COL866: Foundations of Data Science

Ragesh Jaiswal, IITD

Topic Models

Predictive versus Generative

- The Machine Learning discussion we had was predictive in nature. That is, we were interested in building distinguishers for the data without caring about how the data was generated.
- Generative modeling attempts learn the probabilistic process used to generate the observed data. This is a more difficult problem compared to building distinguishers.

Topic Models

- Topic Modeling is the problem is fitting a certain type of stochastic model to a given collection of documents.
- Here are the main assumption of the model:
  - There are r topics.
  - Each of the n documents is a mixture of these topics.
  - The topic mixture of a given document determines the probabilities (frequency) of the d words or terms.
    - A topic is assumed to determine the word frequencies and the frequency of words in a document is a convex combination of the frequency of the topics in the document.

Topic Models

- Topic Modeling is the problem is fitting a certain type of stochastic model to a given collection of documents.
- Here are the main assumption of the model:
  - There are r topics.
  - Each of the n documents is a mixture of these topics.
  - The topic mixture of a given document determines the probabilities (frequency) of the d words or terms.
    - A topic is assumed to determine the word frequencies and the frequency of words in a document is a convex combination of the frequency of the topics in the document.
- The above assumption implies that we are viewing documents in terms of bag of words disregarding the order in which the words appearing in the document. Even though throwing away context information may seem wasteful, but the bag-of-words approach works well in practice.

Topic Models

- In the bag-of-words model, a collection of documents may be represented by a $d \times n$ matrix A called the term-document matrix. This matrix is what is observed.
- In topic modeling, we assumes that there are r topics such that each document is a mixture of these r topics.
- Each document has an associated vector of size r that should give the mixture weights of topics in the document. So, the elements are non-negative with sum equal to 1. These vectors arranged as columns in an $r \times n$ matrix C is called the topic-document matrix.
- There is an $d \times r$ matrix B, called the term-topic matrix, where each column is a vector of expected frequencies of terms in that topic.
- Given B and C, $P = BC$ is a matrix with column $p_{.j}$ denotes the expected frequencies of terms in document j.

- Following is the process to generate n documents each containing m terms (and hence the term-document matrix A).

Document generation process

- Initialise $a_{ij} = 0$ for all $i \in \{1, ..., d\}$ and $j \in \{1, ..., n\}$.
- For $j = 1, ..., n$ in i.i.d. trials do:
  - Let $p_{.j} = Bc_{.j}$
  - For $t = 1$ to m:
    - Generate the t^{th} term x_t of document j by sampling from the set $\{1, ..., d\}$ using the probability vector $p_{.j}$
    - $a_{x_t j} \mathrel{+}= 1/m$

- Following is the process to generate n documents each containing m terms (and hence the term-document matrix A).

Document generation process

- Initialise $a_{ij} = 0$ for all $i \in \{1, ..., d\}$ and $j \in \{1, ..., n\}$.
- For $j = 1, ..., n$ in i.i.d. trials do:
  - Let $p_{.j} = Bc_{.j}$
  - For $t = 1$ to m:
    - Generate the t^{th} term x_t of document j by sampling from the set $\{1, ..., d\}$ using the probability vector $p_{.j}$
    - $a_{x_t j} \mathrel{+}= 1/m$

- The topic modeling problem is to infer B and C from A.
- This problem can also be viewed as non-negative matrix factorisation (NMF) where the goal is to factorise A into B and C with additional constraint that these matrices are non-negative with column sums as 1.
- This problem is computationally hard in general but under suitable assumptions becomes tractable.

- The topic modeling problem is to infer B and C from A.
- This problem can also be viewed as non-negative matrix factorisation (NMF) where the goal is to factorise A into B and C with additional constraint that these matrices are non-negative with column sums as 1.
- This problem is computationally hard in general but under suitable assumptions becomes tractable.
- Here the assumptions of an idealised model where the problem becomes tractable.
    - The pure topic assumption: Each document is purely on a single topic.
    - Separability assumption: The sets of terms occurring in different topics are disjoint.
- Claim: In the above strong model the matrix A has a block structure.

- Pure topics assumption:
  - The pure topic assumption: Each document is purely on a single topic.
  - Separability assumption: The sets of terms occurring in different topics are disjoint.
- Claim: In the above strong model the matrix A has a block structure.
- Let T_l denote the set of documents on topic l and S_l the subset of terms occurring in topic l.
- So, the problem in this strong setting becomes a clustering problem. If we can find the term clusters $S_1, ..., S_r$ or document clusters $T_1, ... T_r$, then we can obtain a good estimate on B and C.
- The data generation assumption provides crucial help in this clustering task.

Topic Models
An idealized model

- Pure topics assumption:
  - The pure topic assumption: Each document is purely on a single topic.
  - Separability assumption: The sets of terms occurring in different topics are disjoint.
- Claim: In the above strong model the matrix A has a block structure.
- Let T_l denote the set of documents on topic l and S_l the subset of terms occurring in topic l.
- So, the problem in this strong setting becomes a clustering problem. If we can find the term clusters $S_1, ..., S_r$ or document clusters $T_1, ... T_r$, then we can obtain a good estimate on B and C.
- The data generation assumption provides crucial help in this clustering task.

Document generation process

- Initialise all $a_{ij} = 0$
- For $j = 1, ..., n$:
  - For $t = 1, ..., m$:
    - Generate the t^{th} term x_t of document j by sampling a term using the probability vector $b_{.l}$, where l is the topic of document j.
    - $a_{x_t j} += 1/m$

Topic Models
An idealized model

- Pure topics assumption:
  - The pure topic assumption: Each document is purely on a single topic.
  - Separability assumption: The sets of terms occurring in different topics are disjoint.
- Let T_l denote the set of documents on topic l and S_l the subset of terms occurring in topic l.
- So, the problem in this strong setting becomes a clustering problem. If we can find the term clusters $S_1, ..., S_r$ or document clusters $T_1, ... T_r$, then we can obtain a good estimate on B and C.
- The data generation assumption provides crucial help in this clustering task.

Document generation process

- Initialise all $a_{ij} = 0$
- For $j = 1, ..., n$:
  - For $t = 1, ..., m$:
    - Generate the t^{th} term x_t of document j by sampling a term using the probability vector $b_{.l}$, where l is the topic of document j.
    - $a_{x_t j} \mathrel{+}= 1/m$

- The clustering problem in the above generative framework is to estimate the clusters $S_1, ..., S_r$ given a matrix A generated using the above process.

- Pure topics assumption:
  - The pure topic assumption: Each document is purely on a single topic.
  - Separability assumption: The sets of terms occurring in different topics are disjoint.
- Let T_l denote the set of documents on topic l and S_l the subset of terms occurring in topic l.
- So, the problem in this strong setting becomes a clustering problem. If we can find the term clusters $S_1, ..., S_r$ or document clusters $T_1, ... T_r$, then we can obtain a good estimate on B and C.
- The data generation assumption provides crucial help in this clustering task.

Document generation process

- Initialise all $a_{ij} = 0$
- For $j = 1, ..., n$:
  - For $t = 1, ..., m$:
    - Generate the t^{th} term x_t of document j by sampling a term using the probability vector $b_{.l}$, where l is the topic of document j.
    - $a_{x_t j} \mathrel{+}= 1/m$

- The clustering problem in the above generative framework is to estimate the clusters $S_1, ..., S_r$ given a matrix A generated using the above process.
- Claim (informal): Under reasonable assumption on the number of documents available, there is an efficient clustering algorithm.

End