CS771A:Machine Learning tools and Techniques

Prof. Piyush Rai

September 22, 2016

Interactive Bayesian Document Clustering

Anivesh Agrawal(14100) Pranay Borkar(14189) Manuj Narang(14373)
Siddhartha Saxena(150719)

1 Problem Definition

With Big Data, there exists multiple ways to interpret it and quantitatively best clustering might not align with the users desired clustering. Thus a clustering algorithm that encodes users prior belief mathematically is favourable.

2 Project Goals

1. Building a clustering model that clusters the given data according to the user feedback via rejection [5]. The feedback mechanism modifies the prior, down-weighing the probability of rejected clusters and increasing for the accepted ones.

2. Building document clustering model, clustering documents according to the topics contained in them (extracted through Latent Dirichlet Allocation[6][7]), and implementing it on the datasets[3][4].

3 Data Sets

[1] CIFAR-10 link. The CIFAR-10 dataset consists of 60000 32x32 colour images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images.

[2] CMU Face. The CMU Pose, Illumination, and Expression (PIE) database, provided by Simon Baker

[3] Reuters. This is a collection of documents that appeared on Reuters newswire in 1987. The documents were assembled and indexed with categories.

[4] AAAI 2014 Accepted Papers. This data set compromises the metadata for the 2014 AAAI conference's accepted papers, including paper titles, authors, abstracts, and keywords of varying granularity

4 References:

[5] Akash Srivastava, James Zou, Charles Sutton. Clustering with a Reject Option: Interactive Clustering as Bayesian Prior Elicitation. (2016)

[6] David M. Blei, Andrew Y. Ng, Michael I. Jordan. Latent Dirichlet Allocation. (2003)

[7] Library for topic modelling with latent dirichlet allocation in python. link