STATS 700 Bayesian Inference and Computation **Fall 2017**

Lecture 20: Nov 29, Dec 4

Lecturer: CHEN Yang *Scribe: Greg Hunt*

20.1 Bayesian Non-parametrics

We say that a method is **parametric** if the number of parameters is the same for each different sample size. On the other hand we say a method is **non-parametric** if the number of parameters grows with the sample size. For example, We have a couple of parameters for each sample point. Alternatively, non-parametric models have an infinite dimensional parameter space.

Bayesian non-parametric models have been posited to solve lots of different problems. For example, a gaussian process for regression, chinese restaurant process for clustering, or a hierarchical dirichlet process for topic modeling.

20.1.1 Dirichlet Process

A dirichlet process is a bayesian non-parametric method that can be used as an alternative to parametric methods. Basically we have a prior of a set of distributions rather than constraining our data to be in a certain family.

We can simulate a Dirichlet process by picking a base distribution H, scaling parameter α, drawing X_1 from H and then with probability $\alpha/(\alpha+n-1)$ drawing X_n from H and with probability $n_x/\alpha+n-1$ setting $X_n = x$ where n_x is the number of previous observations where $j < n$ and $X_j = x$.

The Dirichlet process mixture is specified with some $G \mid \alpha, G_0 \sim D(\alpha, G_0)$ and then

$$X_i \overset{i.i.d.}{\sim} \sum_j p_j f(x \mid \phi_j).$$

We can sample MCMC from a DP mixture with the following steps:

1. $u_i \overset{ind.}{\sim} Unif(0, p_{r_i})$
2. sample $V_{1:q}$ for which $p_q > u_i$

$$V_h \sim B(1+m_h, \alpha + \sum_{k>h} m_k), m_h = \#\{i : r_i = h\}.$$

3. $p(\phi_h) \propto \prod_{r_i=h} f(x_i|\phi_h) G_0(d\phi_h)$.
4. $P(r_i = h) \propto f(x_i \mid h)$ for h such that $p_h > u_i$

20.1.1.1 Extensions

There are extensions to the DP like the Hierarchical Dirichlet Process where we have a collection of data sets $D_1, \ldots, D_m$ each of which is a mixture model but share a common density f and have measures $G_1, \ldots, G_m$. The proposed model is

$$G_1, ..., G_m \sim D(\alpha, G_0)$$

and $G_0|\gamma, H0 \sim D(\gamma, H_0)$.

20.1.2 Gaussian Process

The basic idea is to extend the gaussian distribution to infinite dimensions. It's possible to show that the GP is entirely determined by its mean and covariance function. There are many choices for these functions.

GPs have lots of applications, including GP regression, integration, optimization, unsupervized learning and much much more. We can use GPs for classifction.

Classification can either be broken down into

1. Generative approaches
2. Descriptive approaches

GP classifcation really falls under the second of these two. The main idea of a GP classifier is to have

1. A latent function f that tells us about how the classes are different
2. a map ϕ of f to the probability of the class

The solution is then basically follows as (1) assume a GP proior for f, compute $p(f \mid X, y)$, (2) compute this again for the test data $p(f* \mid X*, f, X)$ (3) compute the probability for the test data $\pi*$

20.2 Variational Inference & Approximate Bayesian Computation

VI is another way of calculating the posterior for interesting models when we can't do it easily. For example, a bayesian mixture of Gaussians since they are comuputationally intractable.

VI is an idea from CS that takes advantage of optimization and is much faster than MCMC. The main idea is to use the KL divergence to find densities closest to a target family.

This has lots of appliations in areas like comp. biology, computer vision, robotics, neuro. or NLP. The main idea is

1. Pick a family of dists over the latent variables q, parameterized by some variational parameter ν
2. find the ν so that q is closest to the true posterior using KL div
3. use that q as an approx for the posterior

There are different ways of doing this, for example Mean fidle Variational Inference:

1. group hidden variables in a nice way so that the family factorizes
2. then cycle through a coordinate ascent to find the max (e.g. ELBO)

Some concerns that might come up here are (1) numerical stability (2) convergence possibilties (3) initalization.

References

References

[1] Teh, Yee Whye. Dirichlet process. Encyclopedia of machine learning. Springer US, 2011. 280-287.

[2] http://stat.columbia.edu/ porbanz/npb-tutorial.html

[3] http://www.gaussianprocess.org/gpml/chapters

[4] Csillry, Katalin, et al. Approximate Bayesian computation (ABC) in practice. Trends in ecology & evolution 25.7 (2010): 410-418.

[5] Marin, Jean-Michel, et al. Approximate Bayesian computational methods. Statistics and Computing (2012): 1-14.

[6] Toni, Tina, et al. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems.

[7] Journal of the Royal Society Interface 6.31 (2009): 187-202. Del Moral, Pierre, Arnaud Doucet, and Ajay Jasra. An adaptive sequential Monte Carlo method for approximate Bayesian computation. Statistics and Computing 22.5 (2012): 1009-1020.