AI4ER 0: Bayesian Linear Regression

Rich Turner
(with thanks to Miguel Hernandez Lobato for the slides)
Motivation

A large number of basis functions can lead to over-fitting of the maximum likelihood estimate: the model fits the training data well but it performs poorly on new test data.

Instead, favor smooth solutions by using Bayes rule with priors that enforce $w$ to be small.

Bayesian inference

Given data $\mathcal{D} = \{(\tilde{x}_n, y_n)\}_{n=1}^N$, we assume the linear regression model

$$y_n = w^T \tilde{x}_n + \epsilon_n, \quad \epsilon_n \sim \mathcal{N}(0, \sigma^2),$$

with unknown $w$. We assume $\sigma^2$ is known to simplify inference.

We also assume a prior distribution $p(w)$ on the model coefficients.

The posterior distribution for $w$ given $\mathcal{D}$ is obtained by Bayes rule:

$$p(w|y, \tilde{X}) = \frac{p(y|\tilde{X}, w)p(w)}{p(y|\tilde{X})}.$$  

The predictive distribution for $y_*$ given a new corresponding $x_*$ is

$$p(y_*|\tilde{x}_*, y, \tilde{X}) = \int p(y_*|\tilde{x}_*, w)p(w|y, \tilde{X}) \, dw.$$  

Exact inference is possible if prior and noise distributions are Gaussian.
Multivariate Gaussian distribution

The density of a $D$-dimensional vector $\mathbf{x}$ is

$$
\mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{V}) = \frac{1}{\sqrt{(2\pi)^D|\mathbf{V}|}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mathbf{m})^T \mathbf{V}^{-1} (\mathbf{x} - \mathbf{m}) \right\}.
$$

The density is proportional to the exponential of a quadratic function of $\mathbf{x}$:

$$
\mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{V}) = \frac{1}{\sqrt{(2\pi)^D|\mathbf{V}|}} \exp \left\{ -\frac{1}{2} \mathbf{x}^T \mathbf{V}^{-1} \mathbf{x} + \mathbf{m}^T \mathbf{V}^{-1} \mathbf{x} - \frac{1}{2} \mathbf{m}^T \mathbf{V}^{-1} \mathbf{m} \right\}
\propto \exp \left\{ -\frac{1}{2} \mathbf{x}^T \mathbf{V}^{-1} \mathbf{x} + \mathbf{m}^T \mathbf{V}^{-1} \mathbf{x} \right\},
$$

(1)

with normalization constant $\sqrt{(2\pi)^D|\mathbf{V}|} \exp\{1/2\mathbf{m}^T \mathbf{V}^{-1} \mathbf{m}\}$.

$\mathbf{m}$ is the $D$-dimensional mean vector and $\mathbf{V}$ is the $D \times D$ covariance matrix:

$$
\mathbf{m} = \mathbb{E}[\mathbf{x}], \quad \mathbf{V} = \mathbb{E}[(\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^T] = \mathbb{E}[\mathbf{x}\mathbf{x}^T] - \mathbf{m}\mathbf{m}^T.
$$

$\hat{\mathbf{x}} \sim \mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{V})$ can be obtained using the matrix square root $\mathbf{V} = \mathbf{V}^{1/2}(\mathbf{V}^{1/2})^T$:

$$
\hat{\mathbf{x}} = \mathbf{m} + \mathbf{V}^{1/2} \mathbf{z},
$$

$\mathbf{z} \sim \mathcal{N}(0, I)$. 

Effect of parameters on density function

Assuming

\[ p(x) = \mathcal{N}(x | \mathbf{m}, \mathbf{V}) = \frac{1}{\sqrt{(2\pi)^D|\mathbf{V}|}} \exp \left\{ -\frac{1}{2}(x - \mathbf{m})^T \mathbf{V}^{-1} (x - \mathbf{m}) \right\}. \]

The parameter \( \mathbf{m} \) determines **mode location** and \( \mathbf{V} \) **scales and rotates** the space.

