Introdução à Amostragem para Inferência Bayesiana - Aula 1

Eliezer de Souza da Silva, PhD
EMAp FGV / BCAM / UFC
eliezer@probabilistic.ai
https:\sereliezer.github.io

30 de Janeiro de 2025

Motivação para Modelagem Bayesiana

Problema: Modelagem de Linguagem Natural

- Suponha que queremos prever a próxima sentença em um diálogo.
- Há múltiplas possibilidades para a próxima sentença.
- Há incerteza tanto sobre o conteúdo exato quanto sobre a parametrização do modelo.

Motivação para Modelagem Bayesiana

Problema: Modelagem de Linguagem Natural

- Suponha que queremos prever a próxima sentença em um diálogo.
- Há múltiplas possibilidades para a próxima sentença.
- Há incerteza tanto sobre o conteúdo exato quanto sobre a parametrização do modelo.

Exemplo: Prevendo a próxima palavra

- Dado um conjunto de palavras anteriores, queremos prever a próxima palavra.
- Podemos modelar essa previsão como uma distribuição categórica condicionada às palavras anteriores.

$$P(w_{t+1}|w_1, w_2, ..., w_t) = \frac{e^{\theta^T \phi(w_1,...,w_t)}}{\sum_{w'} e^{\theta^T \phi(w_1,...,w_t,w')}} \tag{1}$$

Motivação para Modelagem Bayesiana I

Solução Bayesiana:

- Modelamos a incerteza através de uma distribuição sobre sentenças possíveis.
- Também modelamos a incerteza sobre os parâmetros do modelo, θ, usando um prior Bayesiano.
- Permite definir uma família de soluções para esse problema, bem como a incerteza associada a cada solução.

Motivação para Modelagem Bayesiana II

Distribuição sobre Parâmetros

- Definimos um prior para os parâmetros do modelo, por exemplo, um prior Gaussiano:

$$\theta \sim \mathcal{N}(0, \sigma^2 I) \tag{2}$$

- A posteriori dos parâmetros pode ser inferida a partir dos dados observados.

Inferência Bayesiana

- Calculamos a posterior dos parâmetros utilizando a regra de Bayes:

$$P(\theta|D) \propto P(D|\theta)P(\theta) \tag{3}$$

- Como efetivar essa procedimento analiticamente ou numericamente? O estudo de metodologias para inferência Bayesiana se concentra em resolver esse problema.

O que é Inferência Bayesiana?

Modelagem: Conjunto de dados observados $D \subset \mathcal{X}$ para algum domínio $\mathcal{X}$.

- $D = \{X_i, \ldots, X_n\}$ com dados não-anotados, ou com dados anotados onde $D = \{(X_i, Y_i), \ldots, (X_n, Y_n)\}$, onde Y_i corresponde a anotação (classe ou valor numérico, respectivamente para classificação ou regressão).

O que é Inferência Bayesiana?

Modelagem: Conjunto de dados observados $D \subset \mathcal{X}$ para algum domínio $\mathcal{X}$.

- $D = \{X_i, \ldots, X_n\}$ com dados não-anotados, ou com dados anotados onde $D = \{(X_i, Y_i), \ldots, (X_n, Y_n)\}$, onde Y_i corresponde a anotação (classe ou valor numérico, respectivamente para classificação ou regressão).
- $\theta \in \Theta$ para algum domínio de parâmetros, assumindo uma modelagem paramétrica que explique a distribuição dos dados observados.

O que é Inferência Bayesiana?

Modelagem: Conjunto de dados observados $D \subset \mathcal{X}$ para algum domínio $\mathcal{X}$.

- $D = \{X_i, \ldots, X_n\}$ com dados não-anotados, ou com dados anotados onde $D = \{(X_i, Y_i), \ldots, (X_n, Y_n)\}$, onde Y_i corresponde a anotação (classe ou valor numérico, respectivamente para classificação ou regressão).
- $\theta \in \Theta$ para algum domínio de parâmetros, assumindo uma modelagem paramétrica que explique a distribuição dos dados observados.

Inferência Bayesiana: Uso da regra de Bayes para atualizar crenças com base em dados observados.

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \tag{4}$$

O que é Inferência Bayesiana?

Modelagem: Conjunto de dados observados $D \subset \mathcal{X}$ para algum domínio $\mathcal{X}$.

