Regresión lineal con inferencia variacional

Mauricio A. Álvarez

Modelos probabilísticos profundos
AIR Institute

Contenido

Generalidades

- Los métodos variacionales tienen sus orígenes en el siglo 18 con el trabajo de Euler, Lagrange y otros, sobre el cálculo de variaciones.
- En el cálculo estándar se desea encontrar las derivadas de una función.
- Una función se puede entender como un mapeo que toma una variable de entrada y retorna el valor de la función como su salida.
- La derivada luego describe cómo cambia el valor de la salida con cambios infinitesimales del valor de la entrada.

Funcional

- Se puede definir un *funcional* como un mapeo que toma una función como su entrada y retorna el valor del funcional como su salida.

- Un ejemplo es la *entropía* $H(p)$, que toma una función de probabilidad $p(x)$ como la entrada y retorna la cantidad

$$H(p) = -\int p(x) \ln p(x) dx,$$

como la salida.

- Se puede introducir el concepto de *derivada funcional*, que expresa cómo cambia el valor del funcional en respuesta a cambios infinitesimales de la función de entrada.

Cálculo de variaciones (I)

- Las reglas para el cálculo de variaciones son muy parecidas a las reglas del cálculo convencional.
- Muchos problemas en ciencias e ingeniería, pueden expresarse en términos de un problema de optimización en el que la cantidad que se desea optimizar es un *funcional*.
- La solución se obtiene explorando todas las posibles funciones de entrada que maximizan o minimizan el funcional.
- Los métodos variacionales se emplean para encontrar soluciones aproximadas a este tipo de problemas de optimización.

Cálculo de variaciones (II)

- Esto se realiza restringiendo la clase de funciones sobre las cuales se realiza la optimización.
- Por ejemplo, considerando sólo funciones cuadráticas o considerando funciones compuestas por funciones base fijas.
- En inferencia probabilística, la restricción puede tomar la forma de una factorización.

Contenido

Modelo Bayesiano

- Supongamos que se tiene un modelo Bayesiano completo en el que a todos los parámetros se les ha asignado distribuciones prior.
- El modelo podría tener tanto variables latentes como parámetros, conjuntamente denotados como $\mathbf{Z}$.
- Similarmente, se denotan las variables observadas como $\mathbf{X}$.
- El modelo probabilístico especifica la función de distribución conjunta $p(\mathbf{X}, \mathbf{Z})$ y el objetivo es encontrar una aproximación a la distribución posterior $p(\mathbf{Z}|\mathbf{X})$ como a la evidencia del modelo $p(\mathbf{X})$.

Probabilidad marginal logarítmica

- La probabilidad marginal logarítmica $\ln p(\mathbf{X})$ se puede escribir como

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q\|p),$$

donde

$$\mathcal{L}(q) = \int_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} d\mathbf{Z}$$

$$\mathrm{KL}(q\|p) = -\int_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \right\} d\mathbf{Z}.$$

donde $q(\mathbf{Z})$ es la distribución desconocida.

- Se puede maximizar el límite inferior $\mathcal{L}(q)$ optimizando con respecto a la distribución $q(\mathbf{Z})$, que es equivalente a minimizar la divergencia de Kullback-Leibler (KL).

¿Cómo escoger $q(\mathbf{Z})$? (I)

- Si se permite cualquier forma para $q(\mathbf{Z})$, el máximo del límite inferior ocurre cuando la divergencia KL se hace cero, que a su vez ocurre cuando $q(\mathbf{Z})$ iguala a la distribución posterior $p(\mathbf{Z}|\mathbf{X})$.

- En la práctica, se asume que trabajar con la verdadera distribución posterior es intratable.

- En su lugar se considera una familia de distribuciones restringidas $q(\mathbf{Z})$, para las cuales se minimice la divergencia KL.

¿Cómo escoger $q(\mathbf{Z})$? (II)

- Una forma de restringir la familia de distribuciones es usar una distribución paramétrica $q(\mathbf{Z}|\omega)$, gobernada por un conjunto de parámetros ω.

- El límite inferior $\mathcal{L}(q)$ se vuelve entonces una función de ω, y se pueden explotar técnicas de optimización no lineal para determinar los valores óptimos de los parámetros.

