Lecture 5: Stat 238. Winter. 2012

A.L. Yuille

2012-01-20

1 Lecture 4

1. MRF Models: Labeling and Grab-Cut as Examples.
2. Unary and Binary Potentials.
3. Functionals and MRFs.
4. Gibbs Sampling and MCMC.
5. Variational/Mean Field Theory. (CCCP)
6. Example: Geman and Geman model (mixture).
7. Appendix: MCMC.

2 MRF Models

Markov Random Fields (MRFs) are undirected graphical models. The graph $G = V, E$ where V denotes the graph nodes and E the edges between nodes. State variables $W = \{w_i\}$ are defined on the graph nodes – i.e. w_i is the state of node $i \in V$. The edges define a neighborhood $Nbh(i)E$ – so that $j \in Nbh(i)$ if $(i, j) \in E$. The MRF specifies a probability distribution $P(W)$ (or $P(W|I)$ conditioned on the image). The MRF satisfies the Markov condition $P(w_i|W_{/i}) = P(w_i|\{w_j : j \in Nbh(i))$, where $W_{/i}$ denotes the state of all nodes except i – in other words, node i is directly influenced by nodes in its neigborhood $Nbh(i)$ but only indirectly by other nodes. The Hammersley-Clifford theorem states that an MRF can be expressed by a Gibbs distribution – $P(W) = \frac{1}{Z}\exp\{-E(W)\}$, where $E(W)$ is the energy. For more, Google MRFs or read the handout.

In this lecture we treat MRFs where the graph is the image lattice – i.e. nodes are indexed by (x, y) – and the neighborhoods are nearest neighbor – i.e. $Nbh(x, y) = \{(x + 1, y), (x - 1, y), (x, y + 1), (x, y - 1)\}$. Note: these neighborhoods are probably too small to capture the statistical structure of the problem (see next lecture) and they are mostly used because they make the inference task easier – i.e. to estimate $\hat{W} = \arg\max P(W|I)$. They are related to the

energy functionals described in the last lecture (see section – functionals and MRFs). The energy function E is a sum of potentials ϕ defined over the nodes and the edges:

$$P(W|I) = \frac{1}{Z}\exp\{-E(w,I)\}$$

$$E(w,I) = \sum_{(x,y)\in D}\phi(w(x,y),I) + \sum_{(x,y)\in D}\sum_{(x',y')\in Nbh(x,y)}\phi(w(x,y),w(x',y'),I). \quad (1)$$

The *unary potentials* $\phi(w(x,y),I)$ depend only on the state of the individual nodes $w(x,y)$, while the binary potentials $\phi(w(x,y),w(x',y'),I)$ depend on the states of two nodes (or the edge that connects them). If the model only contains unary terms then it reduces to the factorized model studied in lectures 2 and 3 – i.e. $P(W|I) = \frac{1}{Z}\exp\{-\sum_{(x,y)\in D}\phi(w(x,y),I)\} = \frac{1}{Z}\prod_{(x,y)\in D}\exp\{-\phi(w(x,y),I)\}$. Z is the normalization, or partition, function. It is specified by $Z = \sum_w \exp\{-E(w,I)\}$ but this is usually impossible to compute exactly. The number of possible states W is $k^N - 1$, where k is the number of possible states of each $w(x,y)$ and N is the number of pixels in the image. It can be computed if the model is factorizable – only contains unary terms – then $Z = \prod_{(x,y)\in D}\sum_{w(x,y)}\exp\{-\phi(w(x,y),I)\}$, which requires only Nk computations.

In this lecture we focus on image labeling. There are two examples. The first is labeling a pixel as one of a discrete number of labels (e.g., sky, road, vegetation), which is an extension of the models described in lecture 2 and 3 (which used unary terms only). Now we introduce pairwise terms which allow us to include spatial context – e.g., sky pixels are likely to be next to other sky pixels. The second examples is grab-cut (see the handout) where the task is to label pixels as foreground or background. The idea is that a user wants to remove an object from a photograph – so you draw a boundary which surrounds the object, and then want to detect the silhouette directly.

