Bayesian Regression

David S. Rosenberg

New York University

March 20, 2018
Contents

1 Recap: Conditional Probability Models

2 Bayesian Conditional Probability Models
Recap: Conditional Probability Models
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 \( Y \) for each \( x \) in input space \( X \), and \( \theta \) is a parameter in a [finite dimensional] parameter space \( \Theta \).

This is the common starting point for a treatment of classical or Bayesian statistics.
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:

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

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 \( \theta \), 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 \(P_{X \times Y}\).
- 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 \(D : (x_1, y_1), \ldots, (x_n, y_n) \in X \times Y\).
  - i.e. \(D\) could be created by hand, by an adversary, or randomly.
- We rely on the i.i.d. \(P_{X \times Y}\) assumption when it comes to generalization.
To get generalization, we'll still need our usual i.i.d. $\mathcal{P}_{X \times Y}$ 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 $p(y | x_i, \theta)$, for some unknown $\theta \in \Theta$.

We assume the outcomes $y_1, \ldots, y_n$ are independent.
Likelihood Function

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

\[
p(D | x_1, \ldots, x_n, \theta) = \prod_{i=1}^{n} p(y_i | x_i, \theta).
\]

- For fixed \( D \), the function \( \theta \mapsto p(D | x, \theta) \) is the **likelihood function**:

\[
L_D(\theta)
\]

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

\[
\hat{\theta}_{\text{MLE}} = \arg \max_{\theta \in \Theta} L_D(\theta).
\]
Example: Gaussian Linear Regression

- Input space $\mathcal{X} = \mathbb{R}^d$   
  Outcome space $\mathcal{Y} = \mathbb{R}$
- **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?** $\mathbb{R}^d$.
- **Data:** $\mathcal{D} = (y_1, \ldots , y_n)$
- Assume $y_i$’s are **independent**.
Gaussian Likelihood and MLE

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

\[
L_\mathcal{D}(w) = \prod_{i=1}^{n} p(y_i \mid x_i, w) \quad \text{by conditional independence.}
\]

\[
= \prod_{i=1}^{n} \left[ \frac{1}{\sigma \sqrt{2\pi}} \exp\left( - \frac{(y_i - w^T x_i)^2}{2\sigma^2} \right) \right]
\]

- You should see **in your head**\(^1\) that the MLE is

\[
\hat{w}_{\text{MLE}} = \arg \max_{w \in \mathbb{R}^d} L_\mathcal{D}(w)
\]

\[
= \arg \min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} (y_i - w^T x_i)^2.
\]

---

Bayesian Conditional Probability Models
Bayesian Conditional Models

- Input space $\mathcal{X} = \mathbb{R}^d$  
  Outcome space $\mathcal{Y} = \mathbb{R}$

- 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 $\theta$ is

$$p(\theta \mid \mathcal{D}, x_1, \ldots, x_n) \propto p(\mathcal{D} \mid \theta, x_1, \ldots, x_n)p(\theta)$$

$$= \underbrace{L_{\mathcal{D}}(\theta)}_{\text{likelihood}} \underbrace{p(\theta)}_{\text{prior}}$$
Gaussian Example: Priors and Posteriors

- Choose a Gaussian prior distribution $p(w)$ on $\mathbb{R}^d$:
  \[
  w \sim \mathcal{N}(0, \Sigma_0)
  \]
  for some covariance matrix $\Sigma_0 \succ 0$ (i.e. $\Sigma_0$ is spd).

- Posterior distribution
  \[
  p(w | \mathcal{D}, x_1, \ldots, x_n) = \frac{p(w | \mathcal{D}, x_1, \ldots, x_n)}{L_{\mathcal{D}}(w) p(w)} \propto L_{\mathcal{D}}(w) p(w)
  \]
  \[
  = \prod_{i=1}^{n} \left[ \frac{1}{\sigma \sqrt{2\pi}} \exp\left( -\frac{(y_i - w^T x_i)^2}{2\sigma^2} \right) \right] \quad \text{(likelihood)}
  \]
  \[
  \times |2\pi \Sigma_0|^{-1/2} \exp\left( -\frac{1}{2} w^T \Sigma_0^{-1} w \right) \quad \text{(prior)}
  \]
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 \( x \in X \) to a density on \( Y \)
- 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**.
In Bayesian statistics we have two distributions on $\Theta$:
- the prior distribution $p(\theta)$
- the posterior distribution $p(\theta \mid D, x_1, \ldots, x_n)$.
Each of these may be thought of as a distribution on the hypothesis space

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