STATS 230: Computational Statistics
Posterior Approximation

Babak Shahbaba

Department of Statistics, UCI

Outline

- When it is difficult to sample from a distribution, we can approximate it with one of the standard distributions such as normal
- In this lecture, we discuss such approximation methods in general; however, in statistics these methods are usually used for posterior approximation
- In what follows, we start with a simple method called Laplace's approximation
- We then discuss variational methods and illustrate it with a simple application

Laplace's approximation

- To approximate a distribution, we could focus on the high probability regions, more specifically, where the probability is the highest (at least locally) , i.e., the mode ($\hat{x}$) of the distribution.

- For this, we can either obtain $\hat{x}$ analytically or use optimization algorithms such as Newton's method

- We first discuss the univariate case, and then generalize it to multivariate problems

Laplace's approximation

- Assume that we know the distribution (e.g., posterior distribution) up a constant: $f(x) = f^*(x)/Z$.
- Denote the log of unnormalized density as $\ell(x) = log(f^*(x))$.
- Find the first derivative $\ell'(x)$ and the second derivative $\ell''(x)$.
- Start with an initial value x_0.
- At each iteration n, use the Taylor series expansion (up to the quadratic term) around the current value of x_n

$$\ell(x) \simeq \ell(x_n) + \ell'(x_n)(x - x_n) + \frac{1}{2}\ell''(x_n)(x - x_n)^2$$

- Taking the derivative of $\ell(x)$, setting it to zero, and solving for x gives the next guess x_{n+1}

$$x_{n+1} = x_n - \frac{\ell'(x_n)}{\ell''(x_n)}$$

Laplace's approximation

- Now assume that we have found the mode $\hat{x}$.

- We write down the Taylor series expansion (up to the quadratic term) around $\hat{x}$ (note that $\ell'(\hat{x}) = 0$)

$$\ell(x) \simeq \ell(\hat{x}) + \frac{1}{2}\ell''(\hat{x})(x-\hat{x})^2$$

- Rewrite the above approximation as follows

$$\begin{aligned} \ell(x) &\simeq \ell(\hat{x}) - \frac{1}{2c}(x-\hat{x})^2 \\ c &= [-\ell''(\hat{x})]^{-1} \end{aligned}$$

which means

$$f^*(x) \simeq f^*(\hat{x}) \exp[-\frac{1}{2c}(x-\hat{x})^2]$$

Laplace's approximation

- The right hand side is the density of a normal up to a constant.
- Therefore, we can approximate the distribution with

$$N(\hat{x}, [-\ell''(\hat{x})]^{-1})$$

- For multivariate distributions, we have

$$N(\hat{\mathbf{x}}, [-\nabla^2 \ell(\hat{\mathbf{x}})]^{-1})$$

Normalizing constant

- We can use the above approach to approximate an integral

$$Z = \int f^*(x)dx$$

  for example, to find the normalizing constant of the posterior distribution

- As before, assume that the function $f^*(x)$ peaks at a $\hat{x}$. We have

$$\int f^*(x)dx \quad \simeq \quad \int f^*(\hat{x}) \exp[-\frac{1}{2c}(x-\hat{x})^2]dx$$

- Then, we can approximate $Z = \int f^*(x)dx$ as follows:

$$Z = f^*(\hat{x})\sqrt{2\pi c}$$

- When $x = (x_1, \ldots, x_k)$,

$$Z = f^*(\hat{x})\sqrt{(2\pi)^k|\Sigma|}$$

Variational methods

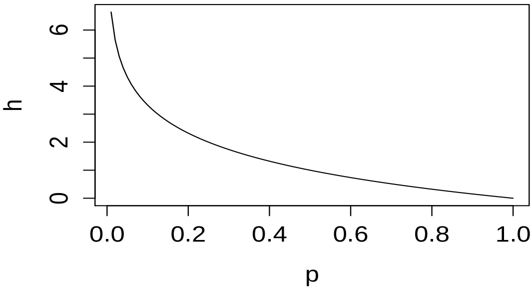
- Normal approximation does not always work.
- In what follows, we discuss a more general approach based on variational methods.
- This approach is inspired by information theory, so we first review some fundamental concepts in this field.

Information theory

- Information theory deals with communication problems.