For these applications the unary potentials $\phi(w(x,y),I)$ provide local evidence for the states $w(x,y)$ of each pixel (x,y) conditioned on the image. These can be learnt by the techniques described in lectures 2 and 3 using training data (for grab-cut the boundary drawn by the user provides training data, contaminated because the object silhouette is not specified precisely by the user).

The binary potentials for these applications are specified as follows (ideally we would learn the unary and binary potentials together – see next lecture). They capture the intuition that neighboring pixels tend to have the same labels, except if these pixels have very different intensities (suggesting an edge between them). This can be specified by:

$$\phi(w(x,y),w(x',y'),I) = A\{1 - \mathcal{I}(w(x,y),w(x',y'))\},$$

$$\phi(w(x,y),w(x',y'),I) = A\{1 - \mathcal{I}(w(x,y),w(x',y'))\}\exp\{-B|I(x,y) - I(x',y')|\}. \quad (2)$$

Here A and B are constants. $\mathcal{I}(.,.)$ is the identity function which takes value 1 if both terms are identical, and value 0 is they are different.

More generally, MRF models involve three aspects: (i) *representating* the problem by specifying the graph structure and the state variables, (ii) *learning* or specifying the potentials of the distribution, and (iii) performing *inference* to estimate $\hat{W} = \arg\max P(W|I)$ or some other estimator of W.

We have discussed how to represent the problem of image labeling, we have specified the potentials, and we will describe some techniques to do inference later in this lecture. Other lectures will describe other applications, other inference algorithms, and learning.

3 Functionals and MRKs

The previous lecture described formulating vision problems in terms of minimizing functionals. How does this relate to MRFs?

Consider the TV norm model: $E[J;I] = \int_D |\vec{\nabla} J| d\vec{x} + \frac{\lambda}{2} \int_D (J(\vec{x}) - I(\vec{x}))^2 d\vec{x}$ defined over $D \subset R^2$. Suppose we approximate D by a discrete lattice of points $\{(x,y) : (x,y) \in L\}$. This replaces the data term $\int_D (J(\vec{x}) - I(\vec{x}))^2 d\vec{x}$ by $\sum_{(x,y)\in L} \{I(x,y) - J(x,y)\}^2$ and the smoothing term $\int_D |\vec{\nabla} J| d\vec{x}$ by $\sum_{(x,y)\in L} \sum_{(x',y')\in Nbh(x,y)} |J(x,y) - J(x',y'|$ (where the neigbhorhood is the nearest neighbors on the lattice). This gives a discrete energy:

$$E(J;I) = \sum_{(x,y)\in L} \{I(x,y) - J(x,y)\}^2 + \lambda \sum_{(x,y)\in L} \sum_{(x',y')\in Nbh(x,y)} |J(x,y) - J(x',y'| \tag{3}$$

We define a continuous-valued MRF by using the Gibbs distribution:

$$P(J|I) = \frac{1}{Z} \exp\{-E(J;I)\}. \tag{4}$$

Note that the neigborhood structure (i.e. graph edges) of the MRFs arise from the derivatives in the TV model (and the way we approximate derivatives by taking differences between pixel values at neighboring pixels). Broadly speaking, first order derivatives transform to nearest neighbor interactions on the lattice. Higher order derivatives translate the longer range interactions.

Note that this translation between functionals and continuous-valued MRFs does not guarantee that: (i) that the MRF is well-defined (e.g., normalizable), and (ii) the solution to the MRF is similar to the solution to the funcational (sometimes there is a simple limit result – e.g., as the number of lattice nodes become infinitesimally close together then the MRF solution converges to the functional solution – but this limiting argument becomes very tricky for some models, like Mumford and Shah).

The MRF version of the TV-norm model can be decomposed into two parts:

$$P(I|J) = \frac{1}{Z_1} \exp\{-\sum_{(x,y)\in L} (J(x,y) - I(x,y))^2\}, P(J) \frac{1}{Z_e} \exp\{-\sum_{(x,y)\in L} \sum_{(x',y')\in Nbh(x,y)} |J(x,y) - J(x',y'|\}.$$

Here $P(I|J)$ is the likelihood function of generative the data I if the state os J. The quadratic form means that it corresponds to assuming that the observed image I is given by J plus additive zero mean gaussian noise which is independent at each pixel.

