Lecture 14: March 11, 2015
cs 573: Probabilistic Reasoning
Professor Nevatia
Spring 2015

Admin

- HW #5 due Mar 30; to be posted by March 12
- Exam 2, April 29, class period; NOT cumulative
- Exam 1
  - Median 21
  - High 26.5
  - Upper ¼ $\geq$ 22.5
  - Lower ¼ $\leq$ 17
- Please return exams on Monday, Mar 23
- For regrading, submit comments in writing

Review

- MAP Inference
  - Max query, Max-Marginal Query
  - Methods similar to marginal queries
    - Replace SUM operations with MAX
    - Traceback required
- Today's objective
  - Variational Approach to Approximation

Approximate Inference

- Three approaches to approximation
- Loopy Belief Propagation
  - Simple, commonly used
  - Convergence or quality of approximation is hard to predict
- Variational Methods
  - Underlying math is complex
  - Convergence is guaranteed but not the quality of the approximation
  - We will cover only lightly in the course
- Sampling Methods
  - Will study in some detail (ch. 12)

Variational Approach

- Treats inference problem as an optimization problem
- Approximate the actual distribution, say P, with a simpler distribution, say Q
  - e.g. fit a 1-D Gaussian to a 1-D arbitrary distribution
  - Fit a multi-variate Gaussian but assume co-variances are nil
  - Assume Q is a product of individual variable distributions
  - Need to select a tractable Q
- How to measure difference between P and Q?
- How to choose parameters of Q to minimize the difference between P and Q?

Distance between two Distributions

- How to measure the distance between two probability distributions
- Consider the discrete case for two distributions P and Q, over
  - The L_1 distance: $\|P - Q\|_1 = \sum_{x_1,\ldots,x_n} |P(x_1,\ldots,x_n) - Q(x_1,\ldots,x_n)|$.
  - The L_2 distance: $\|P - Q\|_2 = \left(\sum_{x_1,\ldots,x_n} (P(x_1,\ldots,x_n) - Q(x_1,\ldots,x_n))^2\right)^{\frac{1}{2}}$.
  - The L_∞ distance: $\|P - Q\|_\infty = \max_{x_1,\ldots,x_n} |P(x_1,\ldots,x_n) - Q(x_1,\ldots,x_n)|$.

- L_2 norm is sensitive to a single, large difference; other two are not differentiable and hence harder to work with. Also, they don't factorize well even if the probability distribution does, so differences based on entropy are used more commonly.

Entropy of a Distribution

- See Appendix A.1 of the KF book
- Notion of entropy derives from coding/information theory
  - How many bits needed to transmit the data in an optimal code
- Definition: **Entropy** of X given distribution P(x)
- $H_P(X) = E_P [\log 1/P(x)] = \sum_x P(x) \log (1/P(x))$

  $= -\sum_x P(x) \log P(x)$
- Consider a fair coin, H_p will be .5log .5 + .5 log .5 = 1 (log base 2)
- What if coin always come up heads: entropy = 0 (no need to transmit the result)
- If the coin is unfair, P(heads) = .9; entropy will be low and fewer bits needed to transmit results (not for one trial but many trials, constructing the optimal code is beyond the scope of this course)
- Another interpretation is how much information do we get from the result, or how much uncertainty is introduced by a distribution
  - Consider uniform vs highly peaked or bi-modal

Joint and Relative Entropy

- Joint Entropy

  $H_P(X_1, X_2 \ldots X_n) = E_P[1/\log P(X_1, X_2 \ldots X_n)]$

  Can relate to how many bits needed to encode joint values.

- Relative entropy for multiple variables
  - $D(P||Q) = E_P[\log \{P(X_1, X_2 \ldots X_n) / Q(X_1, X_2 \ldots X_n)\}]$

