Clustering with Dirichlet Mixture Models

Review and application of Bayesian non-parametric clustering

Alessandro Pierro   Dario Coscia

July 21, 2022

Table of Contents

Clustering Problem

Clustering Problem

Cluster analysis is the task of partitioning a set of observations into sub-groups, called a **clusters**, such that observations in the same clusters are closer with respect to a **similarity measure** to each other than to those in other clusters.

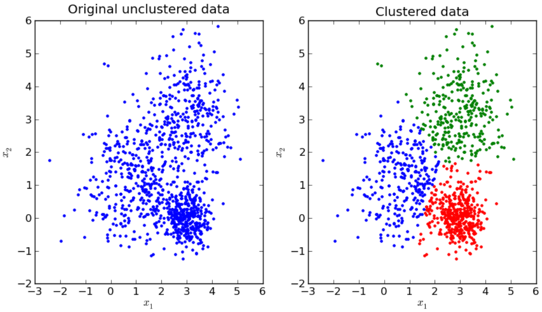

Figure 1: Clustering problem

Finite Mixtures – Classical Formulation

- The classical probabilistic approach to clustering is with **finite mixtures models**

$$\boldsymbol{\pi} \sim Dir(\frac{\alpha}{K}\mathbb{1}) \qquad z_i \sim Cat(\boldsymbol{\pi})$$

$$\boldsymbol{\theta}_k \sim H(\lambda) \qquad \mathbf{x}_i \sim F(\boldsymbol{\theta}_{z_i})$$

- Usually, the prior $H(\lambda)$ is chosen to be conjugate with the distribution $F(\boldsymbol{\theta})$
- In the case of Gaussian Mixture, inference can be performed using **Expectation Maximization**

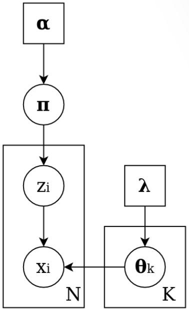

Figure 2: Finite Mixture PGM

Finite Mixtures – Alternative Formulation

- An alternative formulation for the finite mixtures can be given as

$$\boldsymbol{\pi} \sim Dir(\frac{\alpha}{K}\mathbb{1}) \qquad \bar{\boldsymbol{\theta}}_i \sim G(\boldsymbol{\theta})$$
$$\boldsymbol{\theta}_k \sim H(\lambda) \qquad \mathbf{x}_i \sim F(\bar{\boldsymbol{\theta}}_i)$$

where

$$G(\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \delta_{\theta_k}(\boldsymbol{\theta})$$

- If we sample (enough) from G, we will have K different values with probability 1

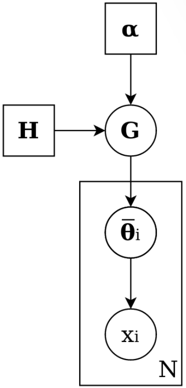

Figure 3: Finite Mixture PGM

Dirichlet Processes Mixture Models

Towards Infinite Mixture Models

- Using finite mixtures (or k-means) requires selecting the **number of clusters**. This can be done using the *elbow method* or comparing model evidences
- We would like to have a (possibly) variable number of clusters, to handle **novelty detection**

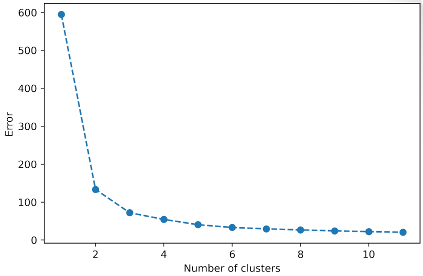

Figure 4: Elbow method for selecting k

How can we generalize to infinite mixtures?

Dirichlet Processes – Definition

Dirichlet Process

A **Dirichlet Process (DP)** is a stochastic process that takes values over probability distributions $G : \Theta \to \mathbb{R}^+$, such that:

$$G(\theta) \geq 0 \quad \forall \theta \in \Theta$$
$$\int_\Theta G(\theta) d\theta = 1$$

For each partition $T_1, \ldots, T_k$ of Θ, a DP is defined implicitly by the requirement

$$(G(T_1), \ldots, G(T_K)) \sim Dir(H(\alpha T_1), \ldots, H(\alpha T_K))$$

where α is the **concentration parameter** and H is the **base measure**.