$P(J)$ is the prior on J – it assumes that images are piecewise smooth (for justification see next lecture). Technically this an *improper prior* because it is not normalizable because it is specified only on the relative values $J(x,y) - J(x',y')$ (it is invariant to $J(x,y) \mapsto J(x,y) + K$ for any constant K and hence cannot be normalized).

$P(J|I) = \frac{P(I|J)P(I)}{P(I)}$ is the posterior distribution of J given I. This can be normalized – and this corresponds to the discretized version of the TV model.

Note that models like Geman+Geman and Blake+Zisserman are similar to the discretized version of the TV norm model. They are not convex – and hence harder to do inference with – but capture some aspects of images better (see handout).

Finally, we can convert a continuous-valued MRF into a discrete-valued MRF by discretizing $w(x,y)$ – e.g., setting $w(x,y) \in \{0,1,...,255\}$.

4 Gibbs Sampling

Gibbs sampling is a procedure to draw samples $W_1, ..., W_m$ from $P(W|I)$. These samples are likely to be from places where $P(W|I)$ and from them we can estimate properties such as $\arg\max P(|I)$ or $\sum_W WP(W|I)$. Gibbs sampling is a special case of Markov Chain Monte Carlo (MCMC) (see MCMC appendix).

Gibbs sampling initializes W at random and then repeats the following step until convergence – select a point (x,y) at random, then sample the state $w(x,y)$ from the conditional distribution $P(w(x,y)|\{w(x',y' : (x',y') \in Nbh(x,y))\})$. The procedure repeats until convergence. The output is a sample W_1. Then the procedure starts again to obtain a new sample W_2, and so on.

The conditional distribution of $w(x,y)$ can be calculated as:

$$P(w(x,y)|\{w(x',y' : (x',y') \in Nbh(x,y))\}) = \frac{\exp\{-\phi(w(x,y),I) - 2\sum_{(x',y')\in Nbh(x,y)} \phi(w(x,y),w(x',y'),I)\}}{\sum_{(x,y)} \exp\{-\phi(w(x,y),I) - 2\sum_{(x',y')\in Nbh(x,y)} \phi(w(x,y),w(x',y'),I)\}}. \quad (5)$$

Sampling from the conditional distribution is practical because of the simple form of the distribution. Its normalization factor can be computed (unlike the normalization constant of the original distribution $P(W|I)$ which is usually impossible to compute).

The most probable samples from $P(w(x,y)|...)$ will be those where $\phi(w(x,y),I)$ is small (i.e. driven by the data) and $\sum_{(x',y')\in Nbh(x,y)} \phi(w(x,y),w(x',y'),I)$ is small (i.e. the state $w(x,y)$ is consistent with the states of its neighbors).

A difficulty with Gibbs sampling, and with MCMC in general, is that it is difficult to be sure when it has converged although there are tests (Jun Liu).

5 Mean Field Theory and Variational Methods

Mean Field Theory attempts to find a probability distribution $Q(W)$ which approximates $P(W|I)$. $Q(W)$ takes a simple form – like $\prod_{(x,y)\in D} q_{(x,y)}(w(x,y))$ – so that estimating $\hat{W}_Q = \arg\max Q(W)$ is straightforward (e.g., $\hat{w}(x,y)_Q = \arg\max q_{(x,y)}(w(x,y))$).

$Q(W)$ is chosen to minimize the Kullback-Leibler divergence:

$$F(Q) = \sum_W Q(W) \log \frac{Q(W)}{P(W|I)}. \tag{6}$$

$F(Q) \geq 0$ with the property that $F(Q) = 0$ only if $Q(W) = P(W|I)$.

