Aluno: Cleiton Moya de Almeida
EST5104 - Inferência Bayesiana
Prof. Josemar Rodrigues
12/05/2025

Lista 4

Enunciado

Gostaria de ver seus comentários nos exemplos 1, 4 e 7 do blog do Daniel

Resolução

Nos exemplos 1, 4 e 7 do Blog do Simpson, ele exemplifica o uso da *Penalised Complexity Prior* (PCP) com o modelo Binomial Negativa.

Binomial Negativa

Na forma mais usual, a distribuição Binomial Negativa conta o número de falhas k que ocorreram em uma sequência de experimentos de Bernoulli antes da ocorrência de um número n (prefixado, não aleatório) de sucessos. A função de massa de probabilidade é dada por:

$$p(X = k|n, p) = \binom{k+n-1}{k} p^n (1-p)^k \tag{1}$$

onde p é a probabilidade de sucesso. Esta distribuição possui média $\mu = \frac{n(1-p)}{p}$ e variância $\sigma^2 = \frac{n(1-p)}{p^2}$.

Uma outra forma usual de parametrização da Binomial Negativa é em termos da média μ e de um parâmetro de dispersão $\alpha = 1/n$. Esta é a forma usada pelo Simpson. A função de massa de probabilidade fica então:

$$\begin{aligned} p(X = k|\mu, \alpha) &= \binom{k+\alpha^{-1}-1}{\alpha^{-1}-1} \left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{1/\alpha} \left(\frac{\mu}{\mu+\alpha^{-1}}\right)^k \\ &= \frac{\Gamma(k+\alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(k+1)} \left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{1/\alpha} \left(\frac{\mu}{\mu+\alpha^{-1}}\right)^k \end{aligned} \tag{2}$$

com a variância dada por $\sigma^2 = \mu(1+\alpha\mu)$.

Modelo hierárquico

O modelo Binomial Negativa, $y_i \sim$ Binomial-Negativa(μ, α) pode ser escrito através do seguinte modelo hierárquico:

$$\begin{aligned} y_i \mid u_i &\sim \text{Poisson}(\mu u_i) \\ u_i &\sim \text{Gamma}(\alpha^{-1}, \alpha^{-1}), \quad i = 1 \dots n \end{aligned} \tag{3}$$

Ao marginalizar o modelo hierárquico eq. (3) com respeito a u_i, o modelo Binomial Negativa é retomado.

Penalised Complexity Prior

A ideia geral da PCP é identificar quais parâmetros tornam o modelo mais complexo e quais parâmetros não influenciam na complexidade. Os parâmetros associados à complexidade são denominados por Simpson como parâmetros de flexibilidade (neste relatório preferimos utilizar o termo *parâmetros de complexidade*). Para estes parâmetros, existe um valor de base tal que o modelo se reduz a um modelo base. Por exemplo, para o modelo Binomial Negativa, o parâmetro que torna o modelo complexo é o α, associado à dispersão. Quando $\alpha = 0$, o modelo se reduz ao modelo de Poisson (modelo base).

Então, a ideia é construir uma priori que favoreça o modelo simples e penalize o modelo complexo. Para isto, Simpson propõe usar a Divergência de Kullback-Leibler, definida por

$$\text{KL}(f||g) = \int_X f(x) \log \frac{f(x)}{g(x)} dx \tag{4}$$

para modelos contínuos ou

$$\text{KL}(p||q) = \sum_K p(k) \log \frac{p(k)}{q(k)} \tag{5}$$

para modelos discreto, onde f (ou p) é modelo complexo e g (ou q) o modelo de base.

Como a divergência KL não é originalmente uma medida de distância, Simpson propõe utilizar a seguinte métrica adaptada:

$$d(\xi) = \sqrt{2\,\text{KL}\left(f(\cdot|\xi)||f(\cdot|\xi_0)\right)} \tag{6}$$

onde ξ é o parâmetro de flexibilidade e ξ_0 é o valor de base.

