Markov random fields, the Ising model, and Gibbs sampling

Ye Li

These notes give a short description of Markov Random Fields, the Ising model for images, and an introduction to Gibbs Markov Chain Monte Carlo (MCMC) in the context of image. These notes assume you're familiar with basic probability and graphical models.

1 Markov Random Field

MRF definition: A Markov Random Field (MRF) or undirected graphical model is a graphical model of a set of undirected random variables having a Markov property, i.e., the conditional distribution of a random variable depends only on its neighbors. The graphical model is consisted of undirected edges(encodes conditional dependencies), nodes (random variables), and observed data. We can model the MRF by a posterior conditional distribution, which is a Gibbs distribution (according to the Hammersley-Clifford theorem), by specifying an energy function. The energy function has two terms in it: 1) unary potential and 2) pairwise potential. The unary potential term solely models the relation between the observed data point and its label and the pairwise potential then models the relation between the label of interest and the labels neighboring it. For example, in a 2-D image the observed data can be the pixel values and the nodes can be the labels associated with each pixel value. To find the the posterior conditional distribution,$P(label|observedData)$, we will need to use sampling techniques such as Gibbs sampling as exact inference on MRF is usually hard because the graph is not a tree and so we can't use any of the sum-pushing tricks for the sum-product algorithm. In the energy function of the Gibbs distribution, the pairwise potential term can include prior assumptions about the local context of the labels of the nodes to impose the neighboring pixels to have similar labels.

2 The Ising model

The Ising model is specified by a Gibbs distribution $P(\underline{\mathbf{S}}|\underline{\mathbf{I}}) = \frac{1}{Z}exp(-E(\underline{\mathbf{S}};\underline{\mathbf{I}}))$ where the energy $E(\underline{\mathbf{S}};\underline{\mathbf{I}})$ can be expressed by:

$$E(\underline{\mathbf{S}};\underline{\mathbf{I}}) = \sum_x (S(x) - I(x))^2 + \lambda \sum_x \sum_{y \in Nbh(x)} (S(x) - S(y))^2$$

Here, $Nbh(x)$ denotes the set of pixel indices neighboring x, $S(x) \in 0, 1$ (states), and $I(x) \in [0, 1]$ (the image).

The Ising model captures spatial context. From the second term of the energy function, we can see that it punishes differences between neighbouring nodes, forcing that the solution to be smooth in most regions, but in many cases when we are doing inference with this model we would like to keep certain structures. For example in an edge detection task, we would certainly want to keep the edges and so the edges should not be punished too harshly. The extend of this punishment is controlled by λ, which weights the importance of the second term relative to the first. The task of the second term (the pairwise potential term) can actually be understood to force pixels to have the same label and it only operates on pixels that are neighbors of x. This follows the nature of a typical image, that pixels nearby tend to have the same label or context. Thus this model helps capture/distinguish context of an image spatially.

From the posterior, we can calculate the likelihood and prior by this relation: $P(\underline{\mathbf{S}}|\underline{\mathbf{I}}) \propto P(\underline{\mathbf{I}}|\underline{\mathbf{S}})P(\underline{\mathbf{S}})$. We can find the likelihood distribution and the prior by dividing posterior as follows ($Z_1 * Z_2 = Z$):

The likelihood distribution:

$$P(\underline{\mathbf{I}}|\underline{\mathbf{S}}) = \frac{1}{Z_1} exp(-\sum_x (S(x) - I(x))^2)$$

The prior distribution:

$$P(\underline{\mathbf{S}}) = \frac{1}{Z_2} exp(-\lambda \sum_x \sum_{y \in Nbh(x)} (S(x) - S(y))^2)$$

3 Gibbs sampling

3.1 Gibbs sampler on the Ising model

For a 1-D Ising model, we'll use $\underline{\mathbf{S}}$ to represent the set $\{S_1, ...S_n\}$ for labels of observed data and $\underline{\mathbf{I}}$ to represent the set $\{I_1, ...I_n\}$ for observed data. We have the target distribution $P(\underline{\mathbf{S}}|\underline{\mathbf{I}})$ and now we would like to design a Gibbs sampling algorithm to generate samples approximately from the target distribution. First we need to choose some random values to initiate $\underline{\mathbf{S^0}}$ (the supper script indicates iteration). After initialization we now need to move on and do the first iteration. We have to draw the n different $S's$ for $\underline{\mathbf{S^1}}$ separately. So we first draw S_1 from $P(S_1|S_2^0, ...S_n^0, \underline{\mathbf{I}})$, while keeping S_2^0 to S_n^0 fixed, given $\underline{\mathbf{I}}$. Then we draw S_2 from $P(S_2|S_1^1, S_3^0...S_n^0, \underline{\mathbf{I}})$, given the newly sampled S_1^1, S_3^0 to S_n^0 , and $\underline{\mathbf{I}}$. For S_3^1 to be precise, we draw S_3 from $P(S_3|S_1^1, S_2^1, S_4^0...S_n^0, \underline{\mathbf{I}})$, given the newly sampled S_1^1,S_2^1, S_4^0 to S_n^0 from the previous iteration (initialization) , and $\underline{\mathbf{I}}$. We can generalize it to S_4,...S_n but now the question is how do we actually draw sample (states; 1 or 0, edge or non-edge) from the conditional distribution.

