Algoritmo de propagação de expectativa

Shuang Wang
Escola de Engenharia Eletrotécnica e de Computadores
Universidade de Oklahoma, Tulsa, OK, 74135
E-mail: {shuangwang}@ou.edu

Esta nota contém três partes. Primeiro, revisaremos algumas preliminares para o EP. Em seguida, o algoritmo EP será descrito na próxima seção. Finalmente, a relação entre PE e outros métodos variacionais será discutida.

I. PRELIMINARES

A. Família exponencial

A família exponencial de distribuições sobre x é um conjunto de distribuições com a forma

$$p(x; \theta) = h(x)g(\theta)\exp(\theta^T u(x)), \quad (1)$$

onde a medição x pode ser escalar ou vetorial, discreta ou contínua, θ são parâmetros da distribuição, h(x)e u(x) são algumas funções de x, e a função g(θ) é um fator de normalização como

$$g(i) \int h(x)\exp(\theta^T u(x))\, dx = 1. \quad (2)$$

Além disso, se as variáveis forem discretas, basta substituir a integração pela soma.

A família exponencial tem muitas propriedades, o que pode simplificar os cálculos. Por exemplo, se uma função de verossimilhança é um dos membros da família exponencial, a posterior pode ser expressa em uma expressão de forma fechada escolhendo um conjugado anterior dentro da família exponencial. Além disso, a família exponencial tem uma ampla gama de membros, como Gaussiano, Bernoulli, multinomial discreto, Poisson e assim por diante, portanto, é aplicável a muitos modelos de inferência diferentes.

B. Divergência de Kullback-Leibler

A divergência de Kullback-Leibler (KL) [1] é uma medida para quantificar a diferença entre uma distribuição probabilísticasp(x) e uma distribuição aproximada q(x). Para as distribuições p(x) e q(x) sobre variáveis contínuas, KLdivergência é definida como

$$DKL(p(x)\|q(x)) = \int p(x)\log \frac{p(x)}{p(x)q(x)}\, dx, \quad (3)$$

onde, para variáveis discretas, basta substituir a integração pela soma. Além disso, a divergência KL é uma medida não simétrica, o que significa DKL(p(x)‖q(x)) 6= DKL(q(x)‖p(x)). Para dar aos leitores uma visão intuitiva sobre a diferença entre as duas formas de divergência KL acima, assumimos que a verdadeira distribuição p(x) é multimodal e a distribuição candidata q(x) é unimodal. Ao minimizar DKL(q(x)‖p(x)), a distribuição aproximada q(x)

escolha um dos modos em p(x), que geralmente é usado no método variacional de Bayes. No entanto, a melhor distribuição aproximada q(x) obtida minimizando DKL(p(x)‖q(x)) será a média de todos os modos. O último caso é usado no procedimento de inferência aproximada de EP. Como este relatório se concentra na revisão do algoritmo EP, estudaremos a propriedade de minimizar DKL (p (x) ‖q (x)) primeiro. Em relação à diferença entre minimizar DKL(p(x)‖q(x))e DKL(q(x)‖p(x)), discutiremos isso mais adiante neste capítulo.

Para garantir uma solução tratável para minimizar a divergência KL DKL(p(x)‖q(x)), a distribuição aproximada q(x)é geralmente restrita dentro de um membro da família exponencial. Assim, de acordo com (1), q(x) pode ser escrito como

$$q(x;\theta) = h(x)g(\theta)\exp(\theta^T u(x)), \tag{4}$$

onde θ são os parâmetros da distribuição dada.

