Bayesian Regression

David Rosenberg

New York University

April 11, 2017

Bayesian Statistics: Recap

The Bayesian Method

1. Define the model:
   - Choose a probability model or "**likelihood model**":
   $$\{p(\mathcal{D} \mid \theta) \mid \theta \in \Theta\}.$$
   - Choose a distribution $p(\theta)$, called the **prior distribution**.
2. After observing $\mathcal{D}$, compute the **posterior distribution** $p(\theta \mid \mathcal{D})$.
3. Choose **action** based on $p(\theta \mid \mathcal{D})$.
   - e.g. $\mathbb{E}[\theta \mid \mathcal{D}]$ as point estimate for θ
   - e.g. interval $[a, b]$, where $p(\theta \in [a, b] \mid \mathcal{D}) = 0.95$

The Posterior Distribution

- By Bayes rule, can write the posterior distribution as

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)p(\theta)}{p(\mathcal{D})}.$$

- **likelihood:** $p(\mathcal{D} \mid \theta)$
- **prior:** $p(\theta)$
- **marginal likelihood**: $p(\mathcal{D})$.
- Note: $p(\mathcal{D})$ is just a normalizing constant for $p(\theta \mid \mathcal{D})$. Can write

$$\underbrace{p(\theta \mid \mathcal{D})}_{\text{posterior}} \propto \underbrace{p(\mathcal{D} \mid \theta)}_{\text{likelihood}} \underbrace{p(\theta)}_{\text{prior}}.$$

Summary

- Prior represents belief about θ 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 $\mathcal{X} = \mathbf{R}^d$ Output space $\mathcal{Y} = \mathbf{R}$
- **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 $\mathcal{X} = \mathbf{R}^d$ Output space $\mathcal{Y} = \mathbf{R}$
- **Conditional probability model**, or **likelihood model**:

$$y \mid x, w \quad \sim \quad \mathcal{N}\left(w^T x, \sigma^2\right),$$

  for some known $\sigma^2 > 0$.
- **Parameter space?** $\mathbf{R}^d$.
- **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.

Conditional Independence (Review)

Definition

We say W and S are **conditionally independent** given R, denoted

$$W \perp S \mid R,$$

if the conditional joint factorizes as

$$p(w, s \mid r) = p(w \mid r)p(s \mid r).$$

Also holds when W, S, and R represent **sets of random variables**.

- Can have conditional independence without independence.
- Can have independence without conditional independence.

Gaussian Likelihood and MLE

- The **likelihood** of $w \in \mathbf{R}^d$ for the data $\mathcal{D}$ is

$$
\begin{aligned}
p(y \mid x, w) &= \prod_{i=1}^{n} p(y_i \mid x_i, w) \qquad \text{by conditional independence.} \\
&= \prod_{i=1}^{n} \left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2} \right) \right]
\end{aligned}
$$

- You should see **in your head**[1] that the **MLE** is

$$
\begin{aligned}
w^*_{\text{MLE}} &= \underset{w \in \mathbf{R}^d}{\arg\max}\, p(y \mid x, w) \\
&= \underset{w \in \mathbf{R}^d}{\arg\min} \sum_{i=1}^{n} (y_i - w^T x_i)^2.
\end{aligned}
$$

[1] See `https://davidrosenberg.github.io/ml2015/docs/8.Lab.glm.pdf`, slide 5.

Priors and Posteriors

- Choose a Gaussian **prior distribution** $p(w)$ on $\mathbf{R}^d$:

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

  for some **covariance matrix** $\Sigma_0 \succ 0$ (i.e. Σ_0 is spd).

- **Posterior distribution**

$$\begin{aligned}
p(w \mid \mathcal{D}) &= p(w \mid x, y) \\
&= p(y \mid x, w)\, p(w) / p(y \mid x) \\
&\propto p(y \mid x, w) p(w) \\
&= \prod_{i=1}^{n} \left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2} \right) \right] \quad \textbf{(likelihood)} \\
&\quad \times |2\pi\Sigma_0|^{-1/2} \exp\left(-\frac{1}{2} w^T \Sigma_0^{-1} w) \right) \quad \textbf{(prior)}
\end{aligned}$$