Dirichlet Processes Mixture Models

- An intuitive way to understand DP mixture models is the **stick-breaking construction**

$$\boldsymbol{\pi} \sim GEM(\alpha) \qquad \bar{\boldsymbol{\theta}}_i \sim G(\boldsymbol{\theta}) = \sum_{k=1}^{\infty} \pi_k \delta_{\boldsymbol{\theta}_k}(\boldsymbol{\theta})$$

$$\boldsymbol{\theta}_k \sim H(\lambda) \qquad \mathbf{x}_i \sim F(\bar{\boldsymbol{\theta}}_i)$$

- The $GEM(\alpha)$ process is defined as

$$\beta_k \sim Beta(1, \alpha)$$

$$\pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_l) = \beta_k \left(1 - \sum_{l=1}^{k-1} \pi_l\right)$$

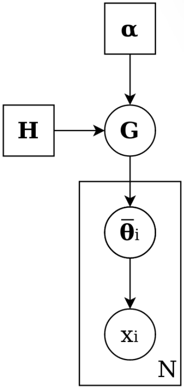

Figure 5: DP mixture PGM

Inference in Dirichlet Mixture Models

- We focus on Dirichlet mixture of Gaussians, and perform learning with variational inference on a stick-breaking formulation

True Model

$$\boldsymbol{\pi} \sim GEM(\alpha)$$

$$\boldsymbol{\theta}_k = \begin{pmatrix} \mu_k \\ \Sigma_k \end{pmatrix} \sim \begin{pmatrix} \mathcal{N}(0, \mathbb{1}) \\ W(D, \mathbb{1}) \end{pmatrix}$$

$$\bar{\boldsymbol{\theta}}_k = \begin{pmatrix} \bar{\mu}_k \\ \bar{\Sigma}_k \end{pmatrix} \sim \begin{pmatrix} G(\boldsymbol{\mu}) = \sum_k \pi_k \delta_{\boldsymbol{\mu}_k}(\boldsymbol{\mu}) \\ G(\boldsymbol{\Sigma}) = \sum_k \pi_k \delta_{\boldsymbol{\Sigma}_k}(\boldsymbol{\Sigma}) \end{pmatrix}$$

$$\mathbf{x}_i \sim \mathcal{N}(\bar{\boldsymbol{\mu}}_i, \bar{\boldsymbol{\Sigma}}_i)$$

Variational Approximation

$$\beta_k \sim Beta(\gamma_{k,1}, \gamma_{k,2})$$

$$\boldsymbol{\theta}_k = \begin{pmatrix} \mu_k \\ \Sigma_k \end{pmatrix} \sim \begin{pmatrix} \mathcal{N}(\nu_{\mu_k}, \mathbb{1}) \\ W(a_k, \mathbb{B}_{\mathbb{k}}) \end{pmatrix}$$

$$\mathbf{z}_i \sim Cat(\nu_{z_i})$$

Practical Applications

Methodology

We apply Dirichlet mixture models in different settings:

- Validation of the model on synthetic dataset

- Application on natural language clustering
  - Text pre-processing and word embedding
  - t-SNE order reduction
  - Model validation on t-SNE projections

We use the `scikit-learn` package in Python, that implements the infinite Gaussian mixture model.

Validation on Synthetic Dataset

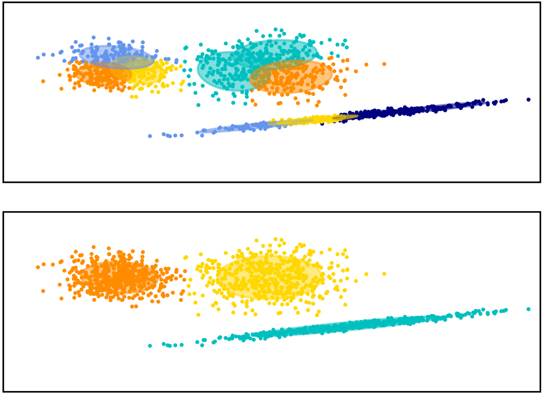

Figure 6: Mixture models. On top, a finite mixture model with ten components. On bottom, an infinite mixture model, with three components inferred.

20newsgroup – Dataset Description

- The 20 newsgroups dataset comprises around 18000 newsgroups posts on 20 topics
- Topics are varied, e.g. science, politics, sports, religion, technology etc.
- Each newsgroup file in the bundle represents a single newsgroup, and each message in a file is the text of some newsgroup document that was posted to that newsgroup.

Text pre-processing and word embedding is used to transform the text into a vector. In our experiments we used tf-idf for text vectorization.

