Bayesian Regression

David S. Rosenberg

New York University

March 20, 2018

Contents

Recap: Conditional Probability Models

Parametric Family of Conditional Densities

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

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

  - where $p(y \mid x, \theta)$ 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.)

Parameters

- A parametric family of conditional densities:

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

- Assume that $p(y \mid x, \theta)$ 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?

The Data: Assumptions So Far in this Course

- Our usual setup is that (x, y) pairs are drawn i.i.d. from $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$.
- How have we used this assumption so far?
  - ties validation performance to test performance
  - ties test performance to performance on new data when deployed
  - motivates empirical risk minimization
- The large majority of things we've learned about ridge/lasso/elastic-net regression, optimization, SVMs, and kernel methods are true for arbitrary training data sets $\mathcal{D} : (x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$.
  - i.e. $\mathcal{D}$ could be created by hand, by an adversary, or randomly.
- We rely on the i.i.d. $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ assumption when it comes to **generalization**.

The Data: Conditional Probability Modeling

- To get generalization, we'll still need our usual i.i.d. $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ assumption.
- This time, for developing the model, we'll make some assumptions about the training data...
- We do not need any assumptions on x's .
  - They can be random, chosen by hand, or chosen adversarially.
- For each input x_i,
  - we observe y_i sampled randomly from $p(y \mid x_i, \theta)$, for some unknown $\theta \in \Theta$.
- We assume the outcomes $y_1, \ldots, y_n$ are independent. (Once we know the x's.)

Likelihood Function

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

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

- 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

$$\hat{\theta}_{\text{MLE}} \quad = \quad \underset{\theta \in \Theta}{\arg\max}\, L_{\mathcal{D}}(\theta).$$

Example: Gaussian Linear Regression

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

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

  for some known $\sigma^2 > 0$.
- **Parameter space?** $\mathbf{R}^d$.
- **Data:** $\mathcal{D} = (y_1, \ldots, , y_n)$
- 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:

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

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

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

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

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**:

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

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

The Posterior Distribution

- The **posterior distribution** for θ is

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

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**

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

The Hypothesis Space

- We have a parametric family of conditional densities:

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

- For fixed $\theta \in \Theta$, $p(y \mid x, \theta)$ is a conditional density, but
- For fixed $\theta \in \Theta$, $x \mapsto p(y \mid x, \theta)$ is also a **prediction function:**
  - maps any input $x \in \mathcal{X}$ to a density on $\mathcal{Y}$
- These prediction functions are usually called **predictive distribution functions**.

- As a set of prediction functions, $\{p(y \mid x, \theta) : \theta \in \Theta\}$ is a **hypothesis space**.

Bayesian Distributions on Hypothesis Space

- In Bayesian statistics we have two distributions on Θ:
  - the prior distribution $p(\theta)$
  - the posterior distribution $p(\theta \mid \mathcal{D}, x_1, \ldots, x_n)$.
- Each of these may be thought of as a distribution on the hypothesis space

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

The Prior Predictive Distribution

- Suppose we have a conditional probability modeling problem.
- We choose a parametric family (i.e. set) of conditional densities

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

- We take a Bayesian approach choose a prior distribution $p(\theta)$ on this set.

- Suppose we need a predictive distribution before we've observed any data.
- The **prior predictive distribution function** is given by

$$x \mapsto p(y \mid x) = \int p(y \mid x; \theta) p(\theta)\, d\theta.$$

- This is an average of all conditional densities in our family, weighted by the prior.

The Posterior Predictive Distribution

- Suppose we've already seen data $\mathcal{D}$.
- The **posterior predictive distribution function** is given by

$$x \mapsto p(y \mid x) = \int p(y \mid x; \theta) p(\theta \mid \mathcal{D})\, d\theta.$$

- This is an average of all conditional densities in our family, weighted by the posterior.
- This is not the only predictive distribution function we can use after seeing $\mathcal{D}$.
- The MAP estimator is another one we'll discuss later.