Substituindo q(x; θ) na divergência KL DKL(p(x)‖q(x)), obtemos

$$DKL(p(x)\|q(x)) = -\ln(\theta) - \theta^T E_{p(x)}[u(x)] + \text{const}, \tag{5}$$

onde o const representa todos os termos que são independentes dos parâmetros θ. Para minimizar a divergência KL, pegue o gradiente de DKL(p(x)‖q(x)) em relação a θ a zero, obtemos

$$-\nabla \ln(\theta) = E_{p(x)}[u(x)]. \tag{6}$$

Além disso, para (2), tomando o gradiente de ambos os lados em relação a θ, obtemos

$$\nabla g(\theta) \int h(x)\exp\{\theta^T u(x)\}\,dx + g(\theta) \int h(x)\exp\{\theta^T u(x)\}\,u(x)dx = 0. \tag{7}$$

Então, reorganizando e reutilizando (2) novamente, obtemos

$$-\nabla \ln(\theta) = E_{q(x)}[u(x)]. \tag{8}$$

Comparando (6) e (8), obtemos

$$E_{p(x)}[u(x)] = E_{q(x)}[u(x)]. \tag{9}$$

Assim, a partir de (9), vemos que a minimização da divergência KL é equivalente a corresponder às estatísticas suficientes esperadas. Por exemplo, para minimizar a divergência KL com uma distribuição gaussiana q(x; θ), precisamos apenas encontrar a média e a covariância de q(x; θ) que são iguais à média e à covariância de p(x; θ), respectivamente.

C. Filtragem de densidade presumida (ADF)

O ADF é uma técnica para construir aproximação tratável para distribuição de probabilidade complexa. O EP pode ser visto como uma extensão do ADF. Assim, primeiro fornecemos uma revisão concisa do ADF e, em seguida, a estendemos ao algoritmo EP.

Consideremos a regra de Bayes e suponhamos que a fatoração da distribuição posterior tenha a seguinte forma

$$p(x|y) = \frac{p(x)p(y|x)}{p(y)}$$
$$= \frac{1}{Z} p0(x) \prod_{i=1}^{N} p(yi|x), \quad (10)$$
$$= \frac{1}{Z} \prod_{i=0}^{N} pi(x),$$

onde Z é uma constante de normalização, pi(x) é uma notação simplificada de cada fator correspondente em (10), onde p0(x) = p0(x) e pi(x) = pi(yi|x) para i > 0. Se assumirmos que a função de verossimilhança p(yi|x) tem uma forma complexa, a avaliação direta da distribuição a posteriori seria inviável. Por exemplo, se cada função de verossimilhança for uma mistura de duas distribuições gaussianas e houver um número total de N = 100 de dados observados. Então, para obter a distribuição posterior, precisamos avaliar a mistura de 2100 gaussianos.

Para resolver esse problema, os métodos de inferência aproximada tentam buscar uma distribuição posterior aproximada que possa ser muito próxima da verdadeira distribuição posterior p(x|y). Normalmente, as distribuições aproximadas são escolhidas dentro da família exponencial para garantir a viabilidade computacional. Então, a melhor distribuição aproximada pode ser encontrada minimizando a divergência KL:

$$\theta* = \arg\min_\theta \mathrm{DKL}(p(x)\|q(x;\theta)). \quad (11)$$

No entanto, podemos ver que é difícil resolver (11) diretamente. O ADF resolve esse problema incluindo iterativamente a função de cada fator na distribuição posterior verdadeira. Assim, a princípio, o ADF escolhe q(x; θ0) para melhor aproximar a função fatorial p0(x) por meio de

$$\theta 0 = \arg\min_\theta \mathrm{DKL}(p0(x)\|q(x;\theta)). \quad (12)$$

Em seguida, o ADF atualizará a aproximação incorporando a próxima função de fator pi(yi|x) até que todas as funções de fator tenham sido envolvidas, o que nos dá a seguinte regra de atualização

$$\theta i = \arg\min_\theta \mathrm{DKL}(pi(x)q(x;\theta i-1)\|q(x;\theta)). \quad (13)$$

Como mostrado na Seção I-B, se q (x; θ) é escolhido da família exponencial, a solução ótima de (13) é combinar as estatísticas suficientes esperadas entre a distribuição aproximada q (x; θi) e a distribuição alvopi (x) q (x; θi − 1). Além disso, de acordo com (13), podemos ver que a melhor aproximação atual é baseada na melhor aproximação anterior. Por esse motivo, o desempenho de estimativa do ADF pode ser sensível à ordem de processo das funções fatoriais, o que pode produzir uma aproximação extremamente pobre em alguns casos. Na próxima seção, forneceremos outra perspectiva da regra de atualização do ADF, que resulta no algoritmo EP e fornece uma maneira de evitar a desvantagem associada ao algoritmo ADF.