Predictive Distributions

- **Likelihood model**: $y \mid x, w \sim \mathcal{N}\left(w^T x, \sigma^2\right)$
- If we knew w, best prediction function (for square loss) is

$$\hat{y}(x) = \mathbb{E}\left[y \mid x, w\right] = w^T x.$$

- In Bayesian statistics we have
  - **Prior distribution:** $w \sim \mathcal{N}(0, \Sigma_0)$, and
  - Given data, we can compute the **posterior distribution**: $p(w \mid \mathcal{D})$.
- 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 $\mathcal{X} = [-1, 1]$ Output space $\mathcal{Y} = \mathbf{R}$
- Given x, the world generates y as

$$y = w_0 + w_1 x + \varepsilon,$$

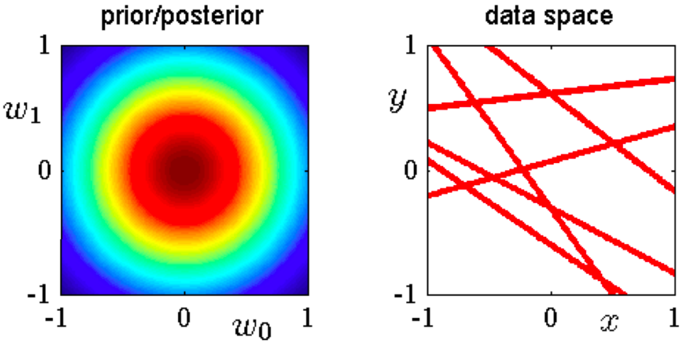
  where $\varepsilon \sim \mathcal{N}(0, 0.2^2)$.
- Written another way, the **likelihood model** is

$$y \mid x, w_0, w_1 \quad \sim \quad \mathcal{N}\left(w_0 + w_1 x\,,\, 0.2^2\right).$$

- What's the parameter space? $\mathbf{R}^2$.
- **Prior distribution:** $w = (w_0, w_1) \sim \mathcal{N}\left(0, \frac{1}{2} I\right)$

Example in 1-Dimension: Prior Situation

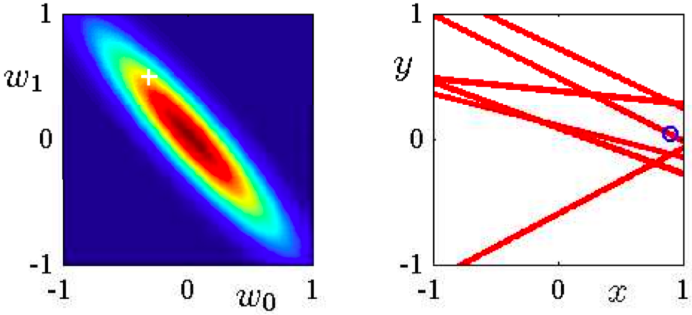
- **Prior distribution:** $w = (w_0, w_1) \sim \mathcal{N}\left(0, \frac{1}{2}I\right)$ (Illustrated on left)



- On right, $y(x) = \mathbb{E}[y \mid x, w] = w_0 + w_1 x$, for randomly chosen $w \sim p(w) = \mathcal{N}\left(0, \frac{1}{2}I\right)$.

Bishop's PRML Fig 3.7

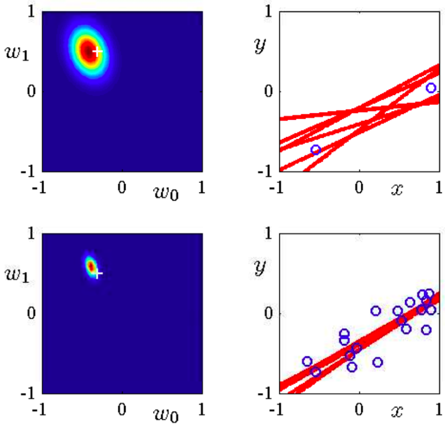
Example in 1-Dimension: 1 Observation



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

Bishop's PRML Fig 3.7

Example in 1-Dimension: 2 and 20 Observations



Bishop's PRML Fig 3.7

Gaussian Regression Continued

Closed Form for Posterior

- Model:

$$\begin{aligned} w &\sim \mathcal{N}(0, \Sigma_0) \\ y_i \mid x, w &\text{ i.i.d. } \mathcal{N}(w^T x_i, \sigma^2) \end{aligned}$$

- Design matrix X; Response column vector y
- **Posterior distribution is a Gaussian distribution:**

$$\begin{aligned} w \mid \mathcal{D} &\sim \mathcal{N}(\mu_P, \Sigma_P) \\ \mu_{\mathbf{P}} &= \left(X^T X + \sigma^2 \Sigma_0^{-1}\right)^{-1} X^T y \\ \Sigma_{\mathbf{P}} &= \left(\sigma^{-2} X^T X + \Sigma_0^{-1}\right)^{-1} \end{aligned}$$

- **Posterior Variance** Σ_P gives us a natural **uncertainty measure.**

See Rasmussen and Williams' *Gaussian Processes for Machine Learning*, Ch 2.1. http://www.gaussianprocess.org/gpml/chapters/RW2.pdf

