Deterministic Approximate Inference

Ron Parr
CPS 271

Deterministic vs. Stochastic

- Deterministic approximations give the same answer every time
- Stochastic approximations are typically based upon sampling:
  - Might give different answers based upon different random number seeds
  - Should converge to the correct answer in the limit

Outline

- Variational approximations
  - Toy application to entropy maximization
  - Example of use in EM
- Expectation propagation

Variational Approximations

- Variational is an overloaded term
- In machine learning/AI, typically refers to:
  - Substitution of one functional form for another
  - Substitution that ensures a one sided bound
- Main idea: Look where the light is!
- If an optimization problem is too hard, replace the problem with an easier one
- Isn't this just cheating?
- Yes, but if you do it in a clever way, you can still provide some guarantees

Maximizing Entropy

- Recall definition of entropy:

$$H(P(X)) = \sum_{X} P(X) \log P(X)$$

- Entropy is a functional (function defined over functions)
- Suppose we have two variables, x and y, and we with to find the joint distribution with highest entropy…

Maximizing Entropy

- Suppose X and Y are binary
- P(XY) is a function from X,Y to [0,1]
- Specified by 3 numbers
- Entropy:

$$H(P(XY)) = \sum_{XY} P(XY) \log P(XY)$$

How to maximize this?

- For simple problems, one can do the maximization directly (set the gradient to 0)
- What if it's hard to do this?
- Idea: Instead of maximizing over all distributions, maximize over just those in which X and Y are independent:

$$P(XY) = P(X)P(Y)$$

Entropy under independence

$$
\begin{aligned}
H(P(XY)) &= \sum_{XY} P(XY)\log P(XY) \\
&= \sum_{XY} P(X)P(Y)\log P(X)P(Y) \\
&= \sum_{XY} P(X)P(Y)(\log P(X) + \log P(Y)) \\
&= \sum_{XY} P(X)P(Y)\log P(X) + \sum_{XY} P(X)P(Y)\log P(Y) \\
&= \sum_{X} P(X)\log P(X)\sum_{Y} P(Y) + \sum_{Y} P(Y)\log P(Y)\sum_{X} P(X) \\
&= \sum_{X} P(X)\log P(X) + \sum_{Y} P(Y)\log P(Y) \\
&= H(P(X)) + H(P(Y))
\end{aligned}
$$

Maximizing Entropy under Independence

$$H(P(XY)) = H(P(X)) + H(P(Y))$$

$$\max_{P(XY)} H(P(XY)) = \max_{P(X)} H(P(X)) + \max_{P(Y)} H(P(Y))$$

- Under assumption of independence:
  - Maximizing entropy for joint distribution decomposes into maximizing entropy for individual distributions
  - H maximized by uniform distribution over X and Y
- This also turns out to be the true maximum, but
- This isn't always guaranteed!

Variational Approximation: Discussion

- Substituting P(X)P(Y) for P(XY) was "safe"
- Could never overestimate true max entropy
- Why:
  - The set of distributions where X and Y are independent is a subset of the set of joint distributions
  - Reinterpretation of independence assumption: We aren't computing the wrong probabilities; we're merely searching a smaller space

EM

- Recall that EM seeks typically seeks to maximize the joint likelihood of the data (X) and parameters θ given some hidden variables (Z)
- Alternates between:
  - Estimating: $Q^{k+1} = p(Z \mid D, \theta^k)$ with w fixed
  - Maximizing: $\theta^{k+1} = \arg\max_{\theta} \sum_{Z} p(Z \mid D, \theta^k) \log p(D, Z \mid \theta^k)$
- Idea: Alternate between estimating hidden parameters, and finding "best fit" model to these parameters
- Example: Gaussian mixtures
  - E step: Estimate membership in clusters
  - M step: Update clusters

A Slightly Different View of EM

Lumping together Z and θ , let's maximize p(x):

$$\log p(X) = L(q) + KL(q \parallel p)$$

$$L(q) = \int q(Z)\log(\frac{p(X,Z)}{q(Z)})dZ$$

$$KL(q \parallel p) = -\int q(Z)\log(\frac{p(Z \mid X)}{q(Z)})dz$$

Really???

