Lecture 10: Bayesian regression

Jacob Steinhardt

September 28, 2021
Announcements

- Jacob’s OH moved to Wednesday this week (1:30-2:30)
- Midterm next Tuesday
- HW party today in Evans 458, 4-6pm
Recap

- Bayesian models
- Inference via sampling (MCMC)
Recap

- Bayesian models
- Inference via sampling (MCMC)

This time: Bayesian perspective on regression
Observe data \((x_1, y_1), \ldots, (x_n, y_n)\), where \(x_i \in \mathbb{R}^d\) and \(y_i \in \mathbb{R}\)

Minimize loss function \(L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, y_i; \theta)\)

Example:
- \(\ell(x, y; \theta) = (y - \theta^\top x)^2\) (least squares regression)
- Other examples?
Observe data \((x_1, y_1), \ldots, (x_n, y_n)\) as before, but this time \(y_i \in \{0, 1\}\) (classification)

Still minimize loss function \(L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, y_i; \theta)\)

\[
\ell(x, y; \theta) = -y \log \sigma(\theta^\top x) - (1 - y) \log(1 - \sigma(\theta^\top x))
\]
Observe data \((x_1, y_1), \ldots, (x_n, y_n)\) as before, but this time \(y_i \in \{0, 1\}\) (classification)

Still minimize loss function

\[
L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, y_i; \theta)
\]

\[
\ell(x, y; \theta) = -y \log \sigma(\theta^\top x) - (1 - y) \log(1 - \sigma(\theta^\top x))
= \log(1 + \exp((-1)^y \theta^\top x))
\]

(Recall \(\sigma(z) = \frac{1}{1 + \exp(-z)}\))
Linear Classification: Review

Observe data \((x_1, y_1), \ldots, (x_n, y_n)\) as before, but this time \(y_i \in \{0, 1\}\) (classification)

Still minimize loss function \(L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, y_i; \theta)\)

\[
\ell(x, y; \theta) = -y \log \sigma(\theta^\top x) - (1 - y) \log(1 - \sigma(\theta^\top x)) \\
= \log(1 + \exp((-1)^y \theta^\top x))
\]

(Recall \(\sigma(z) = \frac{1}{1+\exp(-z)}\))

- Where does logistic loss come from?
- How to generalize (e.g. to counting; \(y \in \{0, 1, 2, \ldots\}\))
Consider linear Gaussian model: $y^{(i)} \mid x^{(i)}, \beta \sim N(\beta^\top x^{(i)}, 1)$

Likelihood function: $p(y \mid x, \beta) = \exp(- (y - \beta^\top x)^2 / 2) / \sqrt{2\pi}$
Consider linear Gaussian model: $y^{(i)} \mid x^{(i)}, \beta \sim N(\beta^\top x^{(i)}, 1)$

Likelihood function: $p(y \mid x, \beta) = \exp\left(-\frac{(y - \beta^\top x)^2}{2}\right) / \sqrt{2\pi}$

Maximum likelihood estimate (MLE):

$$\arg\max_{\beta} p(y^{(1:n)} \mid x^{(1:n)}, \beta) = \arg\min_{\beta} -\log p(y^{(1:n)} \mid x^{(1:n)}, \beta)$$
Consider linear Gaussian model: 
\[ y^{(i)} \mid x^{(i)}, \beta \sim N(\beta ^\top x^{(i)}, 1) \]

Likelihood function: 
\[ p(y \mid x, \beta) = \exp\left(-\frac{(y - \beta ^\top x)^2}{2}\right) / \sqrt{2\pi} \]

Maximum likelihood estimate (MLE): 
\[
\arg\max_\beta p(y^{(1:n)} \mid x^{(1:n)}, \beta) = \arg\min_\beta -\log p(y^{(1:n)} \mid x^{(1:n)}, \beta)
\]
\[
= \arg\min_\beta \sum_{i=1}^n \frac{(y^{(i)} - \beta ^\top x^{(i)})^2}{2} + \log(\sqrt{2\pi})
\]
Consider linear Gaussian model: \( y^{(i)} \mid x^{(i)} , \beta \sim N(\beta^\top x^{(i)} , 1) \)

