Statistical Machine Learning
Reading Assignment 3 Report

Alireza Sadeghi - Seyed Mohsen Shojaee

June 3, 2016

1 Introduction

The Bayesian approach to statistical problems has yielded very promising results. However in the context of nonparametric problem is has been quite unsuccessful. One reason for this problem is the unsatisfied need for a manageable prior. There are two essential properties for a prior distribution to work well in a nonparametric setting. The Bayesian approach to statistical problems has yielded very promising results. However in the context of nonparametric problem is has been quite unsuccessful. One reason for this problem is the unsatisfied need for a manageable prior. There are two essential properties for a prior distribution to work well in a nonparametric setting.

- The support of prior should be large to allow for many possible posteriors.
- Posterior on observed data should be manageable analytically.

Dirichlet distribution introduced in [**?**] a prior (probably the first one) for which the two properties above are simultaneously satisfied.

There are two alternate ways to define Dirichlet process. One by describing the distribution on any finite subset and rely on Kolmogorov's consistency theorem for the existence of the process. The second way is by generalizing Dirichlet distribution to infinite dimension.

In this section, we first review the Dirichlet distribution. Then we provide two formal definitions and finally give some insights about capabilities and shortcomings of Dirichlet Process i.e. the problems it can and can not model.

1.1 Dirichlet Distribution

A k-dimensional Dirichlet distribution is essentially a distribution over all probability measures on k possible outcomes. For Example a 6-dimensional Dirichlet can be used to model the probability distribution of dices. The formal definition is as follows. Let $X_1, X_2, \ldots, X_k$ be Gamma random variables with parameter $(\alpha_i, 1)$. A Dirichlet distribution with parameter $(\alpha_1, \ldots, \alpha_n)$ is defined as distribution of $Y = (Y_1, \ldots, Y_n)$, where

$$Y_i = \frac{Z_i}{\sum_{i=1}^{k} Z_j} \tag{1}$$

As can be seen in the above equation, Y will lie on the $k-$simplex; thus interpreted as a probability distribution over a sample space with k members.

2 Dirichlet Process

We now turn to definition of Dirichlet process. Let $\mathcal{X}$ be a measurable space and $\mathcal{X}$ a $\sigma-$field on it. The process is defined by specifying the joint distribution of any Sequence $A_1, \ldots, A_n \in \mathcal{A}$.

To this end, it is sufficient to specify the distribution on every measurable portioning $B_1, \ldots, B_n$. (We say portioning $(B_1, \ldots, B_n)$ to be measurable if $\forall i, j : B_i \in \mathcal{A}$ and $B_i \cap B_j = \Phi$ and $\bigcup j = 1^n B_j = \mathcal{X}$) From these distributions the distribution for arbitrary sequence $A_1, \ldots, A_n \in \mathcal{A}$ can be uniquely determined. We are almost ready to give a formal definition for Dirichlet distribution, it only remains the parameter of the process which should be a non-null finite measure on $\mathcal{X}$. We use α to denote the parameter.

We say a random process Q is a Dirichlet process with parameter α if for every finite measurable portioning $B_1, \ldots, B_n$ the distribution of $(P(B_1), \ldots, P(B_n))$ is Dirichlet with parameter $(\alpha(B_1), \ldots, \alpha(B_n))$.

An interesting property of Dirichlet distribution that can be considered a version of large support desired property is stated below.

Let Q be a Dirichlet distribution, then for every fixed probability measure T on $\mathcal{X}$ sequence of measurable sets $A_1, \ldots, A_n \in \mathcal{A}$ and for every $\epsilon > 0$:

$$\mathbb{P}(|P(A_i) - T(A_i) < \epsilon) > 0 \tag{2}$$

The following theorem ensures the posterior manageability for the Dirichlet process too,

Therom 1 Let P be a Dirichlet process on $(\mathcal{X}, \mathcal{A})$ with parameter α and let $X_1, \ldots, X_n$ be a sample of size n from P. Then conditional distribution of P given $X_1, \ldots, X_n$, is a Dirichlet process with parameter $\alpha + \sum_1^n \delta_{X_i}$.

3 Sampling

Let y denote the observations or data, and let θ denote the parameter or set of parameters by which the data are to be summarized. Bayesian methods combine prior evidence on the parameters contained in the density $p(\theta)$ with the likelihood $p(y|\theta)$ to produce the entire posterior density $p(\theta|y)$ of $p(\theta)$.

From the posterior density one may extract any information, not simply "the most likely value" of a parameter, as with maximum likelihood (ML) estimators. However, until the advent of Monte Carlo Markov Chain methods it was not straightforward to sample from the posterior density, except in cases where it was analytically defined. Monte Carlo Markov Chain (MCMC) methods are iterative sampling methods that allow sampling from the posterior distribution, meaning $p(\theta|y)$.

