Bayesian Regression

David Rosenberg

New York University

October 29, 2016
Bayesian Statistics: Recap
The Bayesian Method

1. Define the model:
   - Choose a probability model or “likelihood model”:
     \[ \{ p(\mathcal{D} | \theta) \mid \theta \in \Theta \}. \]
   - Choose a distribution \( p(\theta) \), called the prior distribution.

2. After observing \( \mathcal{D} \), compute the posterior distribution \( p(\theta | \mathcal{D}) \).

3. Choose action based on \( p(\theta | \mathcal{D}) \).
   - e.g. \( \mathbb{E}[\theta | \mathcal{D}] \) as point estimate for \( \theta \)
   - e.g. interval \([a, b]\), where \( p(\theta \in [a, b] | \mathcal{D}) = 0.95 \)
The Posterior Distribution

- By Bayes rule, can write the posterior distribution as

\[ p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{p(D)} . \]

- **likelihood:** \( p(D \mid \theta) \)
- **prior:** \( p(\theta) \)
- **marginal likelihood:** \( p(D) \).

- Note: \( p(D) \) is just a normalizing constant for \( p(\theta \mid D) \). Can write

\[ p(\theta \mid D) \propto p(D \mid \theta)p(\theta) . \]
Prior represents belief about $\theta$ before observing data $\mathcal{D}$.

Posterior represents the **rationally “updated” beliefs** after seeing $\mathcal{D}$.

All inferences and action-taking are based on posterior distribution.
Bayesian Gaussian Linear Regression
Bayesian Conditional Models

- Input space $\mathcal{X} = \mathbb{R}^d$  
  Output space $\mathcal{Y} = \mathbb{R}$

- **Conditional probability model, or likelihood model:**

$$\{p(y \mid x, \theta) \mid \theta \in \Theta\}$$

- Conditional here refers to the conditioning on the input $x$.
  - $x$’s are not governed by our probability model.
  - Everything conditioned on $x$ means “$x$ is known”

- **Prior distribution:** $p(\theta)$ on $\theta \in \Theta$
Gaussian Regression Model

- Input space $\mathcal{X} = \mathbb{R}^d$  
  Output space $\mathcal{Y} = \mathbb{R}$

- Conditional probability model, or likelihood model:
  $y \mid x, w \sim \mathcal{N}(w^T x, \sigma^2)$,
  for some known $\sigma^2 > 0$.

- Parameter space? $\mathbb{R}^d$.

- Data: $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
  - Notation: $y = (y_1, \ldots, y_n)$ and $x = (x_1, \ldots, x_n)$.
  - Assume $y_i$’s are conditionally independent, given $x$ and $w$. 
Gaussian Likelihood

- The likelihood of \( w \in \mathbb{R}^d \) for the data \( \mathcal{D} \) is

\[
p(y \mid x, w) = \prod_{i=1}^{n} p(y_i \mid x_i, w) \quad \text{by conditional independence.}
\]

\[
= \prod_{i=1}^{n} \left[ \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{(y_i - w^T x_i)^2}{2\sigma^2} \right) \right]
\]

- You should see in your head\(^1\) that the MLE is

\[
w_{\text{MLE}}^* = \arg \max_{w \in \mathbb{R}^d} p(y \mid x, w)
\]

\[
= \arg \min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} (y_i - w^T x_i)^2.
\]

---

Choose a Gaussian prior distribution \( p(w) \) on \( \mathbb{R}^d \):

\[
w \sim \mathcal{N}(0, \Sigma_0)
\]

for some covariance matrix \( \Sigma_0 \succ 0 \) (i.e. \( \Sigma_0 \) is spd).

Posterior distribution

\[
p(w \mid D) = p(w \mid x, y) \\
= p(y \mid x, w) p(w) / p(y) \\
\propto p(y \mid x, w) p(w)
\]

\[
= \prod_{i=1}^{n} \left[ \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{(y_i - w^T x_i)^2}{2\sigma^2} \right) \right] \quad \text{(likelihood)}
\]

\[
\times |2\pi \Sigma_0|^{-1/2} \exp \left( -\frac{1}{2} w^T \Sigma_0^{-1} w \right) \quad \text{(prior)}
\]
What does the posterior give us?

- **Likelihood model**: \( y \mid x, w \sim \mathcal{N}(w^T x, \sigma^2) \)
- **Prior distribution**: \( w \sim \mathcal{N}(0, \Sigma_0) \)
- Given data, compute **posterior distribution**: \( p(w \mid \mathcal{D}) \).
- If we knew \( w \), best prediction function (for square loss) is
  \[
  \hat{y}(x) = \mathbb{E}[y \mid x, w] = w^T x.
  \]

Prior \( p(w) \) and posterior \( p(w \mid \mathcal{D}) \)
  - give **distributions over prediction functions**!
Gaussian Regression Example
Example in 1-Dimension: Setup

- Input space \( X = [-1, 1] \)  
- Output space \( Y = \mathbb{R} \)
- Let’s suppose for any \( x \), \( y \) is generated as
  \[ y = w_0 + w_1 x + \epsilon, \]
  where \( \epsilon \sim \mathcal{N}(0, 0.2^2) \).
- Written another way, the likelihood model is
  \[ y \mid x, w_0, w_1 \sim \mathcal{N}(w_0 + w_1 x, 0.2^2). \]
- What’s the parameter space? \( \mathbb{R}^2 \).
- Prior distribution: \( w = (w_0, w_1) \sim \mathcal{N}(0, \frac{1}{2} I) \)
Example in 1-Dimension: Prior Situation

- **Prior distribution**: \( w = (w_0, w_1) \sim \mathcal{N}(0, \frac{1}{2}I) \)

On right, \( y(x) = \mathbb{E}[y \mid x, w] \), for randomly chosen \( w \sim \mathcal{N}(0, \frac{1}{2}I) \).