Sanity check…

$$
\begin{aligned}
\log p(X) &= L(q) + KL(q \,\|\, p) \\
&= \int q(Z)\log(\frac{p(X,Z)}{q(Z)})dZ - \int q(Z)\log(\frac{p(Z \mid X)}{q(Z)})dZ \\
&= \int q(Z)\log(\frac{p(X,Z)}{q(Z)}) - \log(\frac{p(Z \mid X)}{q(Z)})dZ \\
&= \int q(Z)\log(\frac{p(X,Z)}{q(Z)}\frac{q(Z)}{p(Z \mid X)})dZ \\
&= \int q(Z)\log(\frac{p(X,Z)}{p(Z \mid X)})dZ \\
&= \int q(Z)\log(p(X))dZ \\
&= \log(p(X))\int q(Z)dZ \\
&= \log(p(X))
\end{aligned}
$$

EM Continued

$$\log p(X) = L(q) + KL(q \,\|\, p)$$

$$L(q) = \int q(Z)\log(\frac{p(X,Z)}{q(Z)})dZ$$

$$KL(q \,\|\, p) = -\int q(Z)\log(\frac{p(X \mid Z)}{q(Z)})dz$$

- EM reinterpreted:
  - E: Maximizie L(q) keeping Z fixed
  - M: Maximize fit between q and p, keeping p fixed
- In terms of Gaussian mixtures:
  - E: Assign points cluster membership probs
  - M: Update cluster centers, variances based on membership

Variation Approximation for EM

- Variational approximation comes into play when the it is hard to maximize the KL distance for a particular form of Q
- Example: Clustering with priors
  - Assume priors over:
    - Cluster membership
    - Cluster means
    - Cluster variances
  - Problem: Q now has an ugly form

Variational Mixture of Gaussians

- Form of distribution with priors:

$$p(X, Z, \pi, \mu, \Lambda) = p(X \mid Z, \mu, \Lambda)\, p(Z \mid \pi)\, p(\pi)\, p(\mu \mid \Lambda)\, p(\Lambda)$$

Dirichlet mixture membership prior:

Gaussian-Wishart priors on mean and precision:

- But how do we minimize KL for: $q(Z, \pi, \mu, \Lambda)$

Variational Approximation Step

- Approximate: $q(Z, \pi, \mu, \Lambda)$
- With: $q(Z)q(\pi, \mu, \Lambda)$
- Do coordinate ascent on these separately:
- Alternate between:
  - Freezing priors and updating Gaussians
  - Freezing Gaussians and updating priors
- Why this is good:
  - Can show that each step is tractable
  - Works well in practice (can be viewed as "solving" the problem of how many clusters are needed)

Summary of Variational Approach for EM

- Replace intractable maximization of Q w/something simpler
- Usually this plays out as follows:
  - Make some independence assumptions that let you factor Q
  - Perform coordinate ascent on the factored version of Q by freezing some terms while optimizing others
- Why this is safe:
  - Factored representations are a subset of the space of original distributions
  - We will never overestimate, but we might fail to find the globally optimal choice
- In general: Assuming independence is not a requirement; it's just a convenient choice

Variational Approximation vs. Independence Assumptions

- Q: Aren't these just the same?
- A: They overlap, but they aren't identical
- Variational approximation:
  - Often involves an independence assumption because doesn't require it
  - Often occurs in the inner loop of an optimization, driven by efficiency of optimization concerns
  - Often applied to latent variables
- Independence assumptions:
  - Usually a high level modeling decision about the observed and latent variables
  - Driven by representation concerns

Expectation Propagation

- EP: Deterministic approximation method
- Less general than variational methods
- Quick and easy to understand an implement
- Main assumption: Distribution is represented as a product of factors:

$$P(D,\theta) = \prod_i f_i(\theta)$$

- Example: A graphical model

EP, continued

- We want to approximate the posterior distribution of the model parameters, given the data:

$$f(\theta) = \frac{1}{Z}\prod_i \tilde{f}_i(\theta)$$

- Idea: Do some kind of coordinate ascent by freezing all factors except one, and then updating the free parameters

EP, continued

Initialize : $\tilde{f}$

Initialize : $q(\theta) \propto \prod_i \tilde{f}_i(\theta)$

Repeat until convergence :

Choose some : $\tilde{f}_j(\theta)$

compute : $q^{\setminus j}(\theta) = \dfrac{q(\theta)}{\tilde{f}_j(\theta)}$

fit $q^{new}(\theta)$ to $q^{\setminus j}(\theta) f_j(\theta)$

$z_j = \int q^{\setminus j}(\theta) f_j(\theta) d\theta$

$\tilde{f}_j(\theta) = z_j \dfrac{q^{new}(\theta)}{q^{\setminus j}(\theta)}$

What's going on here, in English:
Freeze all components of our approximation, except one
Update our approximation locally
Repeat

EP Properties

- Is exact in some special cases
- Can be shown to be equivalent to some message passing algorithms for graphical models

Approximate Inference Conclusions

- Deterministic approximations rely upon some form of simplifying assumption about the model
- Often represent messy distributions with products of simpler distributions (factorization)
- Often replace global optimizations with local optimizations
- Main advantage: Stable, predictable
- Main disadvantage: No "anytime" property