STCS 6702: Foundations of Graphical Models: Reading 0

Maxime TCHIBOZO (MT3390)

September 2020

1 Blei (2014) – Build, Compute, Critique, Repeat: Data Analysis with Latent Variable Models

This paper gives a detailed description of the universe of ideas spanned by Latent Variable Models.

In practice, Machine Learning scientists should follow Box's Loop – the statistician's adaptation of the scientific method. Build models iteratively, with the assumption that a complex observed data set exhibits simpler, but unobserved, patterns.

Examples of problems that latent variable models can help solve: 1. Understanding the structure of social networks, and detecting clusters linking people who would enjoy meeting each other. 2. Understanding themes from documents and how these themes change over time. 3. Reconstructing original population groups that mixed together to form modern observations. This paper recalls the same framework defined in class, utilizing local hidden variables, global hidden variables and observed variables in the case of the Gaussian Mixture Model.

With θ the global variable describing Mixture Membership proportions. μ is drawn from the Gaussian whose parameter is selected from the vector θ according to a Dirichlet distribution.

This framework is used for inference of the latent variables, with the posterior we aim to find being p$(\mu_{1:K}, \theta, z_{1:N}|x_{1:N})$. Given the observed data, we infer the distribution of the hidden structure that probably generated them. The posterior (over the global variables) can also be used to form a posterior predictive distribution function once the latent variables have been inferred. There are 3 ways to reason with latent variable models: 1) The Generative Probabilistic Process

– i.e imagine the process which creates the data we observe. 2) The Joint Distribution – i.e Write down the joint mathematically, derive the posterior. 3) The Graphical Model – i.e Draw a graph of the Data Generating Process (1)

The paper then details specific applications or algorithms of latent variables. Linear Factor Models (PCA, factor analysis, canonical correlation analysis) – i.e dimensionality reduction. They provide a way of using our latent variable framework with high-dimensional data.

Mixed-Membership Models – they allow us to perform unsupervised analysis on data which exhibits specific groups or statistically related observations. These underlying groups can be themes in text, population groups in genetic data, types of respondent in a survey. To reason with the underlying data generating process, we should question how the groups are related between themselves, and within themselves. Matrix Factorization Models – predicting values inside of matrices (is row i connected to column j), for example the viewer-movie matrix in the Netflix problem. This relies on embeddings of both viewers and movies into low dimension vectors. The embedding vectors can then be used to measure distances and recognize clusters. Time Series Models – in Hidden Markov Models, each temporal observation is drawn from an unobserved mixture component, which itself depends on the previous component's unobserved mixture. HMMs are commonly used in Natural Language models. Kalman Filters use a similar graphical representation, but using continuous variables.

Other methods include hierarchical regression models and Bayesian non-parametric models. Once a basic data generating process has been selected, we can update the distributions at play depending on the type of data: we use truncated Gaussians or Gamma distributions for positive data, discrete distributions for categorical data or even Generalized Linear Models. It takes practice and experience to express the type of latent structure we want to recognize and devise the assumptions that help generate the data.

Once we have established a structure for the data generating process, we uncover the hidden parameters given observations by inferring the posterior (usually with Laplace approximations and Markov Chain Monte Carlo sampling). Laplace is intractable in high-dimensional problems. MCMC methods (which include Metropolis-Hastings and Gibbs sampling) simulate the Graphical Model process, drawing samples first from the transition distribution and later the stationary distribution. The generated samples are then used to empirically approximate the posterior.

Today, Variational Inference (which uses stochastic optimization) is preferred to MCMC for its scalability and performance over high-dimensional datasets. To simplify the joint distribution, we make assumptions simplifying its factorization. Usually, we approximate the posterior, and constrain the distribution of x to an exponential family. Working with exponential families allows us to deduce distributions of other variables using conjugacy.

For mixtures, we use any common prior/likelihood pair (gamma+Poisson, Gaussian+Gaussian, Dirichlet+Multinomial).

Variational Inference converts the inference problem into an optimization problem, by minimizing the KL divergence between the posterior and its approximation over the global latent parameter space.

Once a posterior is estimated, Posterior Predictive Checks provide a way to evaluate whether the estimated data generating process is satisfactory by comparing it to real data, and whether this data is compatible with previous assumptions.