

Bayesian Regression

David S. Rosenberg

Bloomberg ML EDU

November 16, 2017

Recap: Conditional Probability Models

Parametric Family of Conditional Densities

- A **parametric family of conditional densities** is a set

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

- where $p(y | x, \theta)$ is a density on **outcome space** \mathcal{Y} for each x in **input space** \mathcal{X} , and
- θ is a **parameter** in a [finite dimensional] **parameter space** Θ .
- This is the common starting point for a treatment of classical or Bayesian statistics.

Density vs Mass Functions

- In this lecture, whenever we say “density”, we could replace it with “mass function.”
- Corresponding integrals would be replaced by summations.
- (In more advanced, measure-theoretic treatments, they are each considered densities w.r.t. different base measures.)

- A parametric family of conditional densities:

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

- Assume that $p(y | x, \theta)$ governs the world we are observing, for some $\theta \in \Theta$.
- If we knew the right $\theta \in \Theta$, there would be no need for statistics.
- Instead of θ , we have data \mathcal{D} ... how is it generated?

- **Data:** Suppose we have n inputs $x_1, \dots, x_n \in \mathcal{X}$.
 - For now, x can be chosen randomly, by hand, or adversarially.
 - Our entire development will consider x 's fixed and known.
- For each input x_i , we observe y_i sampled randomly from $p(y \mid x_i, \theta)$.
- We assume the outcomes y_1, \dots, y_n are independent. (Once we know the x 's.)

Likelihood Function

- **Data:** $\mathcal{D} = (y_1, \dots, y_n)$
- The probability density for our data \mathcal{D} is

$$p(\mathcal{D} \mid x_1, \dots, x_n, \theta) = \prod_{i=1}^n p(y_i \mid x_i, \theta).$$

- For fixed \mathcal{D} , the function $\theta \mapsto p(\mathcal{D} \mid x, \theta)$ is the **likelihood function**:

$$L_{\mathcal{D}}(\theta)$$

- The **maximum likelihood estimator (MLE)** for θ in the model $\{p(y \mid x, \theta) \mid \theta \in \Theta\}$ is

$$\hat{\theta}_{\text{MLE}} = \arg \max_{\theta \in \Theta} L_{\mathcal{D}}(\theta).$$

Example: Gaussian Linear Regression

- Input space $\mathcal{X} = \mathbf{R}^d$ Outcome space $\mathcal{Y} = \mathbf{R}$
- **Family of conditional probability densities:**

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

for some known $\sigma^2 > 0$.

- **Parameter space?** \mathbf{R}^d .
- **Data:** $\mathcal{D} = (y_1, \dots, y_n)$
- Assume y_i 's are **conditionally independent**, given x_i 's and w .

Gaussian Likelihood and MLE

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

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

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

$$\begin{aligned} \hat{w}_{\text{MLE}} &= \arg \max_{w \in \mathbf{R}^d} L_{\mathcal{D}}(w) \\ &= \arg \min_{w \in \mathbf{R}^d} \sum_{i=1}^n (y_i - w^T x_i)^2. \end{aligned}$$

¹See <https://davidrosenberg.github.io/ml2015/docs/8.Lab.glm.pdf>, slide 5.

Bayesian Conditional Probability Models

Bayesian Conditional Models

- Input space $\mathcal{X} = \mathbf{R}^d$ Outcome space $\mathcal{Y} = \mathbf{R}$
- Two components to Bayesian conditional model:
 - A **parametric family of conditional densities**:

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

- A **prior distribution** for $\theta \in \Theta$.
- **Prior distribution**: $p(\theta)$ on $\theta \in \Theta$

- The **posterior distribution** for θ is

$$\begin{aligned} p(\theta \mid \mathcal{D}, x_1, \dots, x_n) &\propto p(\mathcal{D} \mid \theta, x_1, \dots, x_n) p(\theta) \\ &= \underbrace{L_{\mathcal{D}}(\theta)}_{\text{likelihood}} \underbrace{p(\theta)}_{\text{prior}} \end{aligned}$$

Gaussian Example: 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**

$$\begin{aligned} p(w \mid \mathcal{D}, x_1, \dots, x_n) &= p(w \mid \mathcal{D}, x_1, \dots, x_n) \\ &\propto L_{\mathcal{D}}(w) p(w) \\ &= \prod_{i=1}^n \left[\frac{1}{\sigma \sqrt{2\pi}} \exp \left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2} \right) \right] \text{ (likelihood)} \\ &\quad \times |2\pi \Sigma_0|^{-1/2} \exp \left(-\frac{1}{2} w^T \Sigma_0^{-1} w \right) \text{ (prior)} \end{aligned}$$

- We have a parametric family of conditional densities:

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

- Each $p(y | x, \theta)$ is a conditional density, but also a prediction function:
 - For $x \in \mathcal{X}$, the action produced is a probability density on y .
- In Bayesian statistics we have two distributions on Θ :
 - the prior distribution $p(\theta)$
 - the posterior distribution $p(\theta | \mathcal{D}, x_1, \dots, x_n)$.
- Each distribution on Θ induces a **distributions over prediction functions**.
- For any give x , this gives a single distribution on y .
- This distribution is called a **predictive distribution**.
- So we can have a **prior predictive distribution** and a **posterior predictive distribution**.