Contenido

Mean field (I)

- Supongamos que los elementos de $\mathbf{Z}$ se dividen en grupos que no se traslapan, que se denotan como $\mathbf{Z}_i$, $i = 1, \ldots, M$.

- Luego se asume que la distribución $q(\mathbf{Z})$ se factoriza con respecto a estos grupos, tal que

$$q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i).$$

- Nótese que esta es la única restricción que se hace sobre $q(\mathbf{Z})$.

- Esta forma factorizada de inferencia variacional corresponde a una aproximación desarrollada en física conocida como “mean field theory”.

Mean field (II)

- Entre todas las distribuciones $q(\mathbf{Z})$ que tienen la forma factorizada anterior, se busca aquella distribución para la cual el límite inferior $\mathcal{L}(q)$ sea el mayor.

- Se desea realizar una optimización variacional de $\mathcal{L}(q)$, con respecto a todas las distribuciones $q_i(\mathbf{Z}_i)$, que se puede realizar optimizando con respecto a cada uno de los factores a la vez.

$\mathcal{L}(q)$ en función de $q_i(\mathbf{Z}_i)$

El límite inferior $\mathcal{L}(q)$ se puede escribir como

$$\begin{aligned}\mathcal{L}(q) &= \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})} \right\} d\mathbf{Z} \\ &= \int \left(\prod_{i=1}^{M} q_i(\mathbf{Z}_i) \right) \ln \left\{ \frac{p(\mathbf{X},\mathbf{Z})}{\prod_{i=1}^{M} q_i(\mathbf{Z}_i)} \right\} d\mathbf{Z} \\ &= \int \prod_{\forall i} q_i \left\{ \ln p(\mathbf{X},\mathbf{Z}) \right\} d\mathbf{Z} - \int \prod_{\forall i} q_i \sum_{\forall i} \ln q_i d\mathbf{Z}\end{aligned}$$

Distribuciones óptimas $q_j^*(\mathbf{Z}_j)$

- Se puede demostrar que la solución óptima $q_j^*(\mathbf{Z}_j)$ está dada como

$$\ln q_j^*(\mathbf{Z}_j) = \mathbb{E}_{i \neq j}\left[\ln p(\mathbf{X}, \mathbf{Z})\right] + \text{const},$$

  donde $\mathbb{E}_{i \neq j}$ se toma con respecto a todos los factores q_i con $i \neq j$.

- La constante aditiva const se selecciona de forma tal que normalice la distribución $q_j^*(\mathbf{Z}_j)$.

- Tomando la exponencial en ambos lados, en la expresión anterior, se tiene

$$q_j^*(\mathbf{Z}_j) = \frac{\exp\left(\mathbb{E}_{i \neq j}\left[\ln p(\mathbf{X}, \mathbf{Z})\right]\right)}{\int \exp\left(\mathbb{E}_{i \neq j}\left[\ln p(\mathbf{X}, \mathbf{Z})\right]\right) d\mathbf{Z}_j}.$$

- En la práctica, se prefiere trabajar con la expresión del $\ln q_j^*(\mathbf{Z}_j)$ y después normalizar.

Procedimiento

- El valor óptimo de $q_j^*(\mathbf{Z}_j)$ depende de los valores esperados calculados con respecto a los otros factores $q_i(\mathbf{Z}_i)$, para $i \neq j$.
- La solución se obtiene inicializando primero todos los factores $q_i(\mathbf{Z}_i)$, y luego iterando a través de cada factor.
- Cada factor se actualiza usando la expresión obtenida para $q_i^*(\mathbf{Z}_i)$.
- Para actualizar cualquier expresión, se usan los factores que ya se hayan actualizado hasta ese momento.

Contenido

Esquema completamente Bayesiano

- En el problema de inferencia Bayesiana anterior se hizo una estimación puntual de los parámetros α y β.
- En estimación Bayesiana completa, se debería poner distribuciones sobre estos parámetros y calcular la distribución posterior.
- Aunque la integración exacta no es posible, se puede encontrar una aproximación al posterior usando inferencia variacional.
- En lo que sigue, se asumirá que se conoce el valor de β.

