Statistical Machine Learning
Reading Assignment 3 Report

Alireza Sadeghi - Mohsen Shojaee

May 3, 2016

1 Introduction

2 Dirichlet Process

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 $\{u_(1), u_(2), \dots u_(T)\}$ 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)*

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

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

- *Chernoff's Bound (Multiplicative)*

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

$$M \geq 3\frac{\ln(2/\delta)}{p\epsilon^2} \tag{4}$$

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

4 Gaussian Process