Exemplo 1 - Identificando o parâmetro de complexidade

No exemplo 1 do blog, Simpson descreve que, no modelo binomial negativa, o parâmetro de média (μ) não é um parâmetro de complexidade, pois a média não aumenta ou diminui a complexidade do modelo. Já a dispersão (α) é um parâmetro de complexidade. No caso especial em que $\alpha \to 0$ (no limite), o modelo binomial negativa é equivalente ao modelo de Poisson (modelo de base), cuja função de massa de probabilidade é dada por

$$q(k|\mu) = \frac{\mu^k e^{-\mu}}{k!} = \frac{\mu^k e^{-\mu}}{\Gamma(k+1)}. \tag{7}$$

Observe então que $p(k|\mu, \alpha_0) \coloneqq \lim_{\alpha \to 0} p(k|\mu, \alpha) = q(k|\mu)$.

Exemplo 4 - Calculando a divergência KL

No exemplo 4, a divergência KL é calculada para ambos modelos (tradicional e hierárquico). Primeiro, calcula-se para o modelo tradicional (2):

$$
\begin{aligned}
\frac{1}{2}d^2(\alpha) &= \mathrm{KL}(p(k|\mu,\alpha)||q(k|\mu)) \\
&= \sum_{k=0}^{\infty} p(k|\mu,\alpha)\log\frac{p(k|\mu,\alpha)}{q(k|\mu)} \\
&= \sum_{k=0}^{\infty} \frac{\Gamma(k+\alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(k+1)}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{1/\alpha}\left(\frac{\mu}{\mu+\alpha^{-1}}\right)^{k} \\
&\quad\times\left\{\log\left(\frac{\Gamma(k+\alpha^{-1})}{\Gamma(\alpha^{-1})\cancel{\Gamma(k+1)}}\right)+a^{-1}\log\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)+\cancel{k\log(\mu)}-k\log(\mu+\alpha^{-1})\right. \\
&\quad\left.-\log\frac{\cancel{\mu^k}e^{-\mu}}{\cancel{\Gamma(k+1)}}\right\} \\
&= \sum_{k=0}^{\infty} \frac{\Gamma(k+\alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(k+1)}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{1/\alpha}\left(\frac{\mu}{\mu+\alpha^{-1}}\right)^{k} \\
&\quad\times\left\{\log\Gamma(k+\alpha^{-1})-\log\Gamma(\alpha^{-1})+\alpha^{-1}\log\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)-k\log(\mu+\alpha^{-1})+\mu\right\}
\end{aligned}
\tag{8}
$$

Note que o termo acima destacado em vermelho está ligeiramente diferente do apresentado no blog do Simpson. Acreditamos que trata-se de um erro de digitação no blog.

O principal problema de usar o modelo original é que a divergência KL fica em função do parâmetro μ, e não apenas do parâmetro de complexidade α. Para contornar isto, Simpson usa o modelo hierárquico, no qual o componente complexo $u_i \sim \text{Gamma}(\alpha^{-1},\alpha^{-1})$ depende apenas de α.

A divergência entre duas distribuições $P:\text{Gamma}(a^{-1},a^{-1})$ e $Q:\text{Gamma}(b^{-1},b^{-1})$ é dada por por[1]:

$$
\begin{aligned}
\mathrm{KL}(P||Q) &= b^{-1}\log\frac{a^{-1}}{b^{-1}}-\log\frac{\Gamma(a^{-1})}{\Gamma(b^{-1})}+(a^{-1}-b^{-1})\psi(a^{-1})-(a^{-1}-b^{-1})\frac{a^{-1}}{a^{-1}} \\
&= b^{-1}(\log a^{-1}-\psi(a^{-1}))+\log\Gamma(b^{-1})-b^{-1}\log b^{-1}+b^{-1}+o(b^{-1})
\end{aligned}
\tag{9}
$$

onde $\psi(\cdot)$ é a função digamma e $o(b^{-1})$ são os termos que dependem exclusivamente de a^{-1} e que, portanto, possuem ordem de crescimento menor que b^{-1} quando $b^{-1}\to\infty$.