    $= E_P[\log P(X_1, X_2 \ldots X_n)] - E_P[\log Q(X_1, X_2 \ldots X_n)]$

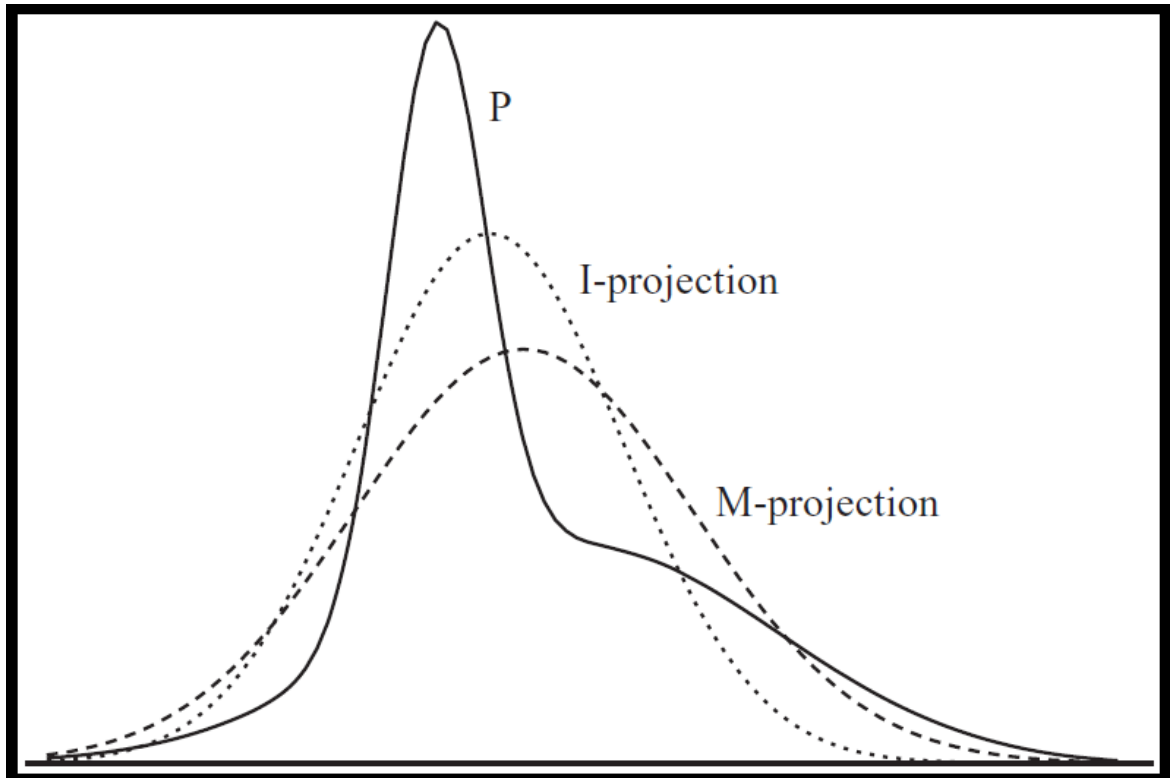
    $= - H_P(X) - E_P[\log Q(X)] = - H_P(X) - \sum_x P(x) \log Q(x)$
  - Extra "cost" imposed by using Q instead of P
  - Kullback-Liebler Divergence (KL-divergence or just KLD)
- $D(P||Q) \geq 0$, it is = 0 *iff* P= Q
  - Proof requires use of Gibbs inequality

    $- \sum_x P(x) \log P(x) \leq - \sum_x P(x) \log Q(x)$
  - See http://en.wikipedia.org/wiki/Gibbs%27_inequality

Joint and Relative Entropy

- $D(Q||P) = - H_Q(X) - \sum_x Q(x) \log P(x)$
- Note $D(P||Q) \neq D(Q||P)$, unless P = Q
- KLD often used for measuring similarity between P and Q.
  - NOT a distance metric; is not symmetric and does not satisfy triangle inequality
- Often, we want to approximate P with a simpler function Q; select the optimal Q from a set of distributions *Q*.
- I and M projections
  - $Q^I = \arg\min_{Q \varepsilon Q} D(Q||P)$; information projection
  - $Q^M \arg\min_{Q \varepsilon Q} D(P||Q)$; moment projection
  - Which is better?
    - Q^M tends to have a higher variance for Gaussian approximations (see example, Figure 8.1)
  - We are going to use Q^I for the variational approach
    - Partly because easier to compute expectations w.r.t. Q

I and M Projections



Example: approx with a Gaussian

Energy Functional

- Minimize $D(Q||P_\Phi) = - H_Q(X) - E_Q[\ln P_\Phi(X)]$

  Where: $P_\Phi(X) = (1/z) \Pi_{\phi \varepsilon \Phi} \phi(U_\phi)$;