Verosimilitud y prior

- Para el problema de regresión lineal Bayesiano anterior, la función de verosimilitud y el prior sobre $\mathbf{w}$ están dados como

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1})$$
$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}).$$

- Ahora se introduce un prior sobre α.

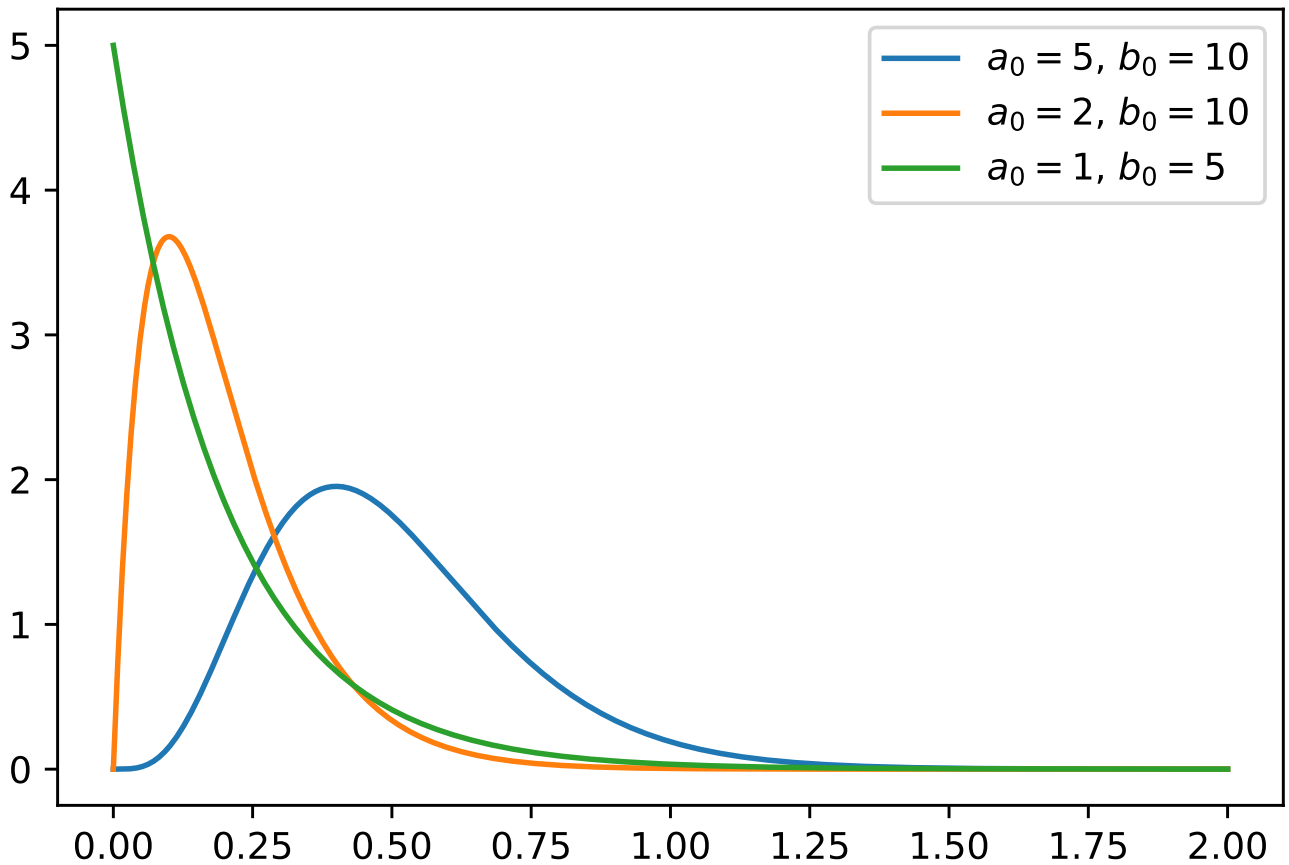
- Se puede demostrar que el prior conjugado a la precisión de una Gaussiana está dado por una distribución gamma,

$$p(\alpha) = \text{Gam}(\alpha|a_0, b_0).$$

Distribución gamma

La distribución gamma $p(\alpha) = \text{Gam}(\alpha|a_0, b_0)$ de parámetros $a_0 > 0$ y b_0 está dada como

$$p(\alpha) = \text{Gam}(\alpha|a_0, b_0) = \frac{1}{\Gamma(a_0)} b_0^{a_0} \alpha^{a_0 - 1} e^{-b_0 \alpha}.$$



Distribución variacional

- La distribución conjunta de todas las variables está dada como

$$p(\mathbf{t}, \mathbf{w}, \alpha) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}|\alpha)p(\alpha).$$

- El objetivo es encontrar una aproximación de la distribución posterior $p(\mathbf{w}, \alpha|\mathbf{t})$.

