Variational Inference

Sargur Srihari
srihari@cedar.buffalo.edu

Topics

- Inference as optimization
  1. Exact Inference revisited
  2. The Energy Functional
  3. Optimizing the Energy Functional

Need for Approximate Inference

- For many networks inference can be performed efficiently
  - But time/space complexity of the clique tree is exponential in tree width of network
  - In such cases exact algorithms become infeasible
- This motivates examination of approximate inference methods
  - Where approximation arises from constructing an approximation to target distribution P_Φ
  - The approximation takes a simpler form that allows inference

Class of Approximate Algorithms

- We consider here a class of approximate inference methods that share a common principle
- Find target class Q of "easy" distributions and
  - Search for an instance within that class that best approximates P_Φ
  - Queries are then answered using inference on Q rather than P_Φ
  - Methods optimize a target function for measuring similarity between Q and P_Φ

Reformulation of Inference Problem

- Inference problem is one of optimizing an objective function over the class Q
- Problem is one of constrained optimization
  - Technique used is based on Lagrange multipliers
- Produces a set of equations that characterize the optima of the objective
  - A set of fixed-point equations that define each variable in terms of others
  - Fixed point equations derived from constrained energy optimization can be viewed as passing messages over a graph object

Categories of methods in this class

1. Message passing on Clique Tree
   - Loopy belief propagation
     - Optimize approximate versions of the energy functional
2. Message passing on Clique Trees with approximate messages
   - Called expectation propagation
     - Maximize exact energy functional but with relaxed constraints on Q
3. Mean-field method
   - Originates in statistical physics
     - Focus on Q that has simple factorization

Exact Inference Revisited

- We have a factorized distribution of the form

$$P_{\Phi}(X) = \frac{1}{Z} \prod_{\phi \in \Phi} \phi(U_{\phi})$$

  - where $U_{\phi} = Scope(\phi)$
  - Factors are:
    - CPDs in a BN or
    - potentials in a MN
- We are interested in answering queries:
  - about marginal probabilities of variables and
  - about the partition function

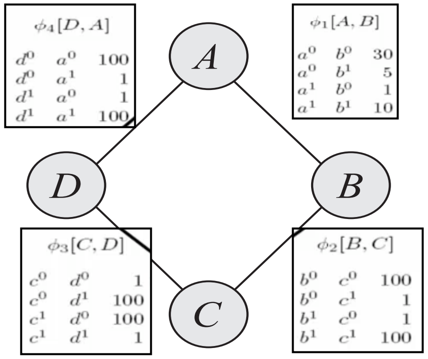
Cluster Tree Representation

- End-product of Belief Propagation is a calibrated cluster tree
- A calibrated set of beliefs represents a distribution
- We view exact inference as searching over the set of distributions Q that are representable by the cluster tree to find a distribution Q^* that matches P_Φ

Cluster graph U for factors Φ over χ is an undirected graph
Each of whose nodes i is associated with a subset $C_i \subseteq \chi$
Each edge between pair of clusters C_i and C_j is associated with a sepset $S_{i,j} \subseteq C_i \cap C_j$
A tree T is a clique tree for graph H if

Each node in T corresponds to a clique in H and each maximal clique in H is a node in T
Each sepset $S_{i,j}$ separates $W_{<Ij,j)}$ and $W<_{(j,i)}$ in H

Ex: Clique Tree representation



1.Gibbs Distribution

$$\tilde{P}_\Phi(A,B,C,D) = \phi_1(A,B)\phi_2(B,C)\phi_3(C,D)\phi_4(D,A)$$

$$P(A,B,C,D) = \frac{1}{Z}\phi_1(A,B)\cdot\phi_2(B,C)\cdot\phi_3(C,D)\cdot\phi_4(D,A)$$

where

$$Z = \sum_{A,B,C,D}\phi_1(A,B)\cdot\phi_2(B,C)\cdot\phi_3(C,D)\cdot\phi_4(D,A)$$