What can you say about

\[
\mathbf{V} = \begin{bmatrix} v_1 & \text{cov} \\ \text{cov} & v_2 \end{bmatrix}
\]

given the following contour plots of \( p(x) \)?
Effect of parameters on density function

Assuming

\[ p(x) = \mathcal{N}(x|m, V) = \frac{1}{\sqrt{(2\pi)^D|V|}} \exp \left\{ -\frac{1}{2} (x - m)^T V^{-1} (x - m) \right\} \, . \]

The parameter \( m \) determines mode location and \( V \) scales and rotates the space.

What can you say about \( V = \begin{bmatrix} v_1 & \text{cov} \\ \text{cov} & v_2 \end{bmatrix} \) given the following contour plots of \( p(x) \)?

\[ v_1 = v_2, \, \text{cov} = 0. \]
Effect of parameters on density function

Assuming

\[ p(x) = \mathcal{N}(x|\mathbf{m}, \mathbf{V}) = \frac{1}{\sqrt{(2\pi)^D|\mathbf{V}|}} \exp \left\{ -\frac{1}{2} (x - \mathbf{m})^T \mathbf{V}^{-1} (x - \mathbf{m}) \right\} \]

The parameter \( \mathbf{m} \) determines **mode location** and \( \mathbf{V} \) **scales and rotates** the space.

What can you say about \( \mathbf{V} = \begin{bmatrix} \nu_1 & \text{cov} \\ \text{cov} & \nu_2 \end{bmatrix} \)?

given the following contour plots of \( p(x) \)?

---

\( \nu_1 = \nu_2, \text{cov} = 0. \)  
\( \nu_1 > \nu_2, \text{cov} = 0. \)
Effect of parameters on density function

Assuming
\[ p(x) = \mathcal{N}(x|m, V) = \frac{1}{\sqrt{(2\pi)^D|V|}} \exp \left\{ -\frac{1}{2} (x - m)^T V^{-1} (x - m) \right\}. \]

The parameter \( m \) determines mode location and \( V \) scales and rotates the space.

What can you say about \( V = \begin{bmatrix} v_1 & \text{cov} \\ \text{cov} & v_2 \end{bmatrix} \) given the following contour plots of \( p(x) \)?

\begin{align*}
\text{Case 1: } & v_1 = v_2, \text{ cov } = 0. \\
\text{Case 2: } & v_1 > v_2, \text{ cov } = 0. \\
\text{Case 3: } & \text{cov} > 0.
\end{align*}
Match each exponential of quadratic function with its contour plot.

i) \( \exp \left\{ -\frac{1}{2} x_1^2 a - \frac{1}{2} x_2^2 a \right\} \),

ii) \( \exp \left\{ -\frac{1}{2} x_1^2 a - \frac{1}{2} x_2^2 b + x_1 x_2 c \right\} \),

iii) \( \exp \left\{ -\frac{1}{2} (y - x_1 - x_2)^2 \right\} \),
Linear combination of Gaussian random variables

Let \( p(x) = \mathcal{N}(x|0, V_1) \) and \( p(e) = \mathcal{N}(x|0, V_2) \) and assume that, for a matrix \( W \),

\[
y = Wx + e.
\]

What is \( p(y) \)? Linear combinations of Gaussian random variables are Gaussian. Therefore, \( p(y) \) is Gaussian with mean vector

\[
m_3 = E[y] = E[Wx + e] = WE[x] + E[e] = 0
\]

and covariance matrix

\[
V_3 = E[yy^T] - E[y]E[y]^T
\]

\[
= E[yy^T]
\]

\[
= E[(Wx + e)(Wx + e)^T]
\]

\[
= E[Wxx^TW^T + ex^TW^T + Wxe^T + ee^T]
\]

\[
= WV_1W^T + V_2.
\]

What if \( p(x) = \mathcal{N}(x|m_1, V_1) \)?
Completing the square

Let

\[ p(x) = \mathcal{N}(x|m, V) = \frac{1}{\sqrt{(2\pi)^D|V|}} \exp \left\{ -\frac{1}{2}(x - m)^T V^{-1}(x - m) \right\}, \]

\[ q(x) = \exp \left\{ -\frac{1}{2}x^T P x + a^T x \right\}. \]

Assume that \( p(x) \propto q(x) \).