Likelihood function: \( p(y \mid x, \beta) = \exp\left(-\frac{(y - \beta \top x)^2}{2}\right) / \sqrt{2\pi} \)

Maximum likelihood estimate (MLE):

\[
\arg\max_\beta p(y^{(1:n)} \mid x^{(1:n)}, \beta) = \arg\min_\beta -\log p(y^{(1:n)} \mid x^{(1:n)}, \beta) \\
= \arg\min_\beta \sum_{i=1}^n \frac{(y^{(i)} - \beta \top x^{(i)})^2}{2} + \log(\sqrt{2\pi}) \\
= \arg\min_\beta \sum_{i=1}^n (y^{(i)} - \beta \top x^{(i)})^2
\]
Consider linear Gaussian model: \( y^{(i)} | x^{(i)}, \beta \sim N(\beta^\top x^{(i)}, 1) \)

Likelihood function: 
\[
p(y | x, \beta) = \exp\left(-\frac{(y - \beta^\top x)^2}{2}\right) / \sqrt{2\pi}
\]

Maximum likelihood estimate (MLE):
\[
\arg\max_{\beta} p(y^{(1:n)} | x^{(1:n)}, \beta) = \arg\min_{\beta} -\log p(y^{(1:n)} | x^{(1:n)}, \beta)
\]
\[
= \arg\min_{\beta} \sum_{i=1}^{n} \frac{(y^{(i)} - \beta^\top x^{(i)})^2}{2} + \log(\sqrt{2\pi})
\]
\[
= \arg\min_{\beta} \sum_{i=1}^{n} (y^{(i)} - \beta^\top x^{(i)})^2
\]

Least squares regression \( \leftrightarrow \) MLE under Gaussian likelihood!
Recall different estimates of $\beta$: MLE, MAP, full posterior distribution
Beyond MLE

Recall different estimates of $\beta$: MLE, MAP, full posterior distribution

MAP: $\arg\max_{\beta} p(\beta \mid x, y) = \arg\max_{\beta} p(\beta)p(y \mid x, \beta)$

Take Gaussian prior over $\beta$: $\beta \sim N(0, \lambda^2 I)$, or $p(\beta) \propto \exp\left(-\frac{1}{2} \|\beta\|^2 / \lambda^2 \right)$. Ridge regression $\leftrightarrow$ MAP under Gaussian likelihood + prior!
Beyond MLE

Recall different estimates of $\beta$: MLE, MAP, full posterior distribution

MAP: $\arg\max_\beta p(\beta \mid x, y) = \arg\max_\beta p(\beta)p(y \mid x, \beta)$

Take Gaussian prior over $\beta$: $\beta \sim N(0, \lambda^2 I)$, or $p(\beta) \propto \exp(-\frac{1}{2}||\beta||^2_2/\lambda^2)$. 

Ridge regression $\leftrightarrow$ MAP under Gaussian likelihood + prior!
Recall different estimates of $\beta$: MLE, MAP, full posterior distribution

**MAP:**

$$\arg\max_{\beta} p(\beta \mid x, y) = \arg\max_{\beta} p(\beta)p(y \mid x, \beta)$$

Take Gaussian prior over $\beta$: $\beta \sim N(0, \lambda^2 I)$, or $p(\beta) \propto \exp(-\frac{1}{2} \|\beta\|^2 / \lambda^2)$.

$$\beta_{MAP} = \arg\min_{\beta} -\log p(\beta) - \log p(y^{(1:n)} \mid x^{(1:n)}, \beta)$$

$$= \arg\min_{\beta} \|\beta\|^2 / \lambda^2 + \sum_{i=1}^{n} (y^{(i)} - \beta^\top x^{(i)})^2$$
Beyond MLE

Recall different estimates of $\beta$: MLE, MAP, full posterior distribution

MAP: $\arg\max_\beta p(\beta \mid x, y) = \arg\max_\beta p(\beta)p(y \mid x, \beta)$