Hence MFT reduces to minimizing $F(Q)$ with respect to Q. This is a continuous minimization problem so we can use steepest descent techniques to estimate Q. If we assume the factorized form $Q(W) = \prod_{(x,y)\in D} q_{(x,y)}(w(x,y))$ then $F(Q)$ can be expressed as:

$$F(Q) = \sum_{(x,y)\in D} \sum_{w(x,y)} q_{(x,y)}(w(x,y)) \log q_{(x,y)}(w(x,y)) + \sum_{(x,y)\in D} q_{(x,y)}(w(x,y))\phi(w(x,y), I)$$
$$+ \sum_{(x,y)\in D} \sum_{(x',y')\in Nbd(x,y)} q_{(x,y)}(w(x,y)) q_{(x',y')}(w(x',y'))\phi(w(x,y), w(x',y'), I). \tag{7}$$

Note that the first term is a convex function – for many potentials the third term is concave (or can be made concave by a trick – Kosowsky and Yuille). Then we can apply CCCP to obtain the update rule:

$$q^{t+1}_{(x,y)}(w(x,y)) =$$
$$\frac{\exp\{-\phi(w(x,y),I) - 2\sum_{(x',y')\in Nbd(x,y)} q_{(x,y)}(w(x,y)) q_{(x',y')}(w(x',y'))phi(w(x,y), w(x',y'), I)\}}{\sum_{w(x,y)} \exp\{-\phi(w(x,y),I) - 2\sum_{(x',y')\in Nbd(x,y)} q_{(x,y)}(w(x,y)) q_{(x',y')}(w(x',y'))phi(w(x,y), w(x',y'), I)\}}. \tag{8}$$

This reduces $F(Q)$ but is not guaranteed to converge to the global minimum. Note that the MFT update rule is very similar to the Gibbs sampler. It can be shown (e.g., Amit) that the MFT can be derived by taking the expectation of the Gibbs sampler. It also combines the unary evidence for state $w(x,y)$ with the consistency of the neighbors.

If $F(Q)$ has many local minima then we can use a continuation method called deterministic annealing. This defines a family of $F_T(Q) = \sum_W Q(W) \log \frac{Q(W)}{P^{1/T}(W)}$, where $P^{1/T}(W)$ is the distribution $P(W)$ taken to the $(1/T)^{th}$ power and then normalized. For large T the distribution $P(W)$ is smooth and $F_T(Q)$ will have a single minimum.

6 Geman and Geman:

See handout.

7 MCMC Appendix

MCMC gives a way to sample from any distribution $P(\vec{x})$. This enables us to estimate quantities such as $\vec{x}^* = \arg\max P(\vec{x})$ or $\sum_{\vec{x}} \vec{\phi}(\vec{x})P(\vec{x})$. The advantage of MCMC is that it does not require knowing the normalization constant Z of the distribution $P(\vec{x}) = (1/Z)\exp\{-E(\vec{x})\}$. But MCMC is an art rather than a science.

A Markov chain is defined by transition kernel $K(\vec{x}|\vec{x}')$, such that $\sum_{\vec{x}} K(\vec{x}|\vec{x}') = 1$, $\forall \vec{x}'$ and $K(\vec{x}|\vec{x}') \geq 0$. We also require the constraint that for any $\vec{x}_0$ and $\vec{x}_N$ there exists a chain $\vec{x}_1, ..., \vec{x}_{N-1}$ such that $K(\vec{x}_i|\vec{x}_{i-1}) > 0$ for $i = 1, ..., N$ (i.e. so that the chain is *irreducible* – you can get to any state from any other state in a finite number of moves).

An MCMC for a distribution $P(\vec{x})$ is a special Markov chain where the transition kernel satisfies $\sum_{\vec{y}} K(\vec{x}|\vec{y})P(\vec{y}) = P(\vec{x})$ – i.e. the target distribution $P(\vec{x})$ is a fixed point of the chain. In practice, most MCMC are designed to satisfy the more restrictive *detailed balance* condition (which implies the fixed point condition):

$$K(\vec{x}|\vec{y})P(\vec{y}) = K(\vec{y}|\vec{x})P(\vec{x}). \tag{9}$$

To run MCMC we give an initial condition $\vec{x}_0$ and repeatedly sample from $K(\vec{x}'|\vec{x})$ to get a sequence $\vec{x}_1, ..., \vec{x}_t, ...$ so that for sufficiently large t $\vec{x}_t$ is a sample from $P(\vec{x})$.