Write \( m \) and \( V \) in terms of \( P \) and \( a \).

We have that

\[ p(x) \propto \exp \left\{ -\frac{1}{2}x^T V^{-1} x + m^T V^{-1} x \right\} \quad \text{from equation (1)}, \]

\[ q(x) = \exp \left\{ -\frac{1}{2}x^T P x + a^T x \right\}. \]

Therefore, \( V = P^{-1} \) and \( m = Va \).

What is the normalization constant of \( q(x) \)?
Product of Gaussian densities

Let \( p(x) = \mathcal{N}(x|m_1, V_1) \) and \( q(x) = \mathcal{N}(x|m_2, V_2) \). What is \( t(x) \propto p(x)q(x) \)?

t(x) is Gaussian \( \mathcal{N}(x|m_3, V_3) \) because the product of exponentials of quadratic functions is also the exponential of a quadratic function. What are \( m_3 \) and \( V_3 \)?

\[
p(x)q(x) = \mathcal{N}(x|m_1, V_1)\mathcal{N}(x|m_2, V_2)
\]

\[
= \frac{1}{\sqrt{(2\pi)^D |V_1|}} \exp \left\{ -\frac{1}{2} x^T V_1^{-1} x + m_1^T V_1^{-1} x - \frac{1}{2} m_1^T V_1^{-1} m_1 \right\}
\]

\[
= \frac{1}{\sqrt{(2\pi)^D |V_2|}} \exp \left\{ -\frac{1}{2} x^T V_2^{-1} x + m_2^T V_2^{-1} x - \frac{1}{2} m_2^T V_2^{-1} m_2 \right\}
\]

\[
\propto \exp \left\{ -\frac{1}{2} x^T \begin{pmatrix} V_1^{-1} + V_2^{-1} \end{pmatrix} x + \left( m_1^T V_1^{-1} + m_2^T V_2^{-1} \right) x \right\}.
\]

Therefore, \( V_3 = (V_1^{-1} + V_2^{-1})^{-1} \) and \( m_3 = V_3 (m_1^T V_1^{-1} + m_2^T V_2^{-1})^T \).

What is the normalization constant of \( p(x)q(x) \)?
Bayesian linear regression

Consider a regression model in which $\sigma^2$ is known: the only unknown is $\mathbf{w}$.

Recall that the likelihood function under Gaussian noise is

$$p(\mathbf{y}|\tilde{\mathbf{X}}, \mathbf{w}) = \prod_{n=1}^{N} \mathcal{N}(y_n|\mathbf{w}^T \tilde{x}_n, \sigma^2) \propto \exp \left\{ -\frac{\mathbf{w}^T \tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \mathbf{w}}{2\sigma^2} + \frac{\mathbf{y}^T \tilde{\mathbf{X}} \mathbf{w}}{\sigma^2} \right\}.$$ 

We choose the prior for $\mathbf{w}$ to be a zero-mean isotropic Gaussian:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \lambda^{-1} \mathbf{I}) \propto \exp \left\{ -\frac{1}{2} \mathbf{x}^T \lambda \mathbf{l} \mathbf{x} \right\}.$$ 

The posterior is then Gaussian:

$$p(\mathbf{w}|\mathbf{y}, \tilde{\mathbf{X}}, \sigma^2) \propto p(\mathbf{y}|\tilde{\mathbf{X}}, \mathbf{w}) p(\mathbf{w})$$

$$\propto \exp \left\{ -\frac{1}{2} \mathbf{w}^T \left( \tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \sigma^{-2} + \lambda \mathbf{I} \right) \mathbf{w} + \frac{\mathbf{y}^T \tilde{\mathbf{X}} \sigma^{-2}}{\mathbf{m}^T \mathbf{V}^{-1}} \mathbf{w} \right\}.$$ 

