LECTURE 3

Topic modeling
Unsupervised text classification

K-MEANS

MODEL

- Data: a vectorized corpus of n documents
  - $X \in \mathbb{R}^{n \times d}$
- Partition the documents in k homogeneous groups
  - k is a hyperparameter
  - Cluster assignments $A \in \{0,1\}^{n \times k}$
    - $a_{ij} = \begin{cases} 1 & if \;\; x_i \;\; belongs \;\; to \;\; cluster \;\; j, \\ 0 & else\,. \end{cases}$
  - Cluster centroids $M \in \mathbb{R}^{k \times m}$
    - $M = \{\mu_1, \mu_2, \ldots, \mu_k\}$

PARAMETER ESTIMATION

- Minimum residual sum of squares (MRSS)
  - $RSS(D; A, M) = \sum_{j=1}^{k} \sum_{i=1}^{n} a_{ij} | x_i - \mu_j |^2$
  - Minimize the intra-cluster sum of squared Euclidean distances between the centroid and the documents
    - $| x_i - \mu_j | = \sqrt{\sum_{c=1}^{m} (x_{ic} - \mu_{jc})^2}$
  - Because A is discrete, this is a non-convex problem
    - The solution is alternating optimization

PARAMETER ESTIMATION

- General algorithm
  1. Initialize the centroids by randomly picking k documents
  2. Repeat until convergence, *i.e.* until M and A remain constant between two iterations :
     1. Determine the optimal cluster assignments, given the centroids
        - $$a_{ij} = \begin{cases} 1 & if \;\; j = \underset{j'}{\text{argmin}}(|x_i - \mu_{j'}|), \\ 0 & else\,. \end{cases}$$
     2. Determine the optimal centroids, given the assignments
- Repeat this algorithm several times and pick the best partitioning (the one that minimizes the *RSS*)

PARAMETER ESTIMATION

- Step 2.2: gradient of the RSS
  - $$\nabla_{\mu_p} RSS = \nabla_{\mu_p}\Big(\sum_{j=1}^{k}\sum_{i=1}^{n} a_{ij}|x_i - \mu_j|^2\Big)$$
    $$= \nabla_{\mu_p}\Big(\sum_{i=1}^{n} a_{ip}(x_i - \mu_p)^2\Big)$$
    $$= \nabla_{\mu_p}\Big(\sum_{i=1}^{n} a_{ip}(x_i^2 - 2x_i\mu_p + \mu_p^2)\Big)$$
    $$= \nabla_{\mu_p}\Big(\sum_{i=1}^{n} a_{ip}(-2x_i\mu_p + \mu_p^2)\Big)$$
    $$= \sum_{i=1}^{n} a_{ip}(-2x_i + 2\mu_p) = -2\sum_{i=1}^{n} a_{ip}x_i + 2\sum_{i=1}^{n} a_{ip}\mu_p$$

PARAMETER ESTIMATION

- Step 2.2: solution of $\nabla_{\mu_p} RSS = 0$ in μ_p
  - $\nabla_{\mu_p} RSS = 0 \Leftrightarrow -2\sum_{i=1}^{n} a_{ip}x_i + 2\sum_{i=1}^{n} a_{ip}\mu_p = 0$
  - $2\sum_{i=1}^{n} a_{ij}\mu_p = 2\sum_{i=1}^{n} a_{ip}x_i$

$$\mu_p \sum_{i=1}^{n} a_{ip} = \sum_{i=1}^{n} a_{ip}x_i$$

$$\mu_p = \frac{\sum_{i=1}^{n} a_{ip}x_i}{\sum_{i=1}^{n} a_{ip}}$$

  - The optimal centroid is the mean of the document vectors assigned to this cluster, hence the name of the method

LATENT SEMANTIC INDEXING

BASICS

- Data: a vectorized corpus of n documents with *tf-idf* weighting
  - $X \in \mathbb{R}^{n \times m}$
- Covariance matrix
  - $A = \frac{1}{n} X^\top X$ (with X centered)
  - a_{ij} is the correlation between term i and term j
  - Because of semantic relationships, some pairs of words are likely to be correlated

EIGEN DECOMPOSITION

- Decomposition
  - $X^\top X = S\Lambda S^\top$
    - S is a m by m unitary matrix, such that $s_{.i}$ is the i^th eigenvector
    - Λ is a diagonal matrix, such that λ_{ii} is the i^th eigenvalue
- New representation of the documents, Z
  - $Z = XS$
- Proof of diagonality of the covariance matrix of Z
  - $A_Z = \frac{1}{n} Z^\top Z = \frac{1}{n}(XS)^\top (XS) = \frac{1}{n} S^\top X^\top XS = \frac{1}{n} S^\top S \Lambda S^\top S$

    $A_Z = \frac{1}{n}\Lambda$

DIMENSION REDUCTION

- Low-dimension representation $Z_d \in \mathbb{R}^{n \times d}, d \ll m$
  - $Z_d = XS_d$ the new dimensions being some sorts of topics
- Computing the eigen decomposition is very expensive
  - Instead we compute the singular value decomposition of X
    - $X = UDV^\top$
      - $U \in \mathbb{R}^{n \times n}$: left-singular vectors
      - $D \in \mathbb{R}^{n \times m}$: diagonal matrix describing the singular values
      - $V \in \mathbb{R}^{m \times m}$: right-singular vectors
  - Low-dimension representation
    - $Z_d = U_d D_d$

NON-NEGATIVE MATRIX FACTORIZATION

BASICS

- Data: a vectorized corpus of n documents either weighted or not
  - $X \in \mathbb{R}^{n \times m}$
- X is non-negative by nature (*tf-idf* is non-negative)
  - Approximate it with a bilinear factorisation $X \simeq WH$
    - $W \in [0; \infty[^{n \times k}$: mixture coefficient vectors, *i.e.* descriptions of documents in terms of topics
    - $H \in [0; \infty[^{k \times m}$: basis vectors, *i.e.* descriptions of topics in terms of words

FACTORIZATION

- The objective is to find W and H so that the error of reconstruction of X is minimal
  - Two common ways of measuring the quality of the approximation
    - Euclidean distance
      - $$||X - WH||^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} (x_{ij} - w_i \cdot h_j)^2$$
    - Kullback-Leibler
      - $$KL(X||WH) = \sum_{i=1}^{n} \sum_{j=1}^{m} \left[X_{ij} \log\left(\frac{x_{ij}}{w_{ij} \cdot h_{ij}} \right) - x_{ij} + w_{ij} \cdot h_{ij} \right]$$

INTERPRETATION

- To get an understanding of a topic
  - Look at the largest coefficients in the related row-vector of H ; a simple sorted plot-bar is a usual representation
- To know what are the topics that underly a document
  - Look at the largest coefficients in the related row-vector W
- To see to which topics a word is related
  - Look at the largest coefficient in the related column-vector of H
- To get topic proportions
  - Compute the normalized column-wise sum of W

TAKE AWAY MESSAGE

- Reducing the dimension of the corpus representation helps capturing topics and semantic information
  - Latent semantic indexing
    - Solid mathematical grounding
    - Difficult to interpret, dense document representations
  - Non-negative matrix factorization
    - Easy to interpret topics and document representations
    - No statistical justification
- Choosing the adequate number of topics remains an open question