- $D = \{X_i, \ldots, X_n\}$ com dados não-anotados, ou com dados anotados onde $D = \{(X_i, Y_i), \ldots, (X_n, Y_n)\}$, onde Y_i corresponde a anotação (classe ou valor numérico, respectivamente para classificação ou regressão).
- $\theta \in \Theta$ para algum domínio de parâmetros, assumindo uma modelagem paramétrica que explique a distribuição dos dados observados.

Inferência Bayesiana: Uso da regra de Bayes para atualizar crenças com base em dados observados.

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \tag{4}$$

Problema: O cálculo da integral $P(D) = \int P(D|\theta)P(\theta)d\theta$ pode ser intratável.

Definição de Conjugação

Modelo Conjugado: Um prior $P(\theta)$ é conjugado para a likelihood $P(D|\theta)$ se a posterior $P(\theta|D)$ pertence à mesma família de distribuições de $P(\theta)$. Em outras palavras, se a forma funcional da priori e da posteriori para a variável θ for a mesma, podemos inferir o valor do termo de normalização da posteriori.

Definição de Conjugação

Modelo Conjugado: Um prior $P(\theta)$ é conjugado para a likelihood $P(D|\theta)$ se a posterior $P(\theta|D)$ pertence à mesma família de distribuições de $P(\theta)$. Em outras palavras, se a forma funcional da priori e da posteriori para a variável θ for a mesma, podemos inferir o valor do termo de normalização da posteriori.

Expressão Proporcional:

$$P(\theta|D) \propto P(D|\theta)P(\theta) \tag{5}$$

Exemplo: Poisson-Gamma

- Likelihood: $Y|\lambda \sim \text{Poisson}(\lambda)$
- Prior: $\lambda \sim \text{Gamma}(\alpha, \beta)$
- Posterior: $\lambda|Y \sim \text{Gamma}(\alpha + Y, \beta + 1)$

Definição de Conjugação

Modelo Conjugado: Um prior $P(\theta)$ é conjugado para a likelihood $P(D|\theta)$ se a posterior $P(\theta|D)$ pertence à mesma família de distribuições de $P(\theta)$. Em outras palavras, se a forma funcional da priori e da posteriori para a variável θ for a mesma, podemos inferir o valor do termo de normalização da posteriori.

Expressão Proporcional:

$$P(\theta|D) \propto P(D|\theta)P(\theta) \tag{5}$$

Exemplo: Poisson-Gamma

- Likelihood: $Y|\lambda \sim \text{Poisson}(\lambda)$
- Prior: $\lambda \sim \text{Gamma}(\alpha, \beta)$
- Posterior: $\lambda|Y \sim \text{Gamma}(\alpha + Y, \beta + 1)$

Benefícios da Conjugação:

- Facilita cálculos analíticos.
- Reduz o custo computacional.
- Aparece naturalmente para certas classes de modelos (ex. família exponencial).

Limitações da Conjugação

Problemas com Modelos Conjugados:

- Nem sempre há um prior conjugado disponível.
- Conjugação pode impor restrições não realistas sobre os parâmetros.
- Modelos conjugados podem ser inflexíveis para capturar complexidade dos dados.
- Uma certa confusão entre propriedade do modelo e facilidade computacional. Propriedade desejável computacionalmente, não significa que deve ser imposta!

Limitações da Conjugação

Problemas com Modelos Conjugados:

- Nem sempre há um prior conjugado disponível.
- Conjugação pode impor restrições não realistas sobre os parâmetros.
- Modelos conjugados podem ser inflexíveis para capturar complexidade dos dados.
- Uma certa confusão entre propriedade do modelo e facilidade computacional. Propriedade desejável computacionalmente, não significa que deve ser imposta!

Soluções:

- Amostragem:
  - Amostragem por Rejeição
  - Monte Carlo via Cadeias de Markov (MCMC), Gibbs Sampling
  - Hamiltonian Monte Carlo (HMC)
- Métodos variacionais (não será coberto nessa aulas).

Exemplo: Modelo Poisson-Gamma I

Contexto: contagem de eventos $X_i \in \mathbb{N}$ com parâmetro não-negativo λ referente a taxa de eventos observados. Exemplo: modelo de fila simples sem memória para intervalos fixos.