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 $\varepsilon \sim \mathcal{N}(0, 0.2^2)$.

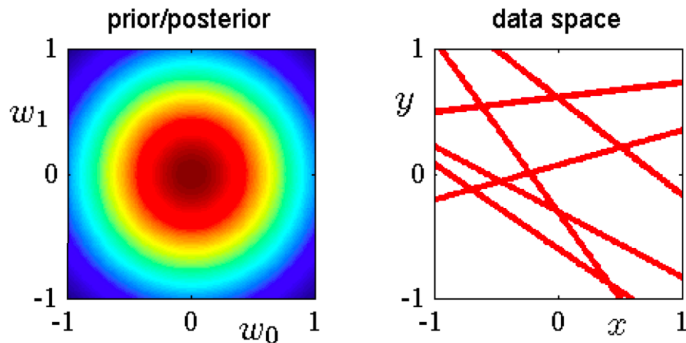
- Written another way, the **conditional probability model** is

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

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

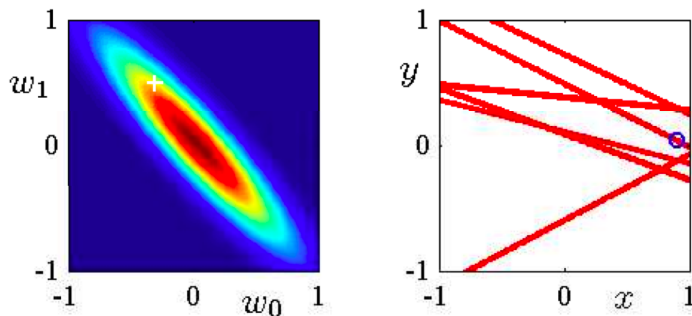
Example in 1-Dimension: Prior Situation

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



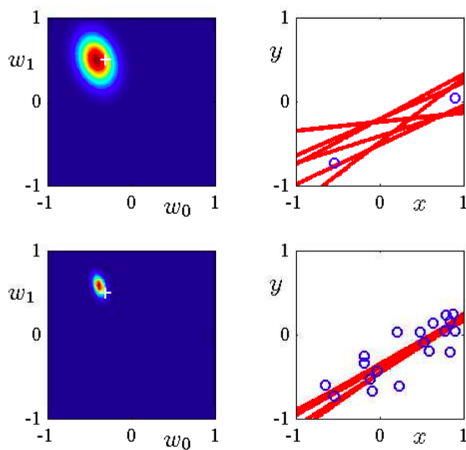
- On right, $y(x) = \mathbb{E}[y | x, w] = w_0 + w_1 x$, for randomly chosen $w \sim p(w) = \mathcal{N}(0, \frac{1}{2}I)$.

Example in 1-Dimension: 1 Observation



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

Example in 1-Dimension: 2 and 20 Observations



Bishop's PRML Fig 3.7

Gaussian Regression Continued

Closed Form for Posterior

- Model:

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

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

$$\begin{aligned}w | \mathcal{D} &\sim \mathcal{N}(\mu_P, \Sigma_P) \\ \mu_P &= (X^T X + \sigma^2 \Sigma_0^{-1})^{-1} X^T y \\ \Sigma_P &= (\sigma^{-2} X^T X + \Sigma_0^{-1})^{-1}\end{aligned}$$

- 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>

- **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 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

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

- By comparison, the posterior variance is

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

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

Posterior Mean and Posterior Mode (MAP)

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

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

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

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

- Which is the ridge regression objective.

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

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

- For Gaussian regression, predictive distribution has closed form.

Closed Form for Predictive Distribution

- **Model:**

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

- **Predictive Distribution**

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

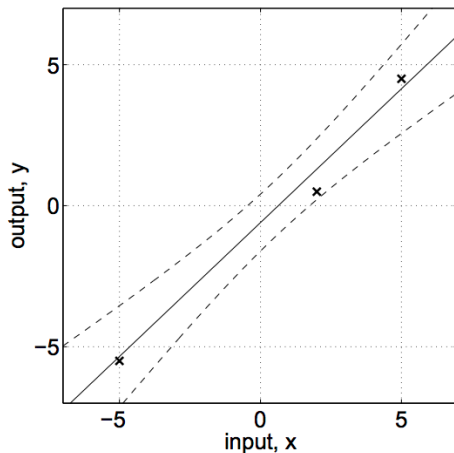
- Averages over prediction for each w , weighted by posterior distribution.

- **Closed form:**

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

Predictive Distributions

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



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