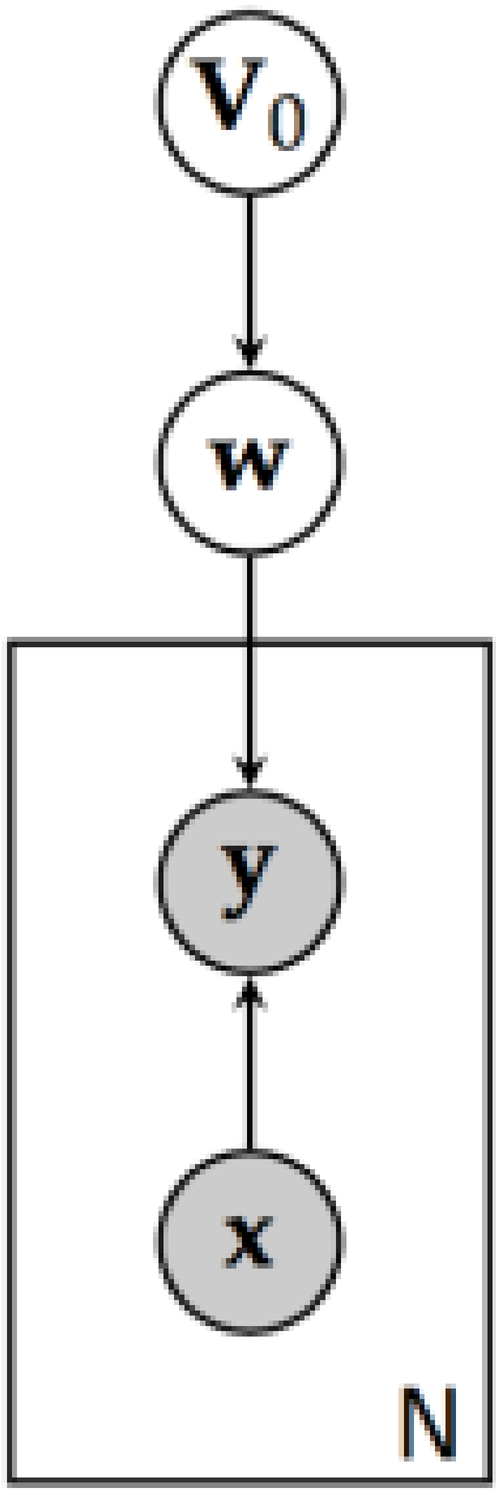
[a] `https://papers.nips.cc/paper/1861-algorithms-for-non-negative-matrix-factorization.pdf`

We chose to minimize the Frobenius norm between the approximation and the actual matrix, but it's equally okay to minimize the fake KL divergence objective they present in the paper. We also chose to use the multiplicative update because it's easier to implement and guarentees monotonicity of the objective (which we check below with the convergence plot). Also below are the most important words for each of the 20 topics. Note that you might have something different since the objective is non-convex, though they should still look coherent.

– topic 1: ['loss' 'vs' 'revs' 'shr' '4th' 'year' 'cts' 'includes' 'lt' 'inc']
– topic 2: ['rate' 'pct' 'bank' 'prime' 'rates' 'to' 'from' 'banks' 'effective' 'cut']
3– topic 3: ['mln' '1986' 'year' 'from' 'in' 'and' 'to' '1985' 'tax' 'extraordinary']
– topic 4: ['fed' 'customer' 'says' 'repurchase' 'reserves' 'federal' 'agreements' 'funds' 'reserve' 'repurchases']
– topic 5: ['to' 'the' 'he' 'trade' 'that' 'and' 'said' 'in' 'would' 'japan']
– topic 6: ['000' 'vs' 'net' 'cts' 'sales' 'avg' 'shrs' 'shr' 'year' 'revs']
– topic 7: ['dlrs' 'quarter' 'share' 'of' 'earnings' '1987' 'or' 'net' 'first' '1986']
– topic 8: ['vs' 'mln' 'net' 'shrs' 'avg' 'shr' 'revs' 'mths' 'nine' '3rd']
– topic 9: ['billion' 'surplus' 'deficit' 'francs' 'marks' 'reserves' 'deposits' 'in' 'trade' 'account']
– topic 10: ['profit' 'vs' 'loss' 'net' 'cts' 'nil' 'revs' 'tax' 'shr' 'six']
– topic 11: ['yen' 'dollar' 'bank' 'japan' 'dealers' 'dollars' 'the' 'tokyo' 'central' 'currency']
– topic 12: ['oil' 'crude' 'prices' 'barrel' 'opec' '50' 'bbl' 'gas' 'bpd' 'to']
– topic 13: ['pct' 'in' 'february' 'january' 'rose' 'year' 'rise' 'index' '1986' 'december']
– topic 14: ['oper' 'excludes' 'vs' 'cts' 'net' 'shr' 'discontinued' 'or' 'note' 'gain']
– topic 15: ['stg' 'bank' 'money' 'market' 'bills' 'england' 'band' 'the' 'assistance' 'mln']
– topic 16: ['tonnes' 'wheat' 'sugar' 'corn' '87' 'export' 'for' 'grain' 'to' 'ec']
– topic 17: ['it' 'of' 'said' 'shares' 'to' 'the' 'stock' 'its' 'lt' 'company']
– topic 18: ['cts' 'div' 'qtly' 'record' 'april' 'pay' 'prior' 'dividend' 'vs' 'sets']
– topic 19: ['the' 'of' 'to' 'in' 'said' 'and' 'on' 'by' 'for' 'was']
– topic 20: ['cts' 'vs' 'shr' 'qtr' 'net' '1st' 'revs' 'inc' '4th' 'lt']

■

Extra Credit (Linear Regression) Consider the Bayesian Linear Regression model with the following generative process:
(1) Draw $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_0)$
(2) Draw $\mathbf{y}_i \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}_i, \sigma^2)$ for $i = 1, 2, \ldots, n$ where σ^2 is known.
Express this model as a directed graphical model using Plate notation. Is $\mathbf{y}_i$ independent of $\mathbf{w}$? Is $\mathbf{y}_i$ independent of $\mathbf{w}$ *given* $\mathcal{D} = \{\mathbf{x}_i\}$? Support these claims.



$\mathbf{V}_0$
$\mathbf{w}$
$\mathbf{y}$
$\mathbf{x}$
N

By d-separation we can see that y is not independent of w regardless of whether we condition on x because y is an observed child of w, effectively observing w. Note that this would change if we are predicting an unobserved y, but that is a different model!

■