Exemplo: Modelo Poisson-Gamma II

Modelo:

- Likelihood: $X_i|\lambda \sim \text{Poisson}(\lambda)$, $P(X_i|\lambda) = \frac{\lambda^{X_i} e^{-\lambda}}{X_i!}$
- Prior: $\lambda \sim \text{Gamma}(\alpha, \beta)$, $P(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$

Exemplo: Modelo Poisson-Gamma III

Posterior: dados observados $D = \{X_1, \ldots, X_n\}$ e $Y = \sum_{i=1}^{n} X_i$

$$
\begin{aligned}
P(\lambda|D) &\propto P(\lambda) \prod_{i=1}^{n} P(X_i|\lambda) \\
&= \frac{\lambda^{\sum_{i=1}^{n} X_i} e^{-\lambda}}{(\sum_{i=1}^{n} X_i)!} \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \\
&\propto \lambda^{(\alpha+Y)-1} e^{-(\beta+n)\lambda}
\end{aligned}
$$

Resultado:

$$
\lambda|X_1, \ldots, X_n \sim \text{Gamma}(\alpha + \sum_{i=1}^{n} X_i, \beta + n)
$$

$$
\mathbb{E}[\lambda|X_1, \ldots, X_n] = \frac{\alpha + \sum_{i=1}^{n} X_i}{\beta + n}
$$

Exemplo: Modelo Poisson-LogNormal I

Contexto: Contagem de eventos $X_i \in \mathbb{N}$ com parâmetro não-negativo λ, onde agora assumimos um prior Log-Normal para λ.

Exemplo: Modelo Poisson-LogNormal II

Modelo:

- Likelihood: $X_i|\lambda \sim \text{Poisson}(\lambda)$, $P(X_i|\lambda) = \frac{\lambda^{X_i} e^{-\lambda}}{X_i!}$
- Prior: $\lambda \sim \text{LogNormal}(\mu, \sigma^2)$, $P(\lambda) = \frac{1}{\lambda\sqrt{2\pi\sigma^2}} e^{-\frac{(\log\lambda - \mu)^2}{2\sigma^2}}$

Exemplo: Modelo Poisson-LogNormal III

Posterior: Dados observados $D = \{X_1, \ldots, X_n\}$ e $Y = \sum_{i=1}^{n} X_i$

$$
\begin{aligned}
P(\lambda|D) &\propto P(\lambda) \prod_{i=1}^{n} P(X_i|\lambda) \\
&\propto \left(\frac{1}{\lambda\sqrt{2\pi\sigma^2}} e^{-\frac{(\log\lambda-\mu)^2}{2\sigma^2}} \right) \times \left(\lambda^Y e^{-n\lambda} \right) \\
&\propto \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\log\lambda-\mu)^2}{2\sigma^2}} \right) \times \left(\lambda^{Y-1} e^{-n\lambda} \right)
\end{aligned}
$$

Exemplo: Modelo Poisson-LogNormal IV

Dificuldade:

- A distribuição posterior resultante não pertence a uma família conhecida, como ocorre no caso Gamma.
- O termo $\lambda^{Y-1}e^{-n\lambda}$ combinado com a exponencial quadrática do Log-Normal impede uma simplificação analítica.

Solução: Amostragem

- Como não há forma fechada para a posterior, precisamos amostrar de uma distribuição não normalizada

$$\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(\log\lambda-\mu)^2}{2\sigma^2}}\right)\times\left(\lambda^{Y-1}e^{-n\lambda}\right)$$

Exemplo: Regressão Bayesiana

Modelo:

$$y_i|\beta \sim \mathcal{N}(X_i^T\beta, \sigma^2), \quad P(y_i|\beta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - X_i^T\beta)^2}{2\sigma^2}} \tag{6}$$

Prior não conjugado:

$$P(\beta) \propto e^{-\frac{|\beta|}{b}} \quad \text{(Laplace)} \tag{7}$$

Posterior: sem solução analítica, exigindo amostragem. A prior não conjugada (Laplace) impede a simplificação.

Exemplo: Regressão Logística Bayesiana

Modelo:

$$P(y_i = 1|x_i, \beta) = \sigma(x_i^T \beta), \quad \sigma(z) = \frac{1}{1 + e^{-z}} \tag{8}$$

Posterior: sem forma fechada, exigindo métodos de amostragem. A não-linearidade da função logística torna a posterior analiticamente intratável.