Although MCMC methods can be encompassed within the broad class of *Monte Carlo* methods, they must be distinguished from conventional Monte Carlo methods that generate independent simulations $\{u_(1), u_(2), \ldots u_(T)\}$ from a target density $\pi(u)$. From such simulations the expectation of a function $g(u)$ under $\pi(u)$, namely

$$\mathbb{E}_\pi[g(u)] = \int g(u)\pi(u)du$$

is estimated by taking the average of the function $g(y)$ evaluated at the sampled u^t, namely

$$\mathbb{E} = \sum_{t=1}^{T} g(u^{(t)})/T$$

Under independent sampling the Monte Carlo estimator g tends to $\mathbb{E}_\pi[g(u)]$ as $T \to \infty$. However, suppose we were to take $\pi(\theta) = p(\theta|y)$ as the density we wanted to find expectation from. We cannot use conventional independent Monte Carlo sampling, as this form of sampling from a posterior density $p(\theta|y)$ is not usually feasible.

When suitably implemented, MCMC methods offer an effective way to generate samples from the joint posterior distribution, $p(\theta|y)$. The "target density" for MCMC samples is the posterior density $\pi(\theta) = p(\theta|y)$, and MCMC sampling is specially relevant when the posterior cannot be stated exactly in analytic form,

e.g. when the prior density assumed for θ is not conjugate with the likelihood $p(y|\theta)$.

3.1 Forward Sampling in Bayesian Networks

Our goal in forward sampling is to compute the marginalized probability distribution over the desired random variable. So our goal is to compute terms of the form $p(Y = y)$. In order to do so, we generate samples from the Bayesian Network and then compute the fraction in which $Y = y$.

Sampling from a Bayesian Network is done in a top-down manner. We start from the root nodes (nodes that have no parents associated with them), and sample each variable from a node's CPD (conditional probability distribution). We then move down the Bayesian Network, sampling subsequent nodes, with respect to the previously sampled parents. Upon getting to the bottom of the graphical model, we have successfully sampled a full realization of the join distribution of parameters. This process can be repeated to generate more samples and thus arrive at more accurate estimates of our favored distribution.

With a little help from the discipline of *Learning Theory*, we can introduce several interesting bounds on the number of samples, required to estimate the value of a parameter, while satisfying specific precision requirements. Two of the most important bounds are:

- *Hoeffding's Bound (Additive)*

$$P_D(T_D \notin [p - \epsilon, p + \epsilon]) \leq 2e^{-2M\epsilon^2} \tag{3}$$

$$M \geq \frac{\ln(2/\delta)}{2\epsilon^2} \tag{4}$$

- *Chernoff's Bound (Multiplicative)*

$$P_D(T_D \notin [p(1 - \epsilon), p(1 + \epsilon)]) \leq 2e^{-Mp\epsilon^2/3} \tag{5}$$

$$M \geq 3\frac{\ln(2/\delta)}{p\epsilon^2} \tag{6}$$

Where in both bounds, M is the number of samples, δ is equal to the right hand side of the (1) and (3) inequalities and T_D is the estimator for the parameter.

3.2 Rejection Sampling in Bayesian Networks

Our goal in rejection sampling is to compute the marginalized probability distribution over a desired random variable, while taking into account and observed evidence, so our goal actually is to compute terms of the form $P(Y = y|E = e)$.

When we have some evidence and we want to do queries with respect to those evidences, we need to take a slightly different approach. The approach we take in this configuration of the problem is called *Rejection Sampling*.

In rejection sampling, we start by sampling the Bayesian Network exactly as we did in normal forward sampling, and finish by computing the fraction where $Y = y$. The only difference is that we throw away all those samples that conflicts our observed evidence.

As obvious, in this case more computational power is required since the number of required samples grows exponentially with the number of observed variables. (since we're keeping a sample with only probability p(e)).

3.3 Markov Chain

A Markov chain defines a probabilistic transition model $T(x \to x')$ over states while satisfying the following property:

$$\forall x : \sum_{x'} T(x \to x') = 1$$

3.3.1 Markov Property

For each $n \geq 1$, if A is an event depending only on any subset of $\{X_{n-1}, X_{n-2}, \ldots, 0\}$, then, for any states i and j in S,

$$P(X_{n+1} = j|X_n = i \quad and \quad A) = P(X_{n+1} = j|X_n = i)$$

More generally, for each $n \geq 1$ and $m \geq 1$, if A is as defined previously, then for any states i and j in S:

$$P(X_{n+m} = j|X_n = i \quad and \quad A) = P(X_{n+m} = j|X_n = i)$$

3.3.2 Temporal Dynamics

The probability of being in state x' at the time step $t + 1$, is the sum over all states, from which one can get to x', multiplied by the probability of the