7.1 Metropolis-Hastings and Gibbs Sampler

Metropolis-Hastings is an ansatz for constructing a transition kernel that obeys the detailed balance condition. It is specified by:

$$K(\vec{y}|\vec{x}) = T(\vec{y}|\vec{x}) \min\{1, \frac{P(\vec{y})T(\vec{x}|\vec{y})}{P(\vec{x})T(\vec{y}|\vec{x})}\}, \quad \text{for } \vec{y} \cdot \vec{x}) \tag{10}$$

where $T(\vec{y}|\vec{x})$ is a conditional distribution and $K(\vec{y}|\vec{y})$ is defined to ensure that $\sum_{\vec{y}} K(\vec{y}|\vec{x}) = 1$ is satisfied for all $\vec{x}$. It can be checked that this satisfies detailed balance. The form of $T(\vec{y}|\vec{x})$ must be chosen to ensure that this is irreducible.

Metropolis-Hastings can be thought of as a two stage process. First, use the *proposal distribution* $T(\vec{y}|\vec{x})$ to generate a *proposal* $\vec{y}$. Accept the proposal with *acceptance probability* $\min\{1, \frac{P(\vec{y})T(\vec{x}|\vec{y})}{P(\vec{x})T(\vec{y}|\vec{x})}\}$. In practice, the convergence speed of Metropolis-Hastings algorithms depends on whether good proposals can be found (we will return to this issue later in the course).

A key property of Metropolis-Hastings algorithms (and other MCMC) is that they do not require knowing the normalization constant Z of the distribution $P(\vec{x}) = (1/Z)\exp\{-E(\vec{x})\}$ (observe that Z cancels in the acceptance probability).

What is the intuition for Metropolis-Hastings? First, we sample from $T(\vec{y}|\vec{x})$ to propose a move to $\vec{y}$. We accept this move with certainty if $E(\vec{y}) < E(\vec{x}) + \log \frac{T(\vec{x}|\vec{y})}{T(\vec{y}|\vec{x})}$ (i.e. the energy decreases after allowing for the proposal probability).

But if $E(\vec{y}) > E(\vec{x}) + \log \frac{T(\vec{x}|\vec{y})}{T(\vec{y}|\vec{x})}$, then the move can still be accepted with probability. Hence, unlike steepest descent the state of an MCMC will not get stuck in a local minima because it can always increase the energy (with probability). However, MCMC will not converge to a fixed point but instead to a probability distribution.

The Gibbs sampler is another MCMC (often very simple to implement). This is usually considered to be slower than Metropolis-Hastings (with good proposal distribution) but is easy to implement. It has transition kernels

$$K_r(\vec{x}|\vec{y}) = P(x_r|y_{N(r)})\delta_{\vec{x}/r,\vec{y}/r}, \quad K(\vec{x}|\vec{y}) = \sum_r \rho(r) K_r(\vec{x}|\vec{y}), \tag{11}$$

where x_r denotes the states of a subset r of nodes, $\vec{x}/r$ is the state of all the nodes except r, $x_{N(r)}$ is the state of all nodes that are neighbors of r, and $P(x_r|y_{N(r)})$ is the distribution of x_r conditioned on its neighbors. $\rho(r)$ is a distribution. In words, we select a subset r of nodes with probability $\rho(\vec{y})$ and update their states by sampling from $P(x_r|y_{N(r)})$ keeping the other states fixed. It can be checked that the Gibbs transition kernel satisfies detailed balance.

Here is a simple illustration of Gibbs sampling. Consider the Ising model defined on $\{x_i\}$ with $x_i \in \{-1, +1\}$.

$$P(x_1, ..., x_d) = \frac{1}{Z} \exp\{\mu \sum_{i=1}^{d-1} x_i x_{i+1}\}. \tag{12}$$

