Variational Bayesian Linear Regression

Seungjin Choi
Department of Computer Science
Pohang University of Science and Technology, Korea
seungjin@postech.ac.kr

Outline

- Linear regression: A quick review
- Bayesian treatment of linear regression
- Variational approximation
- Marginal likelihood maximization

Linear Regression: Revisited

- Linear regression assumes
  \[ y_t = w^T x_t + \epsilon_t, \]
  where the response \( y_t \) is a linear function of covariate \( x_t \in \mathbb{R}^D \) and is linear in the parameters as well.
- Collecting \( N \) response variables (denoted by \( y \in \mathbb{R}^N \)), we have
  \[ y = Xw + \epsilon, \]
  where \( X \in \mathbb{R}^{N \times D} \) is referred to as design matrix.
- Given a training set \( \{X, y\} \), estimate \( w \) so that the response \( y_s \) to a new data point \( x_s \) can be predicted, i.e., \( \mathbb{E}(y_s | x_s) = w^T x_s \).

Graphical Model for Bayesian Linear Regression

- The joint distribution of all the variables is given by
  \[ p(y, X, w, \alpha) = p(y | X, w)p(w | \alpha)p(\alpha). \]
- The likelihood for \( w \) is given by
  \[ p(y | X, w) = \prod_{t=1}^{N} \mathcal{N}(y_t | w^T x_t, \beta^{-1}). \]
- The prior over \( w \) is given by
  \[ p(w | \alpha) = \mathcal{N}(w | 0, \alpha^{-1}I). \]
- The prior over \( \alpha \) is given by
  \[ p(\alpha) = \text{Gam}(\alpha | a_0, b_0). \]
Variational Distribution

- Employ a variational posterior distribution $q(w, \alpha) = q(w)q(\alpha)$ which is an approximation to the posterior distribution $p(w, \alpha|X, y)$.

\[ q(w, \alpha) = q(w)q(\alpha) \approx p(w, \alpha|X, y). \]

- Compute optimized $q^*(w)$ and $q^*(\alpha)$ using VBEM:

\[
\log q^*(w) = \mathbb{E}_w \log p(X, y, w, \alpha) + \text{const},
\]

\[
\log q^*(\alpha) = \mathbb{E}_\alpha \log p(X, y, w, \alpha) + \text{const}.
\]

Optimized $q^*(w)$

\[
\log q^*(w) = \mathbb{E}_\alpha \log [p(y|X, w)p(w|\alpha)p(\alpha)] + \text{const}
\]

\[
= \log p(y|X, w) + \mathbb{E}_\alpha p(w|\alpha) + \text{const}
\]

\[
= -\frac{\beta}{2} \sum_{t=1}^{N} (y_t - w^T x_t)^2 - \frac{1}{2} \mathbb{E}_\alpha \{ \alpha \} w^T w + \text{const}
\]

\[
= -\frac{\beta}{2} (y - Xw)^T (y - Xw) - \frac{1}{2} \mathbb{E}_\alpha \{ \alpha \} w^T w + \text{const}
\]

\[
= -\frac{1}{2} w^T \left( \mathbb{E}_\alpha \{ \alpha \} I + \beta X^T X \right) w + \beta w^T X^T y + \text{const},
\]

which is the log of a Gaussian distribution, leading to:

\[
q^*(w) = \mathcal{N}(w | m_N, S_N),
\]

\[
m_N = \beta S_N X^T y,
\]

\[
S_N = \left( \mathbb{E}_\alpha \{ \alpha \} I + \beta X^T X \right)^{-1},
\]

Optimized $q^*(\alpha)$

\[
\log q^*(\alpha) = \mathbb{E}_w \log p(X, y, w, \alpha) + \text{const}
\]

\[
= \mathbb{E}_w \log [p(y|X, w)p(w|\alpha)p(\alpha)] + \text{const}
\]

\[
= \log p(\alpha) + \mathbb{E}_w \log p(w|\alpha) + \text{const}
\]

\[
= (\alpha_0 - 1) \log \alpha - b_0 \alpha + \frac{D}{2} \log \alpha - \frac{\alpha}{2} \mathbb{E}_w \{ w^T w \} + \text{const}
\]

\[
= \left( a_0 + \frac{D}{2} - 1 \right) \log \alpha - \left( b_0 + \frac{\alpha}{2} \mathbb{E}_w \{ w^T w \} \right) \alpha,
\]

which is the log of a Gamma distribution, leading to:

\[
q^*(\alpha) = \text{Gam}(\alpha | a_N, b_N),
\]

\[
a_N = a_0 + \frac{D}{2},
\]

\[
b_N = b_0 + \frac{1}{2} \mathbb{E}_w \{ w^T w \}.
\]