Therefore, $p(\mathbf{w}|\mathbf{y}, \tilde{\mathbf{X}}, \sigma^2) = \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{V})$ where

$$\mathbf{V} = \left( \tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \sigma^{-2} + \lambda \mathbf{I} \right)^{-1}, \quad \mathbf{m} = \mathbf{V} \sigma^{-2} \tilde{\mathbf{X}}^T \mathbf{y}.$$
The Bayesian predictive distribution

The predictive distribution for the $y_*$ of a given new corresponding $\tilde{x}_*$ is

$$p(y_*|\tilde{x}_*, y, \tilde{X}) = \int p(y_*|\tilde{x}_*, w)p(w|y, \tilde{X})dw = \int N(y_*|w^T\tilde{x}_*, \sigma^2)N(w|m, V)dw.$$  

We have that $y_* = w^T\tilde{x}_* + e_*$, where $w \sim N(m, V)$ and $e_* \sim N(0, \sigma^2)$. Thus,

$$p(y_*|\tilde{x}_*, y, \tilde{X}) = N(y_*|m_*, v_*),$$

where $m_* = m^T\tilde{x}_*$ and $v_* = \tilde{x}_*^TV\tilde{x}_* + \sigma^2$.

Example with polynomial basis functions, $M = 10$, $\lambda = 10^{-5}$, $\sigma^2 = 0.005$:

We reduce overfitting and obtain confidence bands $m_* \pm v_*^{1/2}$ in our predictions!
Example

Prior: $p(w)$

Likelihood: $p(y_1 | \tilde{x}_1, w)$

Posterior: $p(w | y_1, \tilde{x}_1)$

Predictions of posterior samples
Another example with Gaussian basis functions

Functions for posterior samples of $\mathbf{w}$

Predictive distribution

Figure: C. Bishop. *Pattern Recognition and Machine Learning*, 2006.
Maximum a posteriori (MAP) inference

Assumes that the posterior is well approximated by a point mass at its mode:

\[
p(\theta | \mathcal{D}) \approx p(\mathbf{y} | \mathbf{e}_x, y, \mathbf{e}_X) = \int p(\mathbf{y} | \mathbf{e}_x, \mathbf{w}) p(\mathbf{w} | y, \mathbf{e}_X) d\mathbf{w} \]

\[
\mathbf{w}_{\text{MAP}} = \arg \max_{\mathbf{w}} p(\mathbf{w} | y, \mathbf{e}_X)
\]

In particular,

\[
p(y_* | \tilde{x}_*, y, \tilde{X}) = \int p(y_* | \tilde{x}_*, \mathbf{w}) p(\mathbf{w} | y, \mathbf{e}_X) d\mathbf{w} \quad \mathbf{w}_{\text{MAP}} = \arg \max_{\mathbf{w}} p(\mathbf{w} | y, \tilde{X})
\]

\[
p(y_* | \tilde{x}_*, y, \tilde{X}) \approx \int p(y_* | \tilde{x}_*, \mathbf{w}) \delta(\mathbf{w} - \mathbf{w}_{\text{MAP}}) d\mathbf{w} \quad = \arg \max_{\mathbf{w}} p(y | \mathbf{w}, \tilde{X}) p(\mathbf{w})
\]

\[
\approx p(y_* | \tilde{x}_*, \mathbf{w}_{\text{MAP}}), \quad = \arg \max_{\mathbf{w}} \left\{ \log p(y | \mathbf{w}, \tilde{X}) + \log p(\mathbf{w}) \right\}.
\]

MAP inference is a form of regularized MLE. For \( p(\mathbf{w}) = \mathcal{N}(\mathbf{w} | 0, \lambda^{-1} \mathbf{I}) \), we obtain

\[
\mathbf{w}_{\text{MAP}} = \arg \max_{\mathbf{w}} \left\{ \log p(\mathbf{y} | \mathbf{w}, \tilde{X}) - \frac{\lambda}{2} \mathbf{w}^T \mathbf{w} \right\} = (\tilde{X}^T \tilde{X} \sigma^{-2} + \lambda \mathbf{I})^{-1} \sigma^{-2} \tilde{X}^T \mathbf{y}.
\]

MAP inference fails to generate confidence bands in the resulting predictions!