II. PROPAGAÇÃO DE EXPECTATIVA

Adotando outra perspectiva, o ADF pode ser visto como aproximando sequencialmente a função fatorial pi(x) pela função fatorial aproximada p̃i(x), que é restrita dentro da família exponencial, e então atualizando exatamente a distribuição aproximada q(x; θ) multiplicando essas funções fatoriais aproximadas. Essa visão alternativa do ADF pode ser descrita como:

$$\tilde{p}_i(x) \propto \frac{q(x;\theta_i) q(x;\theta_{i-1})}{}, \tag{14}$$

que também produz o algoritmo EP. O algoritmo EP inicializa cada função fatorial pi(x) por uma função fatorial aproximada correspondente p̃i(x). Então, em iterações posteriores, EP revisita cada função fatorial aproximada p̃i(x)e a refina multiplicando todas as melhores estimativas atuais, exceto uma função fatorial verdadeira pi(x) do termo revisitante. Após várias iterações, a aproximação é obtida de acordo com (15).

$$q(x;\theta_*) \propto \prod_{eu} \tilde{p}_i(x). \tag{15}$$

Como esse procedimento não depende da ordem de processo da função fator, o EP fornece uma aproximação mais precisa do que o ADF.

O processo geral de PE é dado da seguinte forma:1) Inicialize o termo aproximação p̃i(x), que pode ser escolhido de um dos membros da família exponencial

com base no problema.2) Calcule a distribuição aproximada

$$q(x;\theta) = \frac{1}{Z} \prod_i \tilde{p}_i(x), \tag{16}$$

onde $Z = \int \prod_i \tilde{p}_i(x)dx$.3) Até que todos p̃i(x) convirjam:

a) Escolha p̃i(x) para refinar o aproximado.b) Remova p̃i(x) da distribuição aproximada atual q(x; θ) com um fator de normalização:

$$\frac{q(x;\theta\backslash i) \propto q(x;\theta)}{\tilde{p}_i(x)}. \tag{17}$$

c) Atualize q(x; θ), onde primeiro combinamos q(x; θ\i), pi(x) atual e um normalizador Zi e, em seguida, minimizamos a divergência KL através da projeção de correspondência de momentos (9) (ou seja, o operador Proj(·)):

$$q(x;\theta) = \text{Proj}\left(\frac{1}{Z_i} q(x;\theta\backslash i) p_i(x)\right). \tag{18}$$

d) Atualize p̃i(x) como

$$\tilde{p}_i(x) = Z_i \frac{q(x;\theta)}{q(x;\theta\backslash i)}. \tag{19}$$

4) Obtenha a distribuição aproximada final por meio de

$$p(x) \approx q(x;\theta_*) \propto \prod_{eu} \tilde{p}_i(x). \tag{20}$$

A. Relação com a BP

Esta seção mostra que o algoritmo BP é um caso especial de EP, onde EP fornece uma aproximação aprimorada para modelos nos quais BP é geralmente intratável.

Vamos primeiro fazer uma rápida revisão do algoritmo BP. O procedimento do algoritmo BP é atualizar iterativamente todos os nós variáveis, depois atualizar todos os nós de fator por meio do envio de mensagens em paralelo e, finalmente, atualizar a crença de cada variável após cada iteração. Ao adotar outro ponto de vista, o PB pode ser visto como uma atualização da crença sobre uma variável xi, incorporando um nó de fator de cada vez. Sob essa perspectiva, a crença da variável xi será atualizada como

$$b(ha)=mXi\rightarrow fs(ha)mfs\rightarrow Xi(ha)Zi, \tag{21}$$

onde Zi = $\int$ mXi→fs (xi) mfs→Xi (xi) dxi é o fator de normalização. Além disso, podemos interpretar vagamentemXi→fs (xi) e mfs→Xi (xi) como a mensagem anterior e de probabilidade, respectivamente.