Algorithm Outline

- Re-estimate $q^*(\alpha)$:

\[
q^*(\alpha) = \text{Gam}(\alpha | a_N, b_N),
\]

\[
a_N = a_0 + \frac{D}{2},
\]

\[
b_N = b_0 + \frac{1}{2} \mathbb{E}_w \{ w^T w \},
\]

\[
\mathbb{E}_w \{ w^T w \} = m_N^T m_N + \text{tr} \{ S_N \}.
\]

- Re-estimate $q^*(w)$:

\[
q^*(w) = \mathcal{N}(w | m_N, S_N),
\]

\[
m_N = \beta S_N X^T y,
\]

\[
S_N = \left( \mathbb{E}_\alpha \{ \alpha \} I + \beta X^T X \right)^{-1},
\]

\[
\mathbb{E}_\alpha \{ \alpha \} = \frac{a_N}{b_N}. 
\]
Predictive Distribution

The predictive distribution over \( y \), given a new input \( x_* \), is evaluated using the Gaussian variational posterior:

\[
p(y_*|x_*, X, y) = \int p(y_*|w, x_*, X, y)dw
= \int p(y_*|w, x_*)p(w|X, y)dw
\approx \int p(y_*|w, x_*)q(w)dw
= \mathcal{N}(y_*|w^T x_*, \beta^{-1}) \mathcal{N}(w|m_N, S_N) dw
= \mathcal{N}(y_*|m_N^T x_*, \beta^{-1} + x^T S_N x_*) .
\]

Marginal and Conditional Gaussians

The last equality was derived using the following Gaussian identify.

Given a marginal Gaussian distribution for \( x \) and a conditional Gaussian distribution for \( y \) given \( x \) in the form

\[
p(x) = \mathcal{N}(x | \mu, \Lambda^{-1}) ,
p(y|x) = \mathcal{N}(y | Ax + b, L^{-1}) ,
\]

the marginal distribution of \( y \) and the conditional distribution of \( x \) given \( y \) are given by

\[
p(y) = \mathcal{N}(y | A\mu + b. L^{-1} + A\Lambda^{-1} A^T) ,
p(x|y) = \mathcal{N}(x | \Sigma \{A^T L (y - b) + \Lambda \mu \}, \Sigma) ,
\]

where

\[
\Sigma = (\Lambda + A^T LA)^{-1} .
\]

Lower Bound

Consider

\[
\log p(y, X) = \log \int \int p(y, X, w, \alpha) dw d\alpha
= \log \int \int q(w, \alpha) \frac{p(y, X, w, \alpha)}{q(w, \alpha)} dw d\alpha \\
\ge \int \int q(w, \alpha) \log \left[ \frac{p(y, X, w, \alpha)}{q(w, \alpha)} \right] dw d\alpha \\
= \mathcal{F}(q) .
\]

Then the lower bound \( \mathcal{F}(q) \) is given by

\[
\mathcal{F}(q) = \int \int q(w, \alpha) \log p(y, X, w, \alpha) dw d\alpha \\
- \int \int q(w, \alpha) \log q(w, \alpha) dw d\alpha .
\]

Lower Bound (Cont’d)

Thus the lower bound \( \mathcal{F}(q) \) is given by

\[
\mathcal{F}(q) = \mathbb{E} \{ \log p(y, X, w, \alpha) \} - \mathbb{E} \{ \log q(w, \alpha) \} \\
= \mathbb{E}_w \{ \log p(y|X, w) \} + \mathbb{E}_{w, \alpha} \{ p(w|\alpha) \} + \mathbb{E}_\alpha \{ p(\alpha) \} \\
- \mathbb{E}_w \{ \log q(w) \} - \mathbb{E}_\alpha \{ \log q(\alpha) \} ,
\]

where the evaluation of each term is straightforward, making use of

\[
\mathbb{E}_\alpha \{ \alpha \} = \frac{a_N}{b_N} ,
\mathbb{E}_w \{ w w^T \} = m_N m_N^T + S_N .
\]
\[
E_w \{\log p(y|X, w)\} = \frac{N}{2} \log \left[ \frac{\beta}{2\pi} \right] - \frac{\beta}{2} y^\top y + \beta m_N^\top X^\top y \\
- \frac{\beta}{2} \text{tr} \left\{ X^\top X(m_N m_N^\top + S_N) \right\}, \\
E_{w, \alpha} \{p(w|\alpha)\} = -\frac{D}{2} \log 2\pi + \frac{D}{2} (\psi(a_N) - \log b_N) \\
- \frac{a_N}{2b_N} (m_N^\top m_N + \text{tr} \{S_N\}), \\
E_\alpha \{p(\alpha)\} = a_0 \log b_0 + (a_0 - 1) [\psi(a_N) - \log b_N] \\
- b_0 \frac{a_N}{b_N} - \log \Gamma(a_N), \\
E_w \{\log q(w)\} = -\frac{1}{2} \log |S_N| - \frac{D}{2} [1 + \log 2\pi], \\
E_\alpha \{\log q(\alpha)\} = - \log \Gamma(a_N) + (a_N - 1) \psi(a_N) + \log b_N - a_N.
\]