Z=7,201,840

Assignment				*Unnormalized*
a^0	b^0	c^0	d^0	300000
a^0	b^0	c^0	d^1	300000
a^0	b^0	c^1	d^0	300000
a^0	b^0	c^1	d^1	30
a^0	b^1	c^0	d^0	500
a^0	b^1	c^0	d^1	500
a^0	b^1	c^1	d^0	5000000
a^0	b^1	c^1	d^1	500
a^1	b^0	c^0	d^0	100
a^1	b^0	c^0	d^1	1000000
a^1	b^0	c^1	d^0	100
a^1	b^0	c^1	d^1	100
a^1	b^1	c^0	d^0	10
a^1	b^1	c^0	d^1	100000
a^1	b^1	c^1	d^0	100000
a^1	b^1	c^1	d^1	100000

2. Clique Tree (triangulated):

1. A,B,D —*{B,D}*— *2. B,C,D*

Initial Potentials:

$$\psi_1(A,B,D) = \phi_1(A,B)\phi_2(B,C)\phi_3(C,D)\phi_4(D,A)$$
$$\psi_2(B,C,D) = \phi_1(A,B)\phi_2(B,C)\phi_3(C,D)\phi_4(D,A)$$

Beliefs (Clique and Sepset)

$$\beta_1(A,B,D) = \tilde{P}_\Phi(A,B,D) = \sum_C \psi_1(A,B,D) = \sum_C \phi_1(A,B)\phi_2(B,C)\phi_3(C,D)\phi_4(D,A)$$

e.g., $\beta_1(a^0,b^0,d^0) = 300,000 + 300,000 = 600,000$

$$\mu_{1,2}(B,D) = \sum_{C_1 - S_{1,2}} \beta_1(C_1) = \sum_A \beta_1(A,B,D)$$

e.g., $\mu_{1,2}(b^0,d^0) = 600,000 + 200 = 600,200$

$$\beta_2(B,C,D) = \tilde{P}_\Phi(B,C,D) = \sum_A \mu_{1,2}(B,D)\cdot\psi_2(B,C,D) = \sum_A \psi_2(B,C,D)$$

e.g., $\beta_2(b^0,c^0,d^0) = 300,000 + 100 = 300,100$

Assignment			$\max_C$
a^0	b^0	d^0	600,000
a^0	b^0	d^1	300,030
a^0	b^1	d^0	5,000,500
a^0	b^1	d^1	1,000
a^1	b^0	d^0	200
a^1	b^0	d^1	1,000,100
a^1	b^1	d^0	100,010
a^1	b^1	d^1	200,000

$\beta_1(A,B,D)$

Assignment		$\max_{A,C}$
b^0	d^0	600,200
b^0	d^1	1,300,130
b^1	d^0	5,100,510
b^1	d^1	201,000

$\mu_{1,2}(B,D)$

Assignment			[illegible]
b^0	c^0	d^0	300,[illegible]
b^0	c^0	d^1	1,300,[illegible]
b^0	c^1	d^0	300,[illegible]
b^0	c^1	d^1	[illegible]
b^1	c^0	d^0	[illegible]
b^1	c^0	d^1	100,[illegible]
b^1	c^1	d^0	5,100,[illegible]
b^1	c^1	d^1	100,[illegible]

$\beta_2(B,C,D)$

$$\tilde{P}_\Phi(a^1,b^0,c^1,d^0) = 100$$

$$\frac{\beta_1(a^1,b^0,d^0)\beta_2(b^0,c^1,d^0)}{\mu_{1,2}(b^0,d^0)} = \frac{200\cdot 300\cdot 100}{600\cdot 200} = 100$$

Distance between Q and P_Φ

- We need methods to optimize distance between Q and P_Φ without answering hard queries about P_Φ
  - The relative entropy (or K-L divergence) allows us to exploit the structure of P_Φ without performing reasoning with it
  - We use Relative entropy of P_1 and P_2 defined as

$$\boxed{D\left(P_1 \,||\, P_2\right) = E_{P_1}\left[\frac{\ln P_1\left[\chi\right]}{\ln P_2\left[\chi\right]}\right]}$$

    - It is always non-negative
    - Equal to 0 if and only if $P_1 = P_2$
- We search for distribution Q that *minimizes* $D(Q \,||\, P_\Phi)$

Summarize the discussion

- Want a distribution Q that minimizes $D(Q \,||\, P_\Phi)$
- To define problem formally, we need to specify objects we want to optimize over
  - Suppose we are given:
    - a clique tree structure T for $P_{\Phi,}$ a set of beliefs

$$Q=\{\beta_i: i \,\varepsilon\, V_T\} \cup \{\mu_{i,j}: (i\text{-}j) \,\varepsilon\, E_T\}$$

where C_i are clusters in T, β_i denote beliefs over C_i and $\mu_{i,j}$ denotes beliefs $S_{i,j}$ of edges in T