Suponhamos que cada mensagem de verossimilhança mfs→Xi (xi) tenha uma forma complexa, por exemplo, uma mistura de múltiplas distribuições gaussianas. Então a complexidade computacional para avaliar as crenças exatas sobre todas as variáveis seria inviável. Em vez de propagar a mensagem de verossimilhança exata mfs→Xi (xi), o EP passa uma mensagem aproximada $\tilde{}$mfs→Xi (xi),onde$\tilde{}$mfs→Xi (xi) é obtido usando a operação de projeção conforme mostrado no processo geral de EP. Além disso $\tilde{}$mfs→Xi(xi) é geralmente escolhido da família exponencial para tornar o problema tratável. Assim, a crença aproximada na PE tem a seguinte forma

$$b(xi) \approx q(xi) \propto \prod_{s\in N(Xi)} \tilde{}mfs\rightarrow Xi(xi). \tag{22}$$

Para mostrar BP como um caso especial de EP, definimos ainda a crença parcial de um nó variável como

$$b(xi)\backslash fs = \frac{b(ha)\tilde{}}{mfs\rightarrow Xi(ha)} \propto \prod_{s'\in N(Xi)\backslash s} \tilde{}\, mfs' \rightarrow Xi\ (ha), \tag{23}$$

e a crença parcial de um nó de fator como

$$b(fs)\backslash Xi = \frac{b(fs)\tilde{}\, mXi\rightarrow fs}{(xi)}, \tag{24}$$

onde b(fs) = $\prod$j∈N(fs)$\tilde{}$mXj→fs (xj) é definido como a crença do nó do fator fs. Comparando com (18) e (19), a regra de atualização da mensagem do nó do fator no EP pode ser escrita como$\tilde{}$mfs→Xi (xi) = Proj (b(xi)\fsmfs→Xi (xi))b(xi)\fs

$$= \text{Projeto}\frac{(\ b(xi)\backslash fs\ \int xs\backslash xi\ fs\ (xs)\ b(fs)\backslash xi)}{b(xi)\backslash fs} \tag{25}$$

onde a integral funciona sobre variável contínua. Para variável discreta, pode-se simplesmente substituir integral por soma. Além disso, a nova crença b(xi) será aproximada como

$$b(xi) \approx qi(xi) = b(xi)\backslash fs\tilde{}\, mfs\rightarrow Xi\ (xi)Zi, \tag{26}$$

onde Zi = $\int$xi b(ha)\fs$\tilde{}$ mfs→Xi(ha).

Agora, se a integral em (25) é tratável (por exemplo, um modelo gaussiano linear) mesmo sem usar a projeção para aproximar mfs→Xi. Então b(xi)\fs em (25) pode ser cancelado. Finalmente, a regra de atualização de mensagem de nó de fator em EP é reduzida ao caso de PN padrão.

III. RELAÇÃO COM OUTROS MÉTODOS DE INFERÊNCIA VARIACIONAL

Nesta seção, descreveremos a relação entre EP e outros algoritmos de inferência variacional, por exemplo, Bayes variacional (VB). O modelo probabilístico bayesiano especifica a distribuição conjunta p(x, y), onde todas as variáveis ocultas em x recebem distribuições anteriores. O objetivo é encontrar a melhor aproximação para a distribuição posterior p(x|y). Vamos dar uma olhada na decomposição da distribuição da junta de toras da seguinte forma

$$\log p(x, y) = \log p(x|y) + \log p(y). \tag{27}$$

Reorganizando (27) e tomando a integral de ambos os lados da equação rearranjada em relação a uma dada distribuição q(x), obtemos a evidência do modelo logarítmico

$$\begin{aligned} \log p(y) &= \int q(x) \log(p(y))dx \\ &= \int q(x) \log(p(x, y)) - \int q(x) \log(p(x|y))dx, \end{aligned} \tag{28}$$

onde ∫ q(x)dx = 1. Então, reformatando (28), obtemos

$$\log p(y) = L(q(x)) + DKL(q(x)||p(x)), \tag{29}$$

onde definimos

$$L(q(x)) = \int q(x) \log\left(\frac{p(x, y)}{q(x)} \right)dx, \tag{30}$$

$$DKL(q(x)||p(x)) = \int q(x) \log\left(\frac{q(x)}{p(x|y)} \right)dx. \tag{31}$$

