Bayesian Regression

Julia Kempe & David S. Rosenberg

CDS, NYU

March 26, 2019

Contents

Recap: Conditional Probability Models

Conditional Probability Modeling

- **Input space** $\mathcal{X}$
- **Outcome space** $\mathcal{Y}$
- **Action space** $\mathcal{A} = \{p(y) \mid p \text{ is a probability distribution on } \mathcal{Y}\}$.
- **Hypothesis space** $\mathcal{F}$ contains prediction functions $f : \mathcal{X} \to \mathcal{A}$.
- **Prediction function** $f \in \mathcal{F}$ takes input $x \in \mathcal{X}$ and produces a **distribution** on $\mathcal{Y}$
- We've been discussing **parametric families of conditional densities**

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

- These are also hypothesis spaces for conditional probability modeling.
- Examples?

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.)

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.
- For developing the model, we'll make some assumptions about the training data...
    - In most of what we've done before, we had no assumptions on the training data.
- It's typical (and most general) to do everything "conditional on the x's"
    - That means, we assume the x's are known
    - In particular, we do not consider them random
    - We don't care how they were generated (randomly, adversarially, chosen by hand)
    - In other words, still no assumptions on x's.

The Data: Conditional Probability Modeling

- So we assume the x's are known.
- 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.
  - But not i.i.d. – Why?
  - Each y_i may be drawn from a different distribution, depending on x_i.

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**:

$$L_{\mathcal{D}}(\theta) = p(\mathcal{D} \mid x, \theta),$$

  where $x = (x_1, \ldots, x_n)$.

Maximum Likelihood Estimator

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

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

- MLE corresponds to ERM for the negative log- likelihood loss (discussed previously).
- The corresponding prediction function is

$$\hat{f}(x) = p(y \mid x, \hat{\theta}_{\text{MLE}}).$$

- We can think of this as a choice of a particular function from the hypothesis space

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

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** $p(\theta)$ on $\theta \in \Theta$.

The Posterior Distribution

- The **prior distribution** $p(\theta)$ represents our beliefs about θ before seeing $\mathcal{D}$.
- The **posterior distribution** for θ is

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

- Posterior represents the **rationally "updated" beliefs** after seeing $\mathcal{D}$.
- Each θ corresponds to a prediction function,
  - i.e. the conditional distribution function $p(y \mid x, \theta)$.

Point Estimates of Parameter

- Suppose for some reason we want point estimates of θ.
- We can use Bayesian decision theory to derive point estimates.
- As discussed last week, we may want to use
  - $\hat{\theta} = \mathbb{E}[\theta \mid \mathcal{D}, x]$ (the posterior mean estimate)
  - $\hat{\theta} = \text{median}[\theta \mid \mathcal{D}, x]$
  - $\hat{\theta} = \arg\max_{\theta \in \Theta} p(\theta \mid \mathcal{D}, x)$ (the MAP estimate)
- depending on our loss function.

Back to the basic question

- Find a function takes input $x \in \mathcal{X}$ and produces a **distribution** on $\mathcal{Y}$?
- Recall frequentist approach:
  - Choose family of conditional probability densities (hypothesis space).
  - Select one conditional probability from family, e.g. by MLE.
  - (MLE has nice properties, so a common choice. See advanced statistics class.)

Bayesian Prediction Function

- In Bayesian setting, **there is no selection** from hypothesis space.
- We chose a parametric family of conditional densities

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

- and a prior distribution $p(\theta)$ on this set.
- Suppose we get an x and we need to predict a distribution for the corresponding y.
- Having set our Bayesian model, there are no more decisions to make – just computation...

The Prior Predictive Distribution

- Suppose we have not yet observed any data.
- In Bayesian setting, we can still produce a prediction function.
- The **prior predictive distribution** 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.
- Such an average is also called a **mixture distribution**.

The Posterior Predictive Distribution

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

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

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

Comparison to Frequentist Approach

- In Bayesian statistics we have two distributions on Θ:
  - the prior distribution $p(\theta)$
  - the posterior distribution $p(\theta \mid \mathcal{D})$.
- We also think of these as distributions on the hypothesis space

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

- In frequentist approach, we choose $\hat{\theta} \in \Theta$, and predict

$$p(y \mid x, \hat{\theta}(\mathcal{D})).$$

- In Bayesian approach, we integrate out over Θ w.r.t. $p(\theta \mid \mathcal{D})$ and predict with

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

What if we don't want a full distribution on y?

