Machine Learning 10-701 Mar 30, 2015
http://alex.smola.org/teaching/10-701-15 due on Apr 06, 2015
Carnegie Mellon University

Homework 9

START HERE: Instructions

- The homework is due at 9:00am on April 6, 2015. Anything that is received after that time will not be considered.
- Answers to every theory questions will be also submitted electronically on Autolab (PDF: Latex or handwritten and scanned). Make sure you prepare the answers to each question separately.
- Collaboration on solving the homework is allowed (after you have thought about the problems on your own). However, when you do collaborate, you should list your collaborators! You might also have gotten some inspiration from resources (books or online etc...). This might be OK only after you have tried to solve the problem, and couldn't. In such a case, you should cite your resources.
- If you do collaborate with someone or use a book or website, you are expected to write up your solution independently. That is, close the book and all of your notes before starting to write up your solution.
- Latex source of this homework: http://alex.smola.org/teaching/10-701-15/homework/hw9_latex.zip.

Mixtures of Exponential Family

In this problem we will study approximate inference on a general Bayesian Mixture Model. In particular, we will derive both Expectation-Maximization (EM) algorithm and Gibbs Sampling for the mixture model.

A typical finite-dimensional mixture model is a hierarchical model consisting of the following components:

- N random variables x_i, $i = 1, \ldots, N$ corresponding to observations, each assumed to be distributed according to a mixture of K components, with each component belonging to the same exponential family of distributions (e.g., all normal, all multinomial, etc.) but with different parameters

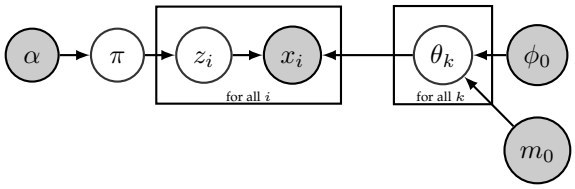
$$p(x_i|\theta) = \exp(\langle \phi(x_i), \theta \rangle - g(\theta)). \tag{1}$$

- N corresponding random latent variables z_i specifying the identity of the mixture component of each observation, each distributed according to a K-dimensional categorical distribution.
- A set of K mixture weights π_k, $k = 1, \ldots, K$, each of which is a probability (a real number between 0 and 1 inclusive), all of which sum to 1.
- A Dirichlet prior on the mixture weights having hyper-parameters α.
- A set of K parameters θ_k, $k = 1, \ldots, K$, each specifying the parameter of the corresponding mixture component. For example, observations distributed according to a mixture of one-dimensional Gaussian distributions will have a mean and variance for each component. Observations distributed according to a mixture of V-dimensional categorical distributions (e.g., when each observation is a word from a vocabulary of size V) will have a vector of V probabilities, collectively summing to 1. Moreover, we put a shared conjugate prior on these parameters:

$$p(\theta; m_0, \phi_0) = \exp\left(\langle \phi_0, \theta \rangle - m_0 g(\theta) - h(m_0, \phi_0)\right). \tag{2}$$

Homework 9

This model can be represented graphically as:



The generative model is given as follows:

1. For each topic $k \in \{1, ..., K\}$
   - Draw $\theta_k \sim p(\cdot|m_0, \phi_0)$
2. Draw mixture weights $\pi \sim \text{Dirichlet}(\alpha)$
3. For each observation $i \in \{1, ..., N\}$
   - Draw a component index $z_i \sim \text{Categorical}(\pi)$
   - Draw a datum $x_i \sim p(\cdot|\theta_{z_i})$

1 Expectation-Maximization on Mixture Models

Expectation-Maximization (EM) is an iterative algorithm for finding MLE or MAP estimates of parameters when the model depends on unobserved latent variables $\mathbf{Z}$. When the variables are fully observed except the model parameter θ, then we can simply run maximize the likeihood defined as $L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z}) = p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ where $\mathbf{X} = \{x_i,\ i = 1, \ldots, N\}$ and $\mathbf{Z} = \{z_i \in \{1, \ldots, K\}, i = 1, \ldots, N\}$. However, when $\mathbf{Z}$ is not observed, we have to integrate out (or sum out) $\mathbf{Z}$ and maximize over the collapsed likelihood i.e. $L(\boldsymbol{\theta}; \mathbf{X}) = p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$.

The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood (ℓ) evaluated using the current estimate for the parameters (initially, random values), and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step.

Question 1

E-step
Write down E-step in terms of π_k, the mixture component weight, and θ_k, the parameter for k-th component, for every component $k = 1, \ldots, K$.

1. Write down the maximization function of posterior given π_k, θ_k.
2. Solve $n_{ik} = \mathbb{E}\left[\delta(z_i = k)\right]$ that maximizes the posterior from previous question.

Question 2

M-step

1. Write down M-step as an optimization problem in terms of $\mathbf{X}$ and $g(\cdot)$, ϕ_0, m_0, α and n_{ik} for every component $k = 1, \ldots, K$.
2. Prove the convexity of the optimization problem.
3. Obtain closed form solution for M-step. Express it in terms of $\eta^{-1}(\cdot)$ and use the fact that $\eta(\theta) = \nabla_\theta g(\theta)$ is an invertible function.