Uma vez que DKL(q(x)||p(x)) é um funcional não negativo, L(q(x)) dá o limite inferior de log p(y). Então a maximização do limite inferior L(q(x)) em relação à distribuição q(x) é equivalente a minimizar DKL(q(x)||p(x)),que acontece quando q(x) = p(x|y). No entanto, trabalhar com a verdadeira distribuição posterior p(x|y) pode ser intratável. Assim, assumimos que os elementos de x podem ser particionados em M grupos disjuntos xi, i = 1, 2, · · · , M . Em seguida, assumimos ainda que a fatoração da distribuição aproximada q (x) em relação a esses grupos tem a forma

$$q(x) = \prod_{eu}^{M} qi\ (xi). \tag{32}$$

Observe que a aproximação fatorada corresponde à teoria da média arquivada, que foi desenvolvida na física. Dadas as suposições acima mencionadas, agora tentamos encontrar qualquer distribuição possível q (x) sobre a qual o limite inferior L (q (x)) é maior. Uma vez que a maximização direta de (30) em relação a q(x) é difícil, em vez disso,

otimizar (30) em relação a cada um dos fatores em (32). Ao substituir (32) por (30), obtemos

$$L(q(x)) = \int q_j(x_j) \left[\int \log\left(p(x, y)\right) \prod_{i \neq j} q_i(x_i)\,dx_i \right] dx_j - \int q_j(x_j) \log(q_j(x_j))\,dx_j + \text{const}$$

$$= -\int q_j(x_j) \log(x_j) \tilde{p}(x_j, y)\,dx_j + \text{const} = -D_{KL}(q_j(x_j) \| \tilde{p}(x_j, y)) + \text{const}, \tag{33}$$

onde definimos $\tilde{p}(x_j, y)$ como

$$\tilde{p}(x_j, y) = \exp\left[\int \log\left(p(x, y)\right) \prod_{i \neq j} q_i(x_i)\,dx_i \right] = E_{i \neq j}[\log p(x, y)]. \tag{34}$$

Assim, se mantivermos todos os fatores $q_i(x_i)$ para $i \neq j$ fixos, então a maximização de (33) em relação a $q_j(x_j)$ é equivalente à minimização de $D_{KL}(q_j(x_j) \| \tilde{p}(x_j, y))$. Na prática, precisamos inicializar todos os fatores $q_i(x_i)$primeiro e, em seguida, atualizar iterativamente cada um dos fatores $q_j(x_j)$ minimizando o $D_{KL}(q_j(x_j) \| \tilde{p}(x_j, y))$, até que o algoritmo se confunda.

Agora podemos ver que a principal diferença entre EP e VB é a maneira de minimizar a divergência KL. A vantagem do VB é que ele fornece um limite inferior durante cada etapa de otimização, portanto, a convergência é garantida. No entanto, o VB pode causar subestimação da variância. Em EP, minimizar $D_{KL}(p(x)||q(x))$ é equivalência à "correspondência de momentos", mas a convergência não é garantida. No entanto, o EP tem um ponto fixo e, se convergir, o desempenho de aproximação do EP geralmente supera o VB.

REFERÊNCIAS

[1] S. Kullback e R. Leibler, "Sobre informação e suficiência", The Annals of Mathematical Statistics, vol. 22, nº 1, pp. 79–86, 1951.
[2] C. Bishop, Reconhecimento de padrões e aprendizado de máquina. Springer Nova York, 2006, vol. 4. [3] T. Minka, "Propagação de expectativa para inferência bayesiana aproximada", em Incerteza em Inteligência Artificial, vol. 17. Citeseer, 2001,

pág. 362-369. [4] ——, http://research.microsoft.com/en-us/um/people/minka/papers/ep/roadmap.html.