- \( y(x) = w_0 + w_1x \) for random \( (w_0, w_1) \sim p(\theta) = \mathcal{N}(0, \frac{1}{2}I) \).

---

Bishop's PRML Fig 3.7
Example in 1-Dimension: 1 Observation

- On left: posterior distribution; white '+' indicates true parameters
- On right: blue circle indicates the training observation

Bishop’s PRML Fig 3.7
Gaussian Regression Example

Example in 1-Dimension: 2 and 20 Observations

Bishop’s PRML Fig 3.7
Gaussian Regression Continued
Closed Form for Posterior

- **Model:**
  \[ w \sim \mathcal{N}(0, \Sigma_0) \]
  \[ y_i \mid x, w \text{ i.i.d. } \mathcal{N}(w^T x_i, \sigma^2) \]

- **Design matrix** \( X \); **Response column vector** \( y \)
- **Posterior distribution is a Gaussian distribution:**
  \[
  w \mid D \sim \mathcal{N}(\mu_P, \Sigma_P)
  \]
  \[
  \mu_P = (X^T X + \sigma^2 \Sigma_0^{-1})^{-1} X^T y
  \]
  \[
  \Sigma_P = (\sigma^{-2} X^T X + \Sigma_0^{-1})^{-1}
  \]

- **Posterior Variance** \( \Sigma_P \) gives us a natural uncertainty measure.

http://www.gaussianprocess.org/gpml/chapters/RW2.pdf
Closed Form for Posterior

- Posterior distribution is a Gaussian distribution:
  \( w | \mathcal{D} \sim \mathcal{N}(\mu_p, \Sigma_p) \)
  \[
  \mu_p = (X^T X + \sigma^2 \Sigma_0^{-1})^{-1} X^T y \\
  \Sigma_p = (\sigma^{-2} X^T X + \Sigma_0^{-1})^{-1}
  \]

- The MAP estimator and the posterior mean are given by
  \[
  \mu_p = (X^T X + \sigma^2 \Sigma_0^{-1})^{-1} X^T y
  \]

- For the prior variance \( \Sigma_0 = \frac{\sigma^2}{\lambda} I \), we get
  \[
  \mu_p = (X^T X + \lambda I)^{-1} X^T y,
  \]
  which is of course the ridge regression solution.
Posterior Mean and Posterior Mode (MAP)

- Posterior density for $\Sigma_0 = \frac{\sigma^2}{\lambda} I$:

$$p(w | D) \propto \exp \left( -\frac{\lambda}{2\sigma^2} \|w\|^2 \right) \prod_{i=1}^{n} \exp \left( -\frac{(y_i - w^T x_i)^2}{2\sigma^2} \right)$$

- To find MAP, sufficient to minimize the negative log posterior:

$$\hat{w}_{\text{MAP}} = \arg\min_{w \in \mathbb{R}^d} \left[ -\log p(w | D) \right]$$

$$= \arg\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} \left( y_i - w^T x_i \right)^2 + \lambda \|w\|^2$$

- Which is the ridge regression objective.
Predictive Distribution

- Given a new input point $x_{\text{new}}$, how to predict $y_{\text{new}}$?

**Predictive distribution**

$$p(y_{\text{new}} \mid x_{\text{new}}, D) = \int p(y_{\text{new}} \mid x_{\text{new}}, \theta, D) p(\theta \mid D) \, d\theta$$

$$= \int p(y_{\text{new}} \mid x_{\text{new}}, \theta) p(\theta \mid D) \, d\theta$$

- For Gaussian regression, predictive distribution has closed form.
Closed Form for Predictive Distribution

- **Model:**
  \[ w \sim \mathcal{N}(0, \Sigma_0) \]
  \[ y_i \mid x, w \text{ i.i.d. } \mathcal{N}(w^T x_i, \sigma^2) \]

- **Predictive Distribution**
  \[ p(y_{\text{new}} \mid x_{\text{new}}, \mathcal{D}) = \int p(y_{\text{new}} \mid x_{\text{new}}, w) p(w \mid \mathcal{D}) d\theta. \]
  - Averages over prediction for each \( \theta \), weighted by posterior distribution.

- **Closed form:**
  \[ y_{\text{new}} \mid x_{\text{new}}, \mathcal{D} \sim \mathcal{N}(\eta_{\text{new}}, \sigma_{\text{new}}) \]
  \[ \eta_{\text{new}} = \mu_P x_{\text{new}} \]
  \[ \sigma_{\text{new}} = \sqrt{x_{\text{new}}^T \Sigma_P x_{\text{new}} + \sigma^2} \]
  - from variance in \( \theta \)
  - inherent variance in \( y \)
Predictive Distributions

With predictive distributions, can give mean prediction with error bands:

Rasmussen and Williams’ *Gaussian Processes for Machine Learning*, Fig. 2.1(b)