Take Gaussian prior over $\beta$: $\beta \sim N(0, \lambda^2 I)$, or $p(\beta) \propto \exp(-\frac{1}{2} \|\beta\|_2^2 / \lambda^2)$.

$$\beta_{\text{MAP}} = \arg\min_\beta - \log p(\beta) - \log p(y^{1:n} \mid x^{1:n}, \beta)$$

$$= \arg\min_\beta \|\beta\|_2^2 / \lambda^2 + \sum_{i=1}^n (y^{(i)} - \beta^\top x^{(i)})^2$$

Ridge regression $\leftrightarrow$ MAP under Gaussian likelihood + prior!
Sampling from the posterior

Suppose we want full posterior over $\beta$. Proportional to:

$$p(\beta \mid x^{(1:n)}, y^{(1:n)}) \propto \exp\left(-\frac{1}{2} \|\beta\|_2^2 / \lambda^2\right) \cdot \prod_{i=1}^{n} \exp\left(-\frac{1}{2} (y^{(i)} - \beta^\top x^{(i)})^2\right).$$

In this case, can show posterior over $\beta$ is Gaussian, compute closed form. But could also do Gibbs sampling:

$$p(\beta_j \mid x^{(1:n)}, y^{(1:n)}, \beta_{-j}) \propto \exp\left(-\frac{1}{2} \beta_j^2 / \lambda^2\right) \cdot \prod_{i=1}^{n} \exp\left(-\frac{1}{2} (y^{(i)} - \beta_{-j}^\top x^{(i)}_{-j} - \beta_j x^{(i)}_j)^2\right)$$

In practice, use an off-the-shelf sampling library such as PyMC3.
Linear regression on wind turbine data

[Jupyter demo]
Number of turbines isn’t an arbitrary real number, but integer count in \{0, 1, 2\ldots\}

What’s a common distribution over count data?

\[
p(\kappa | \mu) = \exp(-\mu) \frac{\mu^\kappa}{\kappa!}
\]
Number of turbines isn’t an arbitrary real number, but integer count in \( \{0, 1, 2 \ldots \} \)

What’s a common distribution over count data?

Poisson distribution: \( p_\mu(k) = \exp(-\mu) \mu^k / k! \)
Number of turbines isn’t an arbitrary real number, but integer count in \{0, 1, 2 \ldots \}

What’s a common distribution over count data?

Poisson distribution: \( p_\mu(k) = \exp(-\mu)\mu^k/k! \)

\[ y \mid x, \beta \sim \text{Poisson}(\beta^\top x) \]
Regression on count data

Number of turbines isn’t an arbitrary real number, but integer count in \( \{0, 1, 2, \ldots \} \)

What’s a common distribution over count data?

Poisson distribution: \( p_\mu(k) = \exp(-\mu)\mu^k / k! \)

\( y \mid x, \beta \sim \text{Poisson}(\exp(\beta^\top x)) \)

link function
Regression on count data

Number of turbines isn’t an arbitrary real number, but integer count in \{0, 1, 2\ldots\}

What’s a common distribution over count data?

Poisson distribution: \( p_\mu(k) = \exp(-\mu) \mu^k / k! \)

\[ y \mid x, \beta \sim \text{Poisson}(\exp(\beta^\top x)) \]

Power of Bayesian thinking: just swap in new likelihood!
Poisson regression on turbine data

[Jupyter demo]
Pitfalls of Bayes

Peril of Bayesian thinking: at the mercy of your model

Poisson distribution too narrow, leads to overconfident posterior

Common issue (esp. with count data): overdispersion
Peril of Bayesian thinking: at the mercy of your model

Poisson distribution too narrow, leads to overconfident posterior

Common issue (esp. with count data): **overdispersion**

Typical fix: negative binomial distribution

\[ p_{\mu,\alpha}(k) \propto \binom{k + \alpha - 1}{k} \left( \frac{\mu}{\mu + \alpha} \right)^k \]

Mean \( \mu \), overdispersion \( \alpha \) (variance \( \mu \cdot (1 + \mu / \alpha) \))
Negative binomial plots