- Once we have a predictive distribution $p(y \mid x, \mathcal{D})$,
  - we can easily generate single point predictions.
- $x \mapsto \mathbb{E}[y \mid x, \mathcal{D}]$, to minimize expected square error.
- $x \mapsto \text{median}[y \mid x, \mathcal{D}]$, to minimize expected absolute error
- $x \mapsto \arg\max_{y \in \mathcal{Y}} p(y \mid x, \mathcal{D})$, to minimize expected 0/1 loss
- Each of these can be derived from $p(y \mid x, \mathcal{D})$.

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

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

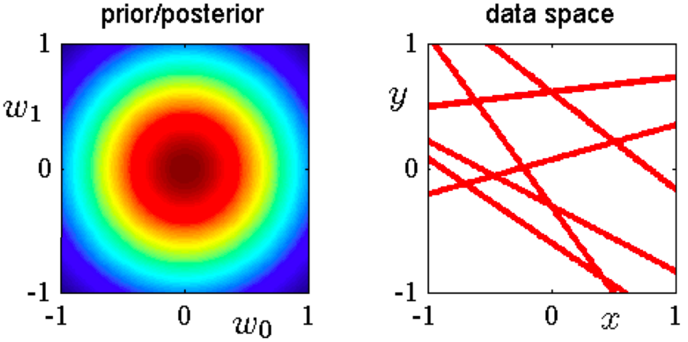
where $\varepsilon \sim \mathcal{N}(0, 0.2^2)$.

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

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

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

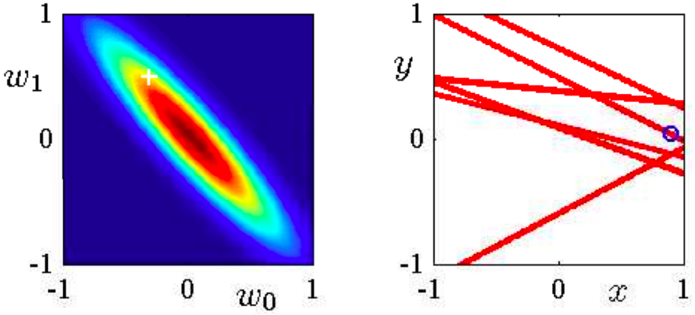
- **Prior distribution:** $w = (w_0, w_1) \sim \mathcal{N}\left(0, \frac{1}{2}I\right)$ (Illustrated on left)



- 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

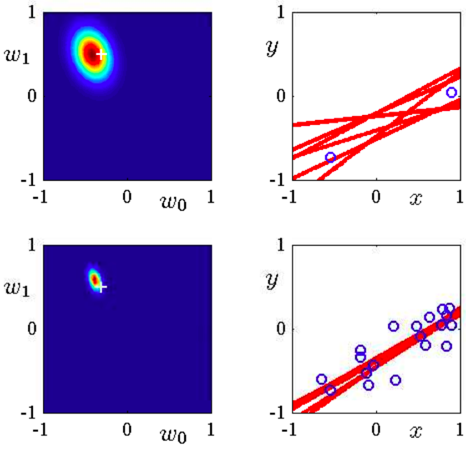
Example in 1-Dimension: 1 Observation



- 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:

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

- 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.**

Closed Form for Posterior

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

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

- If we want point estimates of w, **MAP estimator** and the **posterior mean** are given by

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

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

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

which is of course the ridge regression solution.

Posterior Mean and Posterior Mode (MAP)

- Let's find $\hat{w}_{\text{MAP}}$ another way to elaborate on connection to ridge.
- **Posterior density** on w 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**

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

- For Gaussian regression, predictive distribution has closed form.

Closed Form for Predictive Distribution

- **Model**:

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

- **Predictive Distribution**

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

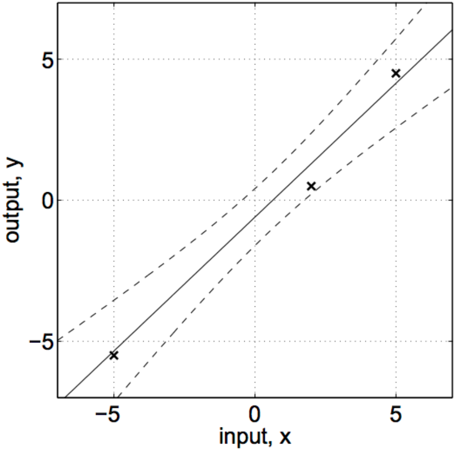
  - Averages over prediction for each w, weighted by posterior distribution.

- **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)