The graphical structure has nearest neighbors – i.e. site i is connected to sites $i+1$ and $i-1$ so $N(i) = \{i-1, i+1\}$ (except for $N(1) = \{2\}$ and $N(d) = \{d-1\}$). We let r correspond to nodes i. Then:

$$P(x_i|\vec{x}/i) = P(x_i|x_{N(i)}) = P(x_i|x_{i+1}, x_{i-1}). \tag{13}$$

To determine this, we write $P(x_i|\vec{x}/i) = P(\vec{x})/P(\vec{x}/i)$. We know $P(\vec{x})$ and $P(\vec{x}/i) = \sum_{x_i} P(\vec{x}) = F(\vec{x}/i)$ where $F(.)$ is some function which we can calculate – but this is not the most direct way. It is better to observe that $P(x_i|\vec{x}/i)$ is a function of x_i and $\vec{x}/i$ divided by a function of $\vec{x}/i$ and must be normalized (i.e., $\sum_{x_i} P(x_i|\vec{x}/i) = 1$). Hence $P(x_i|\vec{x}/i) = \exp\{\mu(x_{i-1}x_i + x_i x_{i+1})\}/f(x_{i-1}, x_{i+1})$ where, by normalization, we have $f(x_{i-1}, x_{i+1}) = \exp\{\mu(x_{i-1}+x_{i+1})\}+\exp\{-\mu(x_{i-1}+x_{i+1})\}$. Hence the conditional distributions are:

$$P(x_i|x_{N(i)}) = \frac{\exp\{\mu(x_{i-1}x_i + x_i x_{i+1})\}}{\exp\{\mu(x_{i-1} + x_{i+1})\} + \exp\{-\mu(x_{i-1} + x_{i+1})\}}. \tag{14}$$

The moral is that the conditional distribution $P(x_i|x_{N(i)})$ is usually straightforward to compute for MRF models. Similarly, we get $P(x_1|x_{N(1)}) = \frac{\exp\{\mu(x_1 x_2)\}}{\exp\{\mu x_2\}+\exp\{-\mu x_2\}}$ and $P(x_d|x_{N(d)}) = \frac{\exp\{\mu(x_{d-1}x_d)\}}{\exp\{\mu x_{d-1}\}+\exp\{-\mu x_{d-1}\}}$.

We now define a Gibbs sampler by selecting a site $i\{1, ..., d\}$ from a uniform distribution $U(.)$ (s.t. $U(i) = 1/d, \ \forall i$). Then we sample from $P(x_i|x_{N(i)})$ to generate a new value for x_i (tossing a biased coin). Then we sample another site and continue.

7.2 Theory of MCMC for detailed balance

It is straightfoward to obtain converge results for MCMC (with detailed balance) but unfortunately they depend on properties of the transition kernel which are often hard or impossible to commute (some very clever people – Diaconis, Strook – have obtained bounds for convergence of MCMC but only with difficulty and – like most bounds – they are only of limited use).

To study MCMC with detailed balance the key observation is that the quantity $Q(x,y) = P(y)^{1/2}K(x|y)P(x)^{-1/2}$ is a symmetric matrix. This enables us to apply linear algebra. In particular, $Q(x,y)$ has d real eigenvectors $\{e^\mu(x)\}$ and eigenvalues $\{\lambda^\mu\}$ (where d is the dimension of the state x and the eigenvalues are ordered by their magnitude), and hence can be expressed as $Q(x,y) = \sum_{\mu=1}^n \lambda^\mu e^\mu(x)e^\nu(x)$. It can be shown that $\lambda^1 = 1$ (corresponding to the fixed point conditions $\sum_y K(x|y)P(y) = P(x)$) and that $|\lambda^i| < 1,\ i = 2, ..., d$. It follows that

$$K^M(y|x)P_0(x) = P(x) + \sum_{\mu=2}^{d} \alpha_\mu \{\lambda^\mu\}^M e^\mu(x)P(x)^{1/2}, \tag{15}$$

where $K^M(y|x)$ is matrix multiplication of the transition kernel with itself M times, $P_0(x)$ is the initial distribution and $\alpha_\mu = \sum_y P_0(y)e^\mu(y)P(y)^{-1/2},\ \ \mu = 2, ..., d$.

The main result is that the second term on the RHS of the equation decay exponentially fast (in M) with decay speed determined by the magnitude of the second biggest eigenvalue λ^2. This implies that samples from the MCMC converge to samples from $P(x)$ exponentially rapidly. The only problem is that computing λ^2 is often impossible.