transition occurring, namely,

$$P^{(t+1)}(X^{(t+1)} = x') = \sum_x P^{(t)}(X^{(t)} = x)T(x \to x')$$

3.3.3 Stationary Distribution

Ultimately, as the process evolves, the probability distribution nearly equalizes. Meaning:

$$P^(t)(x') = p^(t+1)(x') = \sum_x P^{(t)}(x)T(x \to x')$$

So, we introduce a notion of a *Stationary Distribution* as follows

$$\pi(x') = \sum_x \pi(x)T(x \to x')$$

But not all Markov Chains converge to a stationary distribution, actually only *Regular Markov Chains* will eventually arrive at a stationary distribution. We define a Markov chain to be *Regular* in case there exists a k, such that, for every x, x', the probability of getting from x to x' in exactly k steps is a > 0. If this condition holds, we know that the Markov chain converges to a unique stationary distribution regardless of the start state. A Markov chain is regular if both the following conditions hold (Sufficient conditions):

- Every two states are connected
- For every state, there exists a self-transition

3.3.4 MCMC

Our goal using sampling, as explained before is to find the probability distribution $P(X \in S)$ when P is too hard to be compute analytically or to sample directly from. In this case we construct a Markov chain, whose unique stationary distribution is P. Since we only want to use samples that are sampled from a distribution close to P, we only start collecting samples after the chain has run long enough to be *mixing*. So in summary:

- For $c = 1, \ldots, C$
  - Sample $X^{(c,0)}$ from $P^{(0)}$
- Repeat until mixing
  - For $c = 1, \ldots, C$

        * Generate $X^{(c,t+1)}$ from $T(x^{(c,t)} \to x')$
    - Compare window statistics in different chains to determine mixing (It's a heuristic method to check whether the Markov chain has started to mix)
    - $t = t + 1$

Now we know the Markov chain is mixing (we have transitioned in the Markov chain long enough to be sure we've reached its stationary distribution) and with that, we do actual MCMC as follows:

- Repeat until sufficient samples are obtained
    - $D = \emptyset$
    - For $c = 1, \ldots, C$
        * Generate $X^{(c,t+1)}$ from $T(x^{(c,t)} \to x')$
        * $D = D \bigcup X^{(c,t+1)}$
    - $t = t + 1$
- Let $D = x[1], \ldots, x[M]$
- Estimate $E_P[f] = \frac{1}{M} \sum_{m=1}^{M} f(x[m])$

3.4 Summary of MCMC

- Pros:
    - Very general purpose
    - Often easy to implement
    - Good theoretical guarantees when $t \to \infty$
- Cons:
    - Lots of tunable parameters / design choices
    - Can be quite slow to converge
    - Difficult to tell whether it's working (detecting mixing)

4 Mixture of Dirichlet Process Models

In this section we first introduce mixtures of Dirichlet process models and then briefly review three MCMC Sampling Algorithms for DPMM from [**?**].

Consider that y_i's are the data we want to model and there is mixture assumption in data that data points come from a number of mixtures. The data and underlying mixtures can be modeled by Dirichlet process:

$$\begin{aligned} y_i|\theta_i &\sim f_{\theta_i} \\ \theta_i|G &\sim^{iid} G \\ G|\nu, M &\sim DP(M, G_\nu) \\ (\nu, M) &\sim \pi. \end{aligned}$$

In the equation above, f is the family of distribution for each mixture e.g. Gaussian, and θ_i's are the parameters of the i-th mixture. These parameters are themselves governed with a Dirichlet Process G and there is a prior (ν, M) on G. We also introduce the notation of cluster membership indicators, $s_i = j$ iff $i \in S_j$ the j-th cluster.

We now present three MCMC sampling algorithms for the model above, with the conjugate prior assumption. Escobar (1988) proposed the first posterior Gibbs sampler for the DPM model, based on transition probabilities that update θ_i by draws from the complete conditional posterior $P(\theta_i|\theta_{-i}, y)$. However, this Gibbs sampler suffers from a slowly mixing Markov chain. The reason for this is that θ's are updated independently. While in natural data there are often groups of observations that with high probability are associated with the same θ. Since the algorithm cannot change the θ for more than one observation simultaneously, changes to the values of θ in such group can occur only rarely, as they require passage through low probability intermediate state in which observations in the group have different values for θ. This problem can be avoided if we first integrate out the θ and sample the marginal distribution where cluster assignments s_i are sampled instead of mixture parameters θ. In this model new values for θ_i which we denote by θ_i^* are drawn form $P(\theta_i^*|s, y)$ which can be modeled by:

$$P(\theta_i^*|s, y) \propto G(\theta_i^*) \prod_{i \in S_j} f_{\theta_i^*}(y_i) \tag{7}$$

The posterior presented above for θ_i^* is a parametric model with prior $G(\theta)$. To complete this model we should define the posteriors for cluster assignments.