Text Vectorization with tf-idf

- Given a term t and a document d, we define the **term frequency** $tf_{t,d}$ as

$$tf_{t,d} \quad \text{: number of occurrences of } t \text{ in } d$$

- For each term t, we define the **inverse document frequency** idf_t as

$$idf_t = \log \frac{N}{df_t}$$

  where N is the total number of documents in the corpus, and df_t is the number of documents containing t
- Given a vocabulary of terms $\{t_1, \ldots, t_M\}$, a document d can be vectorized as

$$d = (tf_{t_1,d} \cdot idf_{t_1}, \ldots, tf_{t_M,d} \cdot idf_{t_M})$$

Dimensionality Reduction with t-SNE

- Let $(x_1, ..., x_n)$ samples in the input space $\mathcal{X}$, and $(y_1, ..., y_n)$ samples in the embedding $\mathcal{Y}$. The **similarity measure** between point x_i and point x_j is represented by the conditional distribution:

$$p_{i|j} = \frac{\exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-||x_i - x_k||^2/2\sigma_i^2)}$$

- We minimize the KL divergence $K[p_{ij}||q_{ij}]$

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n} \qquad q_{ij} = \frac{(1 + ||y_i - y_j||^2)^-1}{\sum_k (1 + ||y_k - y_l||^2)^-1},$$

where q_{ij} is the similarity measure between point y_i and point y_j.

Visualizing Text Dataset

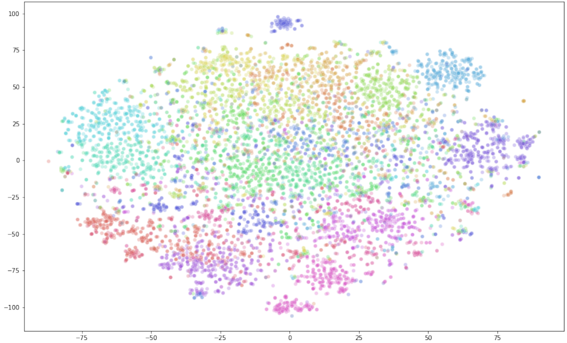

Figure 7: *20 newsgroup dataset* after tf-idf and t-SNE embedding

Text Clustering – Sub-groups Visualization

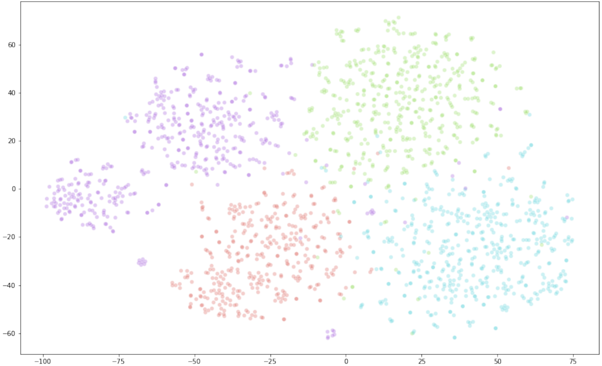

Figure 8: Four random sub-groups from the dataset, indexed by different colors

Text Clustering – Model Results

- The number of initial components is initialised to the number of topics.
- The concentration parameter is initialised very low 10^{-6} to ensure more mass at the edge of the mixture weights simplex.
- The model effectively learns four independent components as expected.

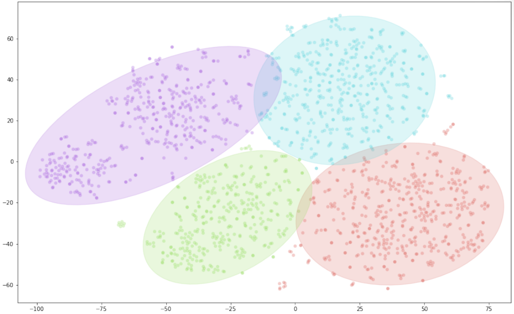

Figure 9: Dirichlet mixture clustering results

Conclusions

Conclusions

- Using infinite mixture models solves the problem of selecting the **optimal number of clusters**
- Inference in probabilistic models is **very slow**, compared to heuristic based algorithm, e.g. K-Means.
- The application to text clustering can be improved using a better (neural) word embedding (e.g. **word2vec**), or additional preprocessing steps (e.g. lemmatization)

References

[1] Zoubin Ghahramani. *Bayesian non-parametrics and the probabilistic approach to modelling*. 2013.

[2] Geoffrey E Hinton and Sam Roweis. *Stochastic neighbor embedding*. 2002.

[3] Laurens Van der Maaten and Geoffrey Hinton. *Visualizing data using t-SNE*. 2008.

[4] Kevin P Murphy. *Machine learning: a probabilistic perspective*. 2012.

[5] Radford M Neal. *Markov chain sampling methods for Dirichlet process mixture models*. 2000.

Thank you for the attention!

Alessandro Pierro and Dario Coscia