Recall the Multivariate Normal distributionfrom homework 5 (http://alex.smola.org/teaching/10-701-15/homework/hw5_sol.pdf). The distribution of Multivariate Normal $\mathcal{N}(\mu, \Sigma)$ is given by

$$p(x|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^{\mathrm{T}} \Sigma^{-1} (x-\mu)\right) \tag{3}$$

where $\mu \in \mathbb{R}^d$ and $\Sigma \succ 0$ is a symmetric positive definite $d \times d$ matrix.

The conjugate prior for Multivariate Normal Distribution can be parametrized as the Normal Inverse Wishart Distribution $\mathcal{NIW}(\mu_0, \kappa_0, \Sigma_0, \nu_0)$. The distribution is given by:

$$\begin{aligned} p(\mu, \Sigma;\ \mu_0, \kappa_0, \Sigma_0, \nu_0) &= \mathcal{N}(\mu|\mu_0, \Sigma/\kappa_0)\mathcal{W}^{-1}(\Sigma|\Sigma_0, \nu_0) \\ &= \frac{\kappa_0^{\frac{d}{2}} |\Sigma_0|^{\frac{\nu_0}{2}} |\Sigma|^{-\frac{\nu_0+d+2}{2}}}{2^{\frac{(\nu_0+1)d}{2}} \pi^{\frac{d}{2}} \Gamma_d(\frac{\nu_0}{2})} e^{-\frac{\kappa_0}{2}(\mu-\mu_0)^{\mathrm{T}}\Sigma^{-1}(\mu-\mu_0) - \frac{1}{2}\operatorname{tr}(\Sigma_0\Sigma^{-1})} \end{aligned} \tag{4}$$

Now, derive the Expectation-Maximiazation rules for the mixture of Mutivariate Normal, $\mathcal{N}(\mu_k, \Sigma_k)$ for $k = 1, \ldots, K$ and with the shared prior $\mathcal{NIW}(\mu_0, \kappa_0, \Sigma_0, \nu_0)$.

Question 3

Multivariate Normal

Write down update rules on E-step and M-step respectively for the Multivariate Gaussian with mixture parameters $\theta_k = (\mu_k, \Sigma_k)$. You can use the solutions from Question 1 and Question 2 and please refer to HW5 for Multivariate Guassians.

2 Gibbs Sampling

In the last sub-problem we used EM for inference and now we turn to Gibbs Sampling, another popular method. Gibbs sampling is a variety of MCMC sampling in which we cycle through all our latent random variables, resampling each conditioned on the currently sampled values of all other random variables.

Homework 9

Question 4

Write out the simple Gibbs sampler for the mixture model, i.e. derive the following conditional probabilities:

- $p(z_i|\text{rest})$
- $p(\theta_k|\text{rest})$
- $p(\pi|\text{rest})$

in terms of $\alpha, m_0, \phi_0, \phi(\cdot), g(\cdot), h(\cdot,\cdot)$. An useful notation would be $n_k = \alpha_k + |\{i : z_i = k\}|$ and $\phi_k = \phi_0 + \sum_{i:z_i=k} \phi(x_i)$. *Hint: Update equations from HW5 might come in handy!*

However, we can do a better job by collapsing out θ_k and π. In general also, collapsing helps the chain mix faster (a consequence of Rao-Blackwell theorem).

Question 5

1. After collapsing out θ_k and π, what would be the Markov blanket for a given z_i?
2. Using this information derive the conditional probability $p(z_i|\text{rest})$. *Hint: HW5 Posterior predictive might come in handy!*

Thus, we need to take care of only two invariants, namely n_k and ϕ_k per component in the inference procedure and a neat program can be written.

Question 6

Your task is to complete the generic program given in Algorithm 1 for Gibbs sampling of mixture models for T iterations.

Algorithm 1 Collapsed Gibbs sampling for mixture models

Initialize $\boldsymbol{z}$ randomly and evaluate initial counts n_k and statistics ϕ_k.
$t \leftarrow 0$
while $t \leq T$ **do**
  for $i = 1 \rightarrow N$ **do**
    Remove datum from current component and update statistics: _____
    Sample z_i using the PMF stored in $p[k] \leftarrow$ _____
    Add datum to the new component and update statistics:_____
  end for
  $t \leftarrow t + 1$
end while

Homework 9

Last but not the least, we derive Gibbs sampler for the Multivariate Gaussian with a Normal Inverse Wishart. Follow the notations of previous sub-problem or HW5.

Question 7

Write out the Gibbs Sampler, i.e. $p(z_i|\text{rest})$ explicitly in terms of data and hyper-parameters $\alpha, \mu_0, \kappa_0, \Sigma_0, \nu_0$. What would be the posterior means of mixture parameters $\theta_k = (\mu_k, \Sigma_k)$ given a sample $\mathbf{Z}$ obtained from our favourite collapsed Gibbs sampling algorithm?

Hopefully, this homework threw light on the commonly used finite mixture models. In all our examples so far, we have chosen some fixed K to be the dimension of π, which effectively enforces a fixed number of clusters. However, we would like to point out a major drawback of this model: the number of clusters present in the data needs to be known apriori, which is not the case often in practice. A novel workaround is to allow unboundedly many clusters, i.e. take $K \to \infty$. This yields a Bayesian nonparametric model known as the Dirichlet Process mixture. Extending our samplers to sample from a mixture of potentially infinitely many components requires only minor changes to the code; see e.g. section 3 of this nice paper http://www.stat.columbia.edu/npbayes/papers/neal_sampling.pdf.