- Se usa el enfoque variacional con una distribución posterior que factoriza como

$$q(\mathbf{w}, \alpha) = q(\mathbf{w})q(\alpha).$$

Distribuciones óptimas

- Se puede demostrar que la distribución óptima $q^*(\alpha)$ está dada com

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

donde

$$a_N = a_0 + \frac{M}{2}, \qquad b_N = b_0 + \frac{1}{2}\mathbb{E}[\mathbf{w}^\top \mathbf{w}].$$

- De igual forma se puede demostrar que la distribución óptima $q^*(\mathbf{w})$ está dada como

$$q^*(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N),$$

donde

$$\mathbf{m}_N = \beta \mathbf{S}_N \mathbf{\Phi}^\top \mathbf{t}, \qquad \mathbf{S}_N^{-1} = \mathbb{E}[\alpha]\mathbf{I} + \beta \mathbf{\Phi}^\top \mathbf{\Phi}.$$

- En las expresiones anteriores,

$$\mathbb{E}[\mathbf{w}\mathbf{w}^\top] = \mathbf{m}_N \mathbf{m}_N^\top + \mathbf{S}_N$$

$$\mathbb{E}[\alpha] = \frac{a_N}{b_N}.$$

Cómo se compara con la solución Bayesiana anterior?

- Consideremos el caso para el cual $a_0 = 0, b_0 = 0$.
- Este caso corresponde a una distribución ancha para α.
- De esta manera

$$\mathbb{E}[\alpha] = \frac{a_N}{b_N} = \frac{M/2}{\mathbb{E}[\mathbf{w}^\top \mathbf{w}]/2} = \frac{M}{\mathbf{m}_N^\top \mathbf{m}_N + \text{trace}(\mathbf{S}_N)}.$$

Función predictiva

- La función predictiva está dada como

$$
\begin{aligned}
p(t|\mathbf{x},\mathbf{t}) &= \int p(t|\mathbf{x},\mathbf{w})p(\mathbf{w}|\mathbf{t})d\mathbf{w} \\
&\simeq \int p(t|\mathbf{x},\mathbf{w})q(\mathbf{w})d\mathbf{w} \\
&= \int \mathcal{N}(t|\mathbf{w}^\top\boldsymbol{\phi}(\mathbf{x}),\beta^{-1})\mathcal{N}(\mathbf{w}|\mathbf{m}_N,\mathbf{S}_N)d\mathbf{w} \\
&= \mathcal{N}(t|\mathbf{m}_N^\top\boldsymbol{\phi}(\mathbf{x}),\sigma^2(\mathbf{x})),
\end{aligned}
$$

donde

$$
\sigma^2(\mathbf{x}) = \beta^{-1} + \boldsymbol{\phi}^\top(\mathbf{x})\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}).
$$

- El límite inferior está dado como

$$
\begin{aligned}
\mathcal{L} &= \mathbb{E}[\ln p(\mathbf{w},\alpha,\mathbf{t})] - \mathbb{E}[\ln q(\mathbf{w},\alpha)] \\
&= \mathbb{E}_{q(\mathbf{w})}[\ln p(\mathbf{t}|\mathbf{w})] + \mathbb{E}_{q(\mathbf{w})q(\alpha)}[\ln p(\mathbf{w}|\alpha)] + \mathbb{E}_{q(\alpha)}[\ln p(\alpha)] \\
&- \mathbb{E}_{q(\mathbf{w})}[\ln q(\mathbf{w})] - \mathbb{E}_{q(\alpha)}[\ln q(\alpha)].
\end{aligned}
$$