To do so, we will need to derive the Gibbs sampling distribution (probabilities corresponding to all states or configurations if there are multiple r.v.s), which is the conditional distribution from which the samples are drawn. Here, we can derive the sampling distribution for first sample S_1^1:

$$P(S_1|\underline{\mathbf{S^0}}_{\neg S_1}, \underline{\mathbf{I}}) = \frac{P(\underline{\mathbf{S^0}}|\underline{\mathbf{I}})}{P(\underline{\mathbf{S^0}}_{\neg S_1}|\underline{\mathbf{I}})}$$

To compute it with the actual values, namely, $\underline{\mathbf{S^0}} = \{s_1, ...s_n\}$, we have (a lot of terms cancel out in the numerator and denominator because of the exponential in the energy function of the Ising model and the only term left out is the one that's operating on the pixel x_1):

$$= \frac{P_{(\underline{\mathbf{S^0}}|\underline{\mathbf{I}})}(s_1, ...s_n|\underline{\mathbf{I}})}{P_{(\underline{\mathbf{S^0}}_{\neg S_1}|\underline{\mathbf{I}})}(s_2, ...s_n|\underline{\mathbf{I}})} = exp(-\sum_{x_1}(S(x) - I(x))^2 - \lambda \sum_{x_1} \sum_{y \in Nbh(x)} (S(x) - S(y))^2)$$

To actually calculate the probability of $P(S_1|\underline{\mathbf{S^0}}_{\neg S_1}, \underline{\mathbf{I}})$, we just need to plug in the states of S_1 and calculate the following probabilities: $P(S_1 = 1|\underline{\mathbf{S^0}}_{\neg S_1}, \underline{\mathbf{I}})$ and $P(S_1 = 0|\underline{\mathbf{S^0}}_{\neg S_1}, \underline{\mathbf{I}})$, and making sure they sum up to 1. With these probabilities, we can then sample S_1^1.

We can then generalize to S_2^1,....S_n^1 to get $\underline{\mathbf{S^1}}$. The "Markov" property comes into play in Gibbs sampling between the iterations. The set of samples ($\underline{\mathbf{S^j}}$) at any iteration is not an independent sequence; there is a Markov chain in that such that the draw at time j depends only on samples at time $j - 1$.

Here $P(S(x)|S(y) : y \in Nbh(x)|\underline{\mathbf{I}})$ is like a generalized form of $P(S_1, \underline{\mathbf{S^0}}_{Nbh(1)}|\underline{\mathbf{I}})$, which only is for the first pixel. The calculation of this sampling distribution is more feasible than the full conditional distribution as 1) this is a distribution of only one random variable and hence it has much less of configurations (in fact just the states of that r.v.) as compared to a full joint distribution on all random variables. 2) there is the neighboring concept so that we only need calculate the S(neighbors) in the pairwise potential term. So the computation here is much less. Conversely, if we want to sample from the full joint conditional distribution, we would need to obtain probabilities of every single possible configuration of the random variables in order to calculate the full joint conditional distribution and then sample from it. This is sometimes hard or impossible because as the number of random variable and states increase the number of configurations goes exponentially and thus the calculation becomes intractable.

3.2 What theoretical results guarantee that Gibbs sampling will converge to samples from the Gibbs distribution?

Gibbs sampling is one of the MCMC algorithms for doing approximate inference. Just like other MCMC algorithms such as the Metropolic-Hastings algorithm, the theory of MCMC guarantees that the samples converge in distribution to a draw from the target joint posterior after the "burn-in" period (Gilks et al., 1996; also see the Computational Cognition Cheat Sheet on Metropolis-Hastings sampling). The proof uses law of large numbers and the central limit theorem and says that the MCMC algorithms allow us to calculate the same monte carlo approximation to the integral below.

$$\frac{1}{J}\sum_{j=1}^{J} h(x^{(j)}) -> E_f[h(X)] = \int_x h(x)f(x)dx$$

where, $f(x)$ is the target density, $x^{(j)}$ is one configuration of the random sequence X, and $h(.)$ is a function operated on X. The theory basically shows that if we take the average of all J samples then that will converge to the expectation in the middle.