Agora, Simpson utiliza a seguinte aproximação assintótica do logaritmo da função Gamma: $\log\Gamma(z)=z\log z-z$, válida para grandes valores de z ($z\to\infty$). Então, a equação eq. (9) reduz-se para

$$
\mathrm{KL}(P||Q) = b^{-1}\left(\log a^{-1}-\psi(a^{-1})\right)+o(b^{-1}). \tag{10}
$$

[1] `https://statproofbook.github.io/P/gam-kl.html`

Note que, se $b \to 0$, a divergência KL tende ao infinito. Então, em princípio, não se poderia utilizar tal métrica para a construção de uma priori. Porém, Simpson diz (sem maiores explicações) que pode-se simplesmente ignorar o termo b^{-1} da eq. (10). Com isto, a métrica de distância fica apenas em função do parâmetro de complexidade:

$$d(\alpha) = \sqrt{2\log\alpha^{-1} - 2\psi(\alpha^{-1})}. \tag{11}$$

A função digamma pode ser aproximada por $\psi(\alpha^{-1}) = \log(\alpha^{-1}) - \alpha/2$. Com isto, a eq. (11) pode ser aproximada por

$$d(\alpha) \approx \sqrt{\alpha}. \tag{12}$$

Esta aproximação revela que a distância adotada ($d(\alpha)$) é aproximadamente igual ao desvio padrão de u_i.

Exemplo 7 - Construção da PCP

Uma vez que se tem a métrica de distância $d(\alpha)$ entre a distribuição complexa e a distribuição de base, o próximo passo é como transformar esta métrica em uma distribuição. Para isto, Simpson parte de uma transformação de variáveis:

$$p(\alpha) = p_d(d(\alpha))\left|\frac{\partial}{\partial\alpha}d(\alpha)\right| \tag{13}$$

onde $p_d(\cdot)$ é a densidade a priori para a parametrização da métrica de distância. Agora, a tarefa é decidir sobre esta densidade a priori. Simpson sugere então utilizar a distribuição exponencial pois, devido à propriedade de falta de memória da exponencial, esta distribuição a priori implica em uma taxa de penalização constante, isto é, a taxa com que a densidade diminui não se altera ao longo do espaço paramétrico. Assim, a eq. (13) fica

$$\begin{aligned} p(\alpha) &= \lambda e^{-\lambda d(\alpha)}\left|\frac{\partial}{\partial\alpha}d(\alpha)\right| \\ &= \frac{\lambda}{\alpha^2}\frac{|\psi'(\alpha-1) - \alpha|}{\sqrt{2\log\alpha^{-1} - 2\psi(\alpha^{-1})}}\exp\left\{-\lambda\sqrt{2\log\alpha^{-1} - 2\psi(\alpha^{-1})}\right\} \end{aligned} \tag{14}$$

Se utilizarmos a forma aproximada da equação da distância eq. (12), a priori simplifica para

$$p(\alpha) = \frac{\lambda}{2\sqrt{\alpha}}e^{-\lambda\sqrt{\alpha}}. \tag{15}$$

Simpson conclui a seção mostrando que esta forma aproximada da eq. (15) é bastante razoável, apresentando um pequeno desvio em relação à eq. (14).

Conclusões

A *Penalised Complexity Prior* é uma proposta interessante de priori fracamente informativa a qual tem como objetivo penalizar a complexidade do modelo. Esta penalização

é realizada através de uma métrica de distância adaptada da Divergência de Kullback-Leibler. No caso do modelo Binomial Negativa, obtém-se com relativa facilidade uma expressão (aproximada) simples para a PCP. Entretanto, apesar de elegante, a construção da priori baseada na divergência KL é questionável porque esta métrica tende ao infinito quando o modelo complexo se aproxima do modelo base.