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

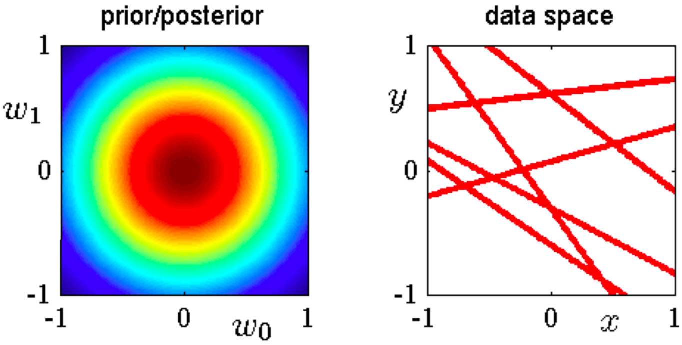
- Written another way, the **conditional probability model** is

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

- 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

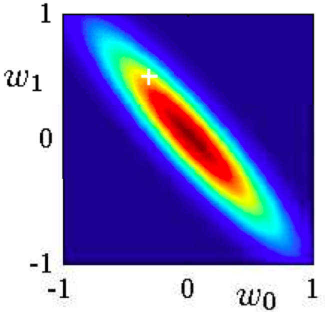
- **Prior distribution:** $w = (w_0, w_1) \sim \mathcal{N}\left(0, \frac{1}{2}I\right)$ (Illustrated on left)

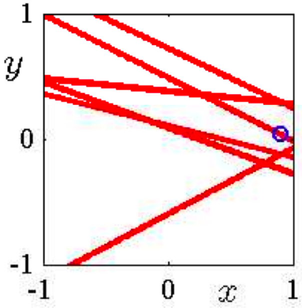


- On right, $y(x) = \mathbb{E}[y \mid x, w] = w_0 + w_1 x$, for randomly chosen $w \sim p(w) = \mathcal{N}\left(0, \frac{1}{2}I\right)$.

Bishop's PRML Fig 3.7

Example in 1-Dimension: 1 Observation

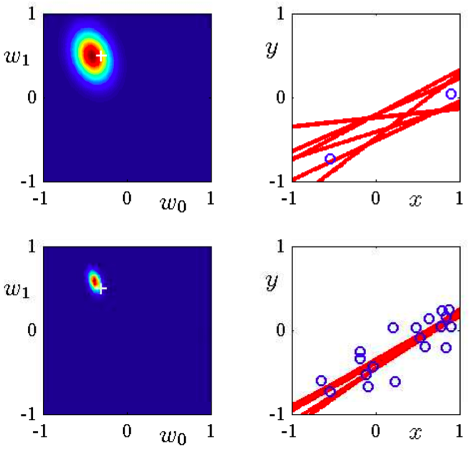




- On left: posterior distribution; white '+' indicates true parameters
- On right: blue circle indicates the training observation

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 $\varepsilon \sim \mathcal{N}\left(0, \sigma^2\right)$.
- 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:

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

- 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**

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

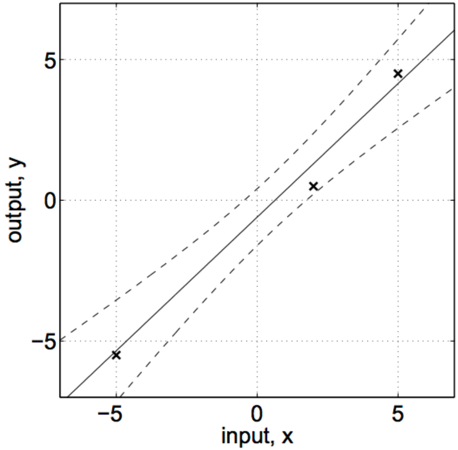
  - Averages over prediction for each w, weighted by posterior distribution.

- **Closed form:**

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

Predictive Distributions

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



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