  $\{U_\phi$: set of variables in scope of $\phi\}$

  $\ln P_\Phi(X) = \sum_{\phi \varepsilon \Phi} \ln \phi(U_\phi) - \ln Z$

$$D(Q||P_\Phi) = - H_Q(X) - E_Q[\sum_{\phi \varepsilon \Phi} \ln \phi(U_\phi)] + E_Q[\ln Z]$$
$$= - H_Q(X) - \sum_{\phi \varepsilon \Phi} E_Q[\ln \phi(U_\phi)] + \ln Z \text{ (Z is indep of Q)}$$
$$= - F[{}^{\sim}P_\Phi, Q] + \ln Z \text{ \{by definition of F\}}$$

$$F[{}^{\sim}P_\Phi, Q] = H_Q(X) + \sum_{\phi \varepsilon \Phi} E_Q[\ln \phi(U_\phi)] \text{ (by definition)}$$

- Find Q that minimizes $D(Q||P_\Phi)$, maximizes $F[{}^{\sim}P_\Phi, Q]$, called *free energy* in physics

Minimize Energy Functional

- Note that F is a *functional* (rather than a function), its arguments are functions and search for optimal is over space of functions
- If $P_\Phi(X)$ is defined over a tree, SP message passing algorithm can be shown to maximize this functional (recovered Q=P; D = 0)
- Loopy BP also maximizes a factored energy functional but this functional is only an approximation of the actual functional.
- Need to find an optimal function Q
  - Not tractable without imposing restrictions on form of Q
- Assume that Q can be factored according to product of variable probabilities, *i.e.* a graph without any edges in it
- Then, $Q(\boldsymbol{X}) = \prod_i Q(X_i)$; also $\sum_i Q(X_i) = 1$
- Minimize energy functional with these constraints
- Note there are no restrictions on $Q(X_i)$

Finding the Minimum

- Energy functional consists of two terms

  $H_Q(X) = \sum_i H_Q(X_i)$;

  $\sum_{\phi \varepsilon \Phi} E_Q[\ln \phi(U_\phi)] = \sum_{\phi \varepsilon \Phi} \{\Pi_{x_i \varepsilon U_\phi} Q(X_i)\} [\ln \phi(U_\phi)]$
- Minimizing the functional is a constrained optimization problem ($\sum_i Q(X_i) = 1$) so use the Lagrange multiplier technique.
- Take derivatives of the Lagrangian with respect to $Q(x_i)$ and set = 0
- Leads to a "Mean-Field" approximation

$$Q(x_i) = \frac{1}{Z_i} \exp\{E_{X_{-i} \sim Q}[\ln P_\Phi(x_i \mid X_{-i})]\}$$

- Note that the term "Mean Field" name comes about because $Q(X_i)$ is the geometric average of the conditional probability of x_i given all the other variables
- Also called a *fixed point equation*
  - At fixed point, x = f(x) in an iterative computation
- Mean-field equation can be further simplified: see next page

Mean Field Algorithm

- Corollary 11.6: shows that $Q(X_i)$ is locally optimal only if:

$$Q(x_i) = \frac{1}{Z_i} \exp\left\{ \sum_{\phi: X_i \in Scope[\phi]} E_{(U_\phi - \{X_i\}) \sim Q}[\ln \phi(U_\phi, x_i)] \right\}.$$

where Z_i is a normalizing constant.

- Notation of the equation is complex but is actually quite simple to apply
  - It sums computes expected values of an assignment of variables in a factor where X_i appears, using the value of Q distribution.
  - If more than one variable, say X_1 and X_2 appear in one factor, $Q(X_1)$ value may affect $Q(X_2)$: hence an iterative algorithm.

Explaining the Notation

- Start with a clique of just two variables (X_1,X_2), given $\phi(X_1,X_2)$
- For X_1 : $Q(X_1=x_1^i) = 1/z_1 \exp(\sum_{x2} Q(x_2) \log \phi(x_1^i,x_2))$
  - Sum is over all possible values of X_2
  - Note above updates the distribution $Q(X_1)$
  - Similar expression for $Q(X_2)$ in terms of $Q(X_1)$; iterate
- Suppose we have a clique (X_1, X_2, X_3)

  For X_1 : $Q(X_1=x_1^i) = 1/z_1 \exp(\sum_{x2,x3} Q(x_2,x_3) \log \phi(x_1^i,x_2,x_3))$

  $= 1/z_1 \exp(\sum_{x2,x3} Q(x_2)Q(x_3) \log \phi(x_1^i,x_2,x_3))$

- If X_1 is also in another clique, say (X_1, X_4), add terms corresponding to $\sum_{x4} Q(x_4) \log \phi(x_1,x_4)$ to the above summation

Next Class

- Read sections 12.1 and 12.2 of the KF book