- Information content (in bits) of an outcome x is

$$h(X = x) = \log_2 \frac{1}{P(X = x)}$$

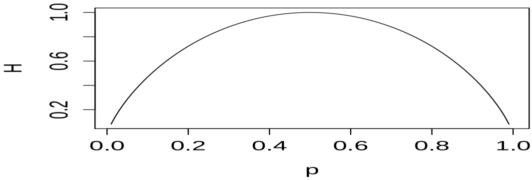


Entropy

- Shannon information content is highest for outcomes with low probability.
- For a set of outcomes, entropy is defined as the average Shannon information:

$$\begin{aligned} H(X) &= \sum_x P(x) \log_2 \frac{1}{P(x)} \\ &= -\sum_x P(x) \log_2 P(x) \end{aligned}$$

- Suppose there are only two possible outcomes with probabilities p and $1-p$,



Relative entropy

- The relative entropy between two probability distributions $P(x)$ and $Q(x)$ is defined as

$$D_{KL}(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

which satisfies Gibbs' inequality

$$D_{KL}(P||Q) \geq 0$$

- Note that this is not symmetric in general so $D_{KL}(P||Q)$ is not the same as

$$D_{KL}(Q||P) = \sum_x Q(x) \log \frac{Q(x)}{P(x)}$$

- This is known as Kullback-Leibler divergence

Variational methods

- Now suppose we have a complex (e.g., high dimensional) probability distribution $P(x)$,

$$P(x) = \frac{1}{z} e^{-E(x)}, \qquad \text{where } x = (x_1, \ldots, x_d)$$

- We want to approximate $P(x)$ by $Q(x, \theta)$ through adjusting θ to get the best approximation.

- θ is called "variational parameters."

- We define "best" in terms of minimum Kullback-Leibler divergence between the two distributions.

Variational Bayes

- As an example, consider the following model:

$$
\begin{aligned}
x|\mu,\sigma^2 &\sim N(\mu,\sigma^2)\\
\mu &\sim N(\mu_0,\sigma_0^2)\\
\sigma^2 &\sim \text{Inv-Gamma}(\alpha,\beta)
\end{aligned}
$$

- Posterior distribution of model parameters is

$$
\begin{aligned}
P(\mu,\sigma^2|x) &= \frac{P(\mu,\sigma^2,x)}{P(x)}\\
&= \frac{P(x|\mu,\sigma^2)P(\mu)P(\sigma^2)}{\int P(x|\mu,\sigma^2)P(\mu)P(\sigma^2)d\mu d\sigma^2}
\end{aligned}
$$

- This is not a tractable distribution in general.
- We want to approximate $P(\mu,\sigma^2|x)$ with a tractable distribution, $Q(\theta)$ that depends on variational parameters θ.

Variational Bayes

- We find a member of $Q(\theta)$ family (i.e., find optimum θ) as an approximation to the posterior distribution by minimizing the KL divergence,

$$
\begin{aligned}
D &= \int_Q \log \frac{Q(\mu, \sigma^2|\theta)}{P(\mu, \sigma^2|x)} Q(\mu, \sigma^2|\theta) d\mu d\sigma^2 \\
&= E_Q[\log Q(\mu, \sigma^2|\theta)] - E_Q[\log P(\mu, \sigma^2, x)] + \log P(x) \\
&= -\mathcal{L}(Q) + \text{ constant}
\end{aligned}
$$

- To minimize D, we need to minimize $-\mathcal{L}(Q)$, or alternatively, maximize $\mathcal{L}(Q)$,

$$
\mathcal{L}(Q) = E_Q[\log P(\mu, \sigma^2, x)] - E_Q[\log Q(\mu, \sigma^2|\theta)]
$$

Variational Bayes

- In the above example, using Jensen's inequality and concavity of the logarithm function, we have

$$
\begin{aligned}
\log P(x) &= \log \int P(\mu, \sigma^2, x) d\mu d\sigma^2 \\
&= \log \int P(\mu, \sigma^2, x) \frac{Q(\mu, \sigma^2|\theta)}{Q(\mu, \sigma^2|\theta)} d\mu d\sigma^2 \\
&= \log E_Q[\frac{P(\mu, \sigma^2, x)}{Q(\mu, \sigma^2|\theta)}] \\
&\geq E_Q[\log \frac{P(\mu, \sigma^2, x)}{Q(\mu, \sigma^2|\theta)}] \\
&= E_Q[\log P(\mu, \sigma^2, x)] - E_Q[\log Q(\mu, \sigma^2|\theta)] = \mathcal{L}(Q)
\end{aligned}
$$

- Therefore, $\mathcal{L}(Q)$ is the lower bound of $\log P(x)$, i.e., the logarithm of the marginal probability of the observed data

Variational Bayes

- Minimizing $-\mathcal{L}(Q)$ in general would not be easy.
- To make this simple, we usually assume that $Q(\mu, \sigma^2|\theta)$ factorizes.
- For the above example, we assume

$$Q(\mu, \sigma^2|\theta) \quad = \quad Q(\mu|m, v)Q(\sigma^2|a, b)$$

- More specifically,

$$\begin{aligned} \mu|m, v &\sim N(\mu|m, v) \\ \sigma^2|a, b &\sim \mathrm{Inv\text{-}Gamma}(\sigma^2|a, b) \end{aligned}$$

- We can minimize $-\mathcal{L}(Q)$ using gradient descent (or coordinate descent, when possible) methods,

$$\theta^{t+1} = \theta^t - \eta \nabla_\theta \mathcal{L}(Q)$$

Variational Bayes

Approximate posterior distribution using variational Bayes. Left panel: True posterior distribution; Right panel: Variational Bayes approximation.