Exemplo: Fatoração Matricial Probabilística

Modelo:

$$U_{ik} \sim \text{Gamma}(\alpha, \beta)$$

$$V_{jk} \sim \text{Gamma}(\alpha, \beta)$$

$$\mathbb{E}[R_{ij}] \approx U_i^T V_j \tag{9}$$

Likelihood:

$$P(R_{ij}|U, V) = \frac{(U_i^T V_j)^{R_{ij}} e^{-U_i^T V_j}}{R_{ij}!} \quad \text{(Poisson)} \tag{10}$$

Posterior: intratável, exigindo métodos de amostragem. A complexidade do modelo (muitos parâmetros U e V) torna a posterior intratável.

Exemplo: Regressão Poisson-Gamma com features

Modelo:

- Likelihood: $Y_i|\lambda_i \sim \text{Poisson}(\lambda_i)$, $P(Y_i|\lambda_i) = \frac{\lambda_i^{Y_i} e^{-\lambda_i}}{Y_i!}$
- $\lambda_i = \exp(\mathbf{x_i}^T \boldsymbol{\beta})$ onde $\boldsymbol{\beta} \sim \text{Gamma}(\alpha, \beta)$

Posterior:

$$P(\boldsymbol{\beta}|\mathbf{Y}) \propto P(\mathbf{Y}|\boldsymbol{\beta})P(\boldsymbol{\beta})$$

$$= \prod_i \frac{\exp(\mathbf{x_i}^T \boldsymbol{\beta})^{Y_i} e^{-\exp(\mathbf{x_i}^T \boldsymbol{\beta})}}{Y_i!} \prod_j \frac{\beta^{\alpha_j}}{\Gamma(\alpha_j)} \beta_j^{\alpha_j - 1} e^{-\beta \beta_j}$$

Resultado: Não possui forma fechada. A presença do exponencial dentro da função de probabilidade da Poisson, junto com a multiplicação do vetor gama pelas features, impede que a posterior simplifique para uma distribuição conhecida.

Introdução às Técnicas de Amostragem I

Motivação: Muitas distribuições posteriores não possuem uma forma fechada conhecida, tornando a amostragem uma ferramenta essencial para inferência Bayesiana.

Definição Matemática:

- Seja $X \sim p(x)$ uma variável aleatória com distribuição desconhecida.
- Amostragem consiste em gerar N amostras $\{X_1, ..., X_N\}$ de $p(x)$ tal que a distribuição empírica:

$$\hat{p}_N(x) = \frac{1}{N} \sum_{i=1}^{N} \delta(X_i - x) \tag{11}$$

  aproxima $p(x)$ conforme $N \to \infty$.
- Além disso, queremos que, para qualquer função mensurável g, a esperança seja aproximada por:

$$\mathbb{E}_p[g(X)] \approx \mathbb{E}_{\hat{p}_N}[g(X)] = \frac{1}{N} \sum_{i=1}^{N} g(X_i) \tag{12}$$

Introdução às Técnicas de Amostragem III

Principais Métodos:

- Inversão da CDF
- Amostragem por Rejeição
- Amostragem por Importância
- Métodos MCMC (Metropolis-Hastings, HMC)

Amostragem via Inversão da CDF

Definição: Se temos a função de distribuição acumulada (CDF) $F(x)$ de uma variável aleatória contínua, podemos gerar amostras de sua distribuição invertendo F.

Amostragem via Inversão da CDF

Definição: Se temos a função de distribuição acumulada (CDF) $F(x)$ de uma variável aleatória contínua, podemos gerar amostras de sua distribuição invertendo F.

Teorema: Se $U \sim \text{Uniform}(0, 1)$, então $X = F^{-1}(U)$ segue a distribuição desejada.

Amostragem via Inversão da CDF

Definição: Se temos a função de distribuição acumulada (CDF) $F(x)$ de uma variável aleatória contínua, podemos gerar amostras de sua distribuição invertendo F.

Teorema: Se $U \sim \text{Uniform}(0, 1)$, então $X = F^{-1}(U)$ segue a distribuição desejada.

Exemplo: Distribuição Exponencial

$$F(x) = 1 - e^{-\lambda x} \Rightarrow x = -\frac{\log(1 - U)}{\lambda} \tag{13}$$

Conexão com Normalizing Flows I

Mudança de Variável: Em modelos como Normalizing Flows, utilizamos a regra de mudança de variável para transformar distribuições simples em distribuições complexas.
Se temos uma transformação $X = T(U)$, então a densidade de probabilidade se transforma como:

$$p_X(x) = p_U(T^{-1}(x)) \left| \frac{d}{dx} T^{-1}(x) \right| \tag{14}$$

Aplicando essa ideia à inversão da CDF, podemos justificar

Conexão com Normalizing Flows II

matematicamente a geração de amostras.

