DATA.STAT.840 Statistical Methods for Text Data Analysis

Exercises for Lecture 4: Document clustering, and introduction to N-grams

Exercise 4.1: Paragraph clustering for The Wonderful Wizard of Oz.

Lecture 4 showed how vector-space models, TF-IDF representation, and Gaussian mixture models could be used to find clusters of similar documents. In this exercise we show how the same approach can be used to discover clusters of similar paragraphs within a single document.

(a) Using the Python **`re`** library, this command can be used to split a string into paragraphs: **`mytext_paragraphs=re.split('\n[ \n]*\n', mytext)`**
where mytext is the string to split. The result is a list of strings (paragraphs), the first and last element list are empty paragraphs, the rest are the actual paragraphs. Modify the processing pipeline discussed on Lecture 2 to read a single text, split it into paragraphs, and process those paragraphs as individual documents.

(b) Download the Project Gutenberg .TXT ebook of Frank L. Baum's "The Wonderful Wizard of Oz", and process its paragraphs using the pipeline you created. (See https://en.wikipedia.org/wiki/The_Wonderful_Wizard_of_Oz for information about the work.)

(c) Create a TF-IDF vector-space representation for the documents, using Length-normalized frequency for the TF part and Smoother logarithmic inverse document frequency for the IDF part.

(d) Create a Gaussian mixture model with 10 mixture components, and fit it to the documents (paragraphs).

(e) For the resulting Gaussian mixture, print for each mixture component the 10 words with highest absolute value in the mean.

(f) For the resulting Gaussian mixture, print for each mixture component the document (paragraph) with highest membership probability in that component.

(g) Discuss the results in (e) and (f).

Report the component information asked for in (e) and (f), your comments in (g), and your code. You can use any programming language and any library, including any implementation of Gaussian mixture modeling.

Exercise 4.2: Comparison of TF-IDF variants.

(a) Find the longest paragraph of "The Wonderful Wizard of Oz". Report the text of the paragraph.

(b) Using the same processing as in Exercise 4.1 (a)-(b), create a TF-IDF representation for the longest paragraph, with three options:

- Length-normalized frequency (TF) and Smoothed logarithmic inverse document frequency (IDF)
- Logarithm of the count (TF) and Smoothed logarithmic inverse document frequency (IDF)
- Count relative to most frequent term (TF) and Version proportional to most common term (IDF)

For each option, report the 20 words with highest TF-IDF weight. Discuss the results; does any one of the options seem to match the semantic meaning of the paragraph better?

Report the requested text, statistics, discussion, and your code.

(exercises continue on the next page)

Exercise 4.3: The Expectation-Maximization algorithm.

The Expectation-Maximization (EM) algorithm is used for maximum likelihood optimization of many statistical models. In the general form, the algorithm optimizes the (log) likelihood of a set of observations, together denoted $\boldsymbol{X}$, which are assumed to be generated given a set of underlying latent variables, together denoted $\boldsymbol{Z}$. The log-likelihood that EM is supposed to optimize is

$$\log p(\boldsymbol{X};\boldsymbol{\Theta})$$

where $\boldsymbol{\Theta}$ denotes the parameters of the statistical model. Each iteration t of the EM algorithm is supposed to update the parameters in a way that increases (or at least never decreases) this likelihood, so that $\log p(\boldsymbol{X};\boldsymbol{\Theta}_{t+1}) \geq \log p(\boldsymbol{X};\boldsymbol{\Theta}_t)$. *But how can we know that it does that?*

In each iteration t, EM performs an E-step and an M-step. The E-step of the algorithm computes the probabilities $p(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\Theta}_t)$. The M-step computes new parameters as:

$$\boldsymbol{\Theta}_{t+1} = max_{\boldsymbol{\Theta}} \; Q(\boldsymbol{\Theta};\boldsymbol{\Theta}_t)$$

$$\text{where} \quad Q(\boldsymbol{\Theta};\boldsymbol{\Theta}_t) = \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\Theta}_t) \log p(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\Theta}) \; .$$

For example, the update equations of the Gaussian mixture model weights, means, and covariances shown on the Lecture 4 slides are a specific case of the above equation. But the equation that the M-step is maximizing is not the same as $\log p(\boldsymbol{X};\boldsymbol{\Theta})$!

Prove that the EM algorithm still works so that $\log p(\boldsymbol{X};\boldsymbol{\Theta}_{t+1}) \geq \log p(\boldsymbol{X};\boldsymbol{\Theta}_t)$ for every iteration. Follow the steps on the next page:

(a) Write the difference of the log-likelihoods $\log p(\boldsymbol{X};\boldsymbol{\Theta}_{t+1})$ and $\log p(\boldsymbol{X};\boldsymbol{\Theta}_t)$. We must prove this difference is greater than zero.

(b) Write the equation of the expected value of this difference over the distribution $p(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\Theta}_t)$. Show that this expected value is the same as the value in step (a). Hint: does $\boldsymbol{Z}$ appear in the original difference?

(c) In the expected value there are logarithms. Inside the logarithm that involves $\boldsymbol{\Theta}_t$, multiply and divide by $p(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\Theta}_t)$. Inside the logarithm that involves $\boldsymbol{\Theta}_{t+1}$, multiply and divide by $p(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\Theta}_{t+1})$.

(d) Show that the equation from (c) can be written as a sum of two terms: $Q(\boldsymbol{\Theta}_{t+1};\boldsymbol{\Theta}_t) - Q(\boldsymbol{\Theta}_t;\boldsymbol{\Theta}_t)$ and a *Kullback-Leibler divergence*. In general, a Kullback-Leibler divergence is an equation of the form $\sum_k p(k) \log \frac{p(k)}{q(k)}$ where $p(k)$ and $q(k)$ are two probability distributions; the divergence measures the amount of difference between the distributions and is always zero or greater. Hint: in general, $\log(a \cdot b) = \log(a) + \log(b)$ and $\log(a/b) = \log(a) - \log(b)$.

(e) Show that because of (d), we know that $\log p(\boldsymbol{X};\boldsymbol{\Theta}_{t+1}) \geq \log p(\boldsymbol{X};\boldsymbol{\Theta}_t)$. Hint: why is $Q(\boldsymbol{\Theta}_{t+1};\boldsymbol{\Theta}_t) - Q(\boldsymbol{\Theta}_t;\boldsymbol{\Theta}_t) \geq 0$?

In your report, show your derivations for each step.