Closed Form for Posterior

- **Posterior distribution is a Gaussian distribution:**

$$\begin{aligned} w \mid \mathcal{D} &\sim \mathcal{N}(\mu_P, \Sigma_P) \\ \mu_{\mathbf{P}} &= \left(X^T X + \sigma^2 \Sigma_0^{-1}\right)^{-1} X^T y \\ \Sigma_{\mathbf{P}} &= \left(\sigma^{-2} X^T X + \Sigma_0^{-1}\right)^{-1} \end{aligned}$$

- The **MAP estimator** and the **posterior mean** are given by

$$\mu_P = \left(X^T X + \sigma^2 \Sigma_0^{-1}\right)^{-1} X^T y$$

- For the prior variance $\Sigma_0 = \frac{\sigma^2}{\lambda} I$, we get

$$\mu_P = \left(X^T X + \lambda I\right)^{-1} X^T y,$$

which is of course the ridge regression solution.

Posterior Variance vs. Traditional Uncertainty

- Traditional regression: OLS estimator (also the MLE) is a random variable – why?
  - Because estimator is a function of data $\mathcal{D}$ and data is random.
- Common assumption: data are iid with Gaussian noise: $y = w^T x + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.
- Then OLS estimator $\hat{w}$ has a **sampling distribution** that is Gaussian with mean w and

$$\mathrm{Cov}(\hat{w}) = \left(\sigma^{-2} X^T X\right)^{-1}$$

- By comparison, the posterior variance is

$$\Sigma_P = \left(\sigma^{-2} X^T X + \Sigma_0^{-1}\right)^{-1}.$$

- When we take $\Sigma_0^{-1} = 0$, we get back $\mathrm{Cov}(\hat{\theta})$ (i.e. like our prior variance goes to ∞.)
- Σ_P is "smaller" than $\mathrm{Cov}(\hat{w})$ because we're using a "more informative" prior.

Posterior Mean and Posterior Mode (MAP)

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

$$p(w \mid \mathcal{D}) \quad \propto \quad \underbrace{\exp\left(-\frac{\lambda}{2\sigma^2}\|w\|^2\right)}_{\textbf{prior}} \underbrace{\prod_{i=1}^{n} \exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2}\right)}_{\textbf{likelihood}}$$

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

$$\begin{aligned} \hat{w}_{\text{MAP}} &= \underset{w \in \mathbf{R}^d}{\arg\min} \left[-\log p(w \mid \mathcal{D})\right] \\ &= \underset{w \in \mathbf{R}^d}{\arg\min} \underbrace{\sum_{i=1}^{n} (y_i - w^T x_i)^2}_{\text{log-likelihood}} + \underbrace{\lambda \|w\|^2}_{\text{log-prior}} \end{aligned}$$

- Which is the ridge regression objective.

Predictive Distribution

- Given a new input point x_{new}, how to predict y_{new} ?
- **Predictive distribution**

$$
\begin{aligned}
p(y_{\textbf{new}} \mid x_{\textbf{new}}, \mathcal{D}) &= \int p(y_{\textbf{new}} \mid x_{\textbf{new}}, w, \mathcal{D}) p(w \mid \mathcal{D})\, dw \\
&= \int p(y_{\textbf{new}} \mid x_{\textbf{new}}, w) p(w \mid \mathcal{D})\, dw
\end{aligned}
$$

- For Gaussian regression, predictive distribution has closed form.

Closed Form for Predictive Distribution

- **Model**:

$$\begin{aligned} w &\sim \mathcal{N}(0, \Sigma_0) \\ y_i \mid x, w &\ \text{i.i.d.}\ \mathcal{N}(w^T x_i, \sigma^2) \end{aligned}$$

- **Predictive Distribution**

$$p(y_{\text{new}} \mid x_{\text{new}}, \mathcal{D}) = \int p(y_{\text{new}} \mid x_{\text{new}}, w) p(w \mid \mathcal{D})\, dw.$$

  - Averages over prediction for each w, weighted by posterior distribution.

- **Closed form:**

$$\begin{aligned} y_{\text{new}} \mid x_{\text{new}}, \mathcal{D} &\sim \mathcal{N}(\eta_{\text{new}}, \sigma_{\text{new}}) \\ \eta_{\text{new}} &= \mu_P^T x_{\text{new}} \\ \sigma_{\text{new}} &= \underbrace{x_{\text{new}}^T \Sigma_P x_{\text{new}}}_{\text{from variance in } w} + \underbrace{\sigma^2}_{\text{inherent variance in } y} \end{aligned}$$

Predictive Distributions

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