  - Set of beliefs in T defines a distribution Q by
  - The beliefs correspond to marginals of Q

$$\boxed{Q(\chi) = \frac{\prod_{i \in V_T} \beta_i}{\prod_{(i-j) \in V_T} \mu_{i,j}}}$$

  - We are now searching over a set of distributions Q
  - that are representable by a set of beliefs Q over the cliques and sepsets in a particular clique tree structure Q

Statement of Inference as Optimization

- Exact inference is one of maximizing $-D(Q \,||\, P_{\Phi})$ over the space of calibrated sets Q

Ctree-Optimize-KL

- **Find** $Q=\{\beta_i : i \in V_T\} \cup \{\mu_{i,j} : (i\text{-}j) \in E_T\}$
- **Maximizing** $-D(Q \,||\, P_{\Phi})$
- **Subject to**

$$\boxed{\begin{array}{l} \mu_{i,j}\left[s_{i,j}\right] = \sum_{C_i - S_{i,j}} \beta_i\left(c_i\right) \quad \forall\left(i-j\right) \in E_T, \forall s_{i,j} \in Val\left(S_{i,j}\right) \\ \sum_{c_i} \beta_i\left(c_i\right) = 1 \qquad \forall \mathrm{i} \in \mathrm{V}_T \end{array}}$$

- Theorem: If T is an I-map of P_{Φ} then there is a unique solution to Ctree-Optimize-KL

Possible approach

- Examine different configurations of beliefs that satisfy marginal consistency constraints
  - Select the configuration that maximizes the objective
  - Such as exhaustive examination is impossible to perform
- Instead of searching over a space of all calibrated trees we can search over a space of simpler distributions
  - We will not find a distribution equivalent to P_Φ but one that is reasonably close

Defining the Energy Functional

- However directly evaluating $D(Q \| P_\Phi)$ is unwieldy

$$D(P_1 \| P_2) = E_{P_1}\left[\frac{\ln P_1[\chi]}{\ln P_2[\chi]}\right] = \sum_{\chi} P_1[\chi]\left[\frac{\ln P_1[\chi]}{\ln P_2[\chi]}\right]$$

  - Because summation over all χ is infeasible in practice
- Instead use equivalent form $D(Q \| P_\Phi) = \ln Z - F(\tilde{P}_\Phi, Q)$
  - Where F is the energy functional

$$F[\tilde{P}_\Phi, Q] = E_Q[\ln \tilde{P}(\chi)] + H_Q(\chi) = \sum_{\phi \in \Phi} E_Q[\ln \phi] + H_Q(\chi)$$

- Since the term $\ln Z$ does not depend on Q,
  - minimizing relative entropy $D(Q \| P_\Phi)$ is equivalent to maximizing the energy functional $F(\tilde{P}_\Phi, Q)$
- Energy functional $F[\tilde{P}_\Phi, Q] = \sum_{\phi \in \Phi} E_Q[\ln \phi] + H_Q(\chi)$ has two terms:
  - energy term (expectation of logs of factors in Φ) and entropy term

Optimizing the Energy Functional

- From here onward we pose the problem of finding a good Q as one of maximizing the energy functional
  - Equivalently minimizing the relative entropy
  - Importantly energy functional involves expectations in Q
  - By choosing Q that allow efficient inference we can evaluate/ optimize the energy functional
- Moreover, energy Functional is a lower bound on partition function
  - Since $D(Q||P_\Phi) \geq 0$ we have $\ln Z \geq F\left[\tilde{P}_\Phi, Q\right]$
  - Useful since partition function is usually the hardest part of inference
    - Plays important role in learning

Strategies for optimizing energy functional

- Methods are referred to as Variational Methods
- Refers to a strategy in which we introduce new parameters that increase the degrees of freedom
- Each choice of these parameters gives a different approximation
- We attempt to optimize the variational parameters to get the best approximation
- Variational calculus: finding optima of a functional
  - E.g., distribution that maximizes entropy

Further Topics in Variational Methods

- Exact Inference
- Propagation-Based Approximations
- Propagation with Approximate Messages
- Structured Variational Approximations