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.

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:**

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

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

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

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

[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

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