[Credit: PyMC3 docs]
Negative binomial regression on turbine data

[Jupyter demo]
Recall loss function for logistic regression: $L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(x^{(i)}, y^{(i)}; \beta)$, where

$$\ell(x, y; \beta) = -y \log \sigma(\beta^{\top} x) - (1 - y) \log (1 - \sigma(\beta^{\top} x))$$
Recall loss function for logistic regression: 

$$ L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(x^{(i)}, y^{(i)}; \beta), $$

where

$$ \ell(x, y; \beta) = -y \log \sigma(\beta^\top x) - (1 - y) \log(1 - \sigma(\beta^\top x)) $$

Negative log-likelihood of Bernoulli (coin flip) model:

$$ y | x, \beta \sim \text{Bernoulli}(\sigma(\beta^\top x)) $$
Logistic regression revisited

Recall loss function for logistic regression: 

\[ L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(x^{(i)}, y^{(i)}; \beta), \]

where

\[ \ell(x, y; \beta) = -y \log \sigma(\beta^\top x) - (1 - y) \log(1 - \sigma(\beta^\top x)) \]

Negative log-likelihood of Bernoulli (coin flip) model:

\[ y \mid x, \beta \sim \text{Bernoulli}(\sigma(\beta^\top x)) \]
Recall loss function for logistic regression: \( L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(x^{(i)}, y^{(i)}; \beta) \), where

\[
\ell(x, y; \beta) = -y \log \sigma(\beta^\top x) - (1 - y) \log(1 - \sigma(\beta^\top x))
\]

Negative log-likelihood of Bernoulli (coin flip) model:

\[
y \mid x, \beta \sim \text{Bernoulli}(\sigma(\beta^\top x))
\]

Logistic regression \(\leftrightarrow\) Bernoulli model with sigmoid link function
Logistic regression revisited

Recall loss function for logistic regression: $L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(x^{(i)}, y^{(i)}; \beta)$, where

$$\ell(x, y; \beta) = -y \log \sigma(\beta^\top x) - (1 - y) \log(1 - \sigma(\beta^\top x))$$

Negative log-likelihood of Bernoulli (coin flip) model:

$$y \mid x, \beta \sim \text{Bernoulli}(\sigma(\beta^\top x))$$

Logistic regression $\leftrightarrow$ Bernoulli model with sigmoid link function

Why sigmoid? ($\sigma(z) = \frac{1}{1 + \exp(-z)} = \frac{\exp(z)}{1 + \exp(z)}$)
Logistic regression revisited

Recall loss function for logistic regression: \( L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(x^{(i)}, y^{(i)}; \beta) \), where

\[
\ell(x, y; \beta) = -y \log \sigma(\beta^\top x) - (1 - y) \log(1 - \sigma(\beta^\top x))
\]

Negative log-likelihood of Bernoulli (coin flip) model:

\( y \mid x, \beta \sim \text{Bernoulli}(\sigma(\beta^\top x)) \)

Logistic regression \( \leftrightarrow \) Bernoulli model with sigmoid link function

Why sigmoid? \( \sigma(z) = \frac{1}{1 + \exp(-z)} = \frac{\exp(z)}{1 + \exp(z)} \)

- Exponentiate to make positive, normalize to add up to 1
- Generalization: softmax \( \exp(z_j) / \sum_{j'} \exp(z_{j'}) \)
Generalized Linear Models

(Inverse) Link function + likelihood. Many libraries handle them!

<table>
<thead>
<tr>
<th>Regression</th>
<th>Inverse link function</th>
<th>Link function</th>
<th>Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>identity</td>
<td>identity</td>
<td>Gaussian</td>
</tr>
<tr>
<td>Logistic</td>
<td>sigmoid</td>
<td>logit</td>
<td>Bernoulli</td>
</tr>
<tr>
<td>Poisson</td>
<td>exponential</td>
<td>log</td>
<td>Poisson</td>
</tr>
<tr>
<td>Negative binomial</td>
<td>exponential</td>
<td>log</td>
<td>Negative binomial</td>
</tr>
</tbody>
</table>
What other modeling assumptions might be violated for the turbine data?