Amostragem por Rejeição

Definição: Um método para gerar amostras de uma distribuição-alvo $p(x)$ usando uma distribuição proposta $g(x)$ e um fator de escala M.

Amostragem por Rejeição

Definição: Um método para gerar amostras de uma distribuição-alvo $p(x)$ usando uma distribuição proposta $g(x)$ e um fator de escala M.

Passos:

- Escolha $g(x)$ tal que $p(x) \leq Mg(x)$ para todo x.

Amostragem por Rejeição

Definição: Um método para gerar amostras de uma distribuição-alvo $p(x)$ usando uma distribuição proposta $g(x)$ e um fator de escala M.

Passos:

- Escolha $g(x)$ tal que $p(x) \leq Mg(x)$ para todo x.
- Amostre $x \sim g(x)$.

Amostragem por Rejeição

Definição: Um método para gerar amostras de uma distribuição-alvo $p(x)$ usando uma distribuição proposta $g(x)$ e um fator de escala M.
Passos:

- Escolha $g(x)$ tal que $p(x) \leq Mg(x)$ para todo x.
- Amostre $x \sim g(x)$.
- Aceite x com probabilidade $p(x)/(Mg(x))$.

A quantidade expressa através da razão de duas densidades $p(x)/(Mg(x))$ aparece em diferentes métodos relacionados a estimação de densidade e teste de hipóteses.

Amostragem por Rejeição e Estimação de Densidade com GANs

Amostragem por rejeição: amostramos da distribuição proposta $x \sim g(x)$ e aceitamos com probabilidade $\frac{p(x)}{Mg(x)}$.

Amostragem por Rejeição e Estimação de Densidade com GANs

Amostragem por rejeição: amostramos da distribuição proposta $x \sim g(x)$ e aceitamos com probabilidade $\frac{p(x)}{Mg(x)}$.

Generative Adversarial Networks (GANs), o discriminador modela a probabilidade de aceitar uma amostra:

$$D(x) = \frac{p(x)}{p(x) + g(x)} \tag{15}$$

Podemos reescrever isso como uma **estimação da razão de densidades**:

$$\frac{p(x)}{g(x)} = \frac{D(x)}{1 - D(x)} \tag{16}$$

Modelamos implicitamente a razão de densidade sem precisar da normalização explícita de $p(x)$.

Amostragem por Rejeição e Estimação de Densidade com GANs

Amostragem por rejeição: amostramos da distribuição proposta $x \sim g(x)$ e aceitamos com probabilidade $\frac{p(x)}{Mg(x)}$.
Generative Adversarial Networks (GANs), o discriminador modela a probabilidade de aceitar uma amostra:

$$D(x) = \frac{p(x)}{p(x) + g(x)} \tag{15}$$

Podemos reescrever isso como uma **estimação da razão de densidades**:

$$\frac{p(x)}{g(x)} = \frac{D(x)}{1 - D(x)} \tag{16}$$

Modelamos implicitamente a razão de densidade sem precisar da normalização explícita de $p(x)$.
O gerador $G(z)$ aprende a mapear uma distribuição latente para a distribuição de interesse.

Exemplo Computacional

Código

🔗 Google Colab

Introdução ao MCMC

Markov Chain Monte Carlo (MCMC): Uma classe de métodos para amostragem de distribuições complexas, especialmente quando a normalização é intratável.

Introdução ao MCMC

Markov Chain Monte Carlo (MCMC): Uma classe de métodos para amostragem de distribuições complexas, especialmente quando a normalização é intratável.

Principais Ideias:

- Utiliza cadeias de Markov para gerar amostras dependentes.
- Converge para a distribuição alvo após um número suficiente de iterações.

Introdução ao MCMC

Markov Chain Monte Carlo (MCMC): Uma classe de métodos para amostragem de distribuições complexas, especialmente quando a normalização é intratável.

Principais Ideias:

- Utiliza cadeias de Markov para gerar amostras dependentes.
- Converge para a distribuição alvo após um número suficiente de iterações.

Exemplos de Algoritmos MCMC:

- Metropolis-Hastings: propõe novos pontos e aceita com uma certa taxa.
- Gibbs Sampling: amostragem condicional de uma variável por vez.
- Hamiltonian Monte Carlo (HMC): usa gradientes para explorar o espaço de amostragem de maneira mais eficiente.