Lecture 10: Bayesian regression

Jacob Steinhardt

September 29, 2020
Recap

- Bayesian models
- Inference via sampling (MCMC)
Recap

- Bayesian models
- Inference via sampling (MCMC)

This time: Bayesian perspective on regression
Linear Regression: Review

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

Example:
- \(\ell(x, y; \beta) = (y - \beta^\top x)^2\) (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 
\[
L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(x^{(i)}, y^{(i)}; \beta)
\]

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

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

\[
= \log(1 + \exp((-1)^y \beta^\top x))
\]

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

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

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

\[
= \log(1 + \exp((-1)^y \beta^\top x))
\]

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

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

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

Maximum likelihood estimate (MLE):

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

$$= \arg\min_\beta \sum_{i=1}^n (y^{(i)} - \beta^\top x^{(i)})^2 / 2 + \log(\sqrt{2\pi})$$
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: $y^{(i)} \mid x^{(i)}, \beta \sim N(\beta^\top x^{(i)}, 1)$

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

Maximum likelihood estimate (MLE):

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

$$= \arg\min_\beta \sum_{i=1}^{n} \left(y^{(i)} - \beta^\top x^{(i)}\right)^2 / 2 + \log(\sqrt{2\pi})$$

$$= \arg\min_\beta \sum_{i=1}^{n} \left(y^{(i)} - \beta^\top x^{(i)}\right)^2$$

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

MAP: $\arg\max_{\beta} p(\beta \mid x, y) = \arg\max_{\beta} p(\beta)p(y \mid x, \beta)$
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 $p(\beta) \propto \exp(-\frac{1}{2} \|\beta\|^2 / \lambda^2)$. 
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 $p(\beta) \propto \exp(-\frac{1}{2} ||\beta||^2_2 / \lambda^2)$.

$$
\begin{align*}
\beta_{\text{MAP}} &= \arg\min_\beta - \log p(\beta) - \log p(y^{(1:n)} \mid x^{(1:n)}, \beta) \\
&= \arg\max_\beta ||\beta||^2_2 / \lambda^2 + \sum_{i=1}^{n} (y^{(i)} - \beta^\top x^{(i)})^2
\end{align*}
$$

Ridge regression $\leftrightarrow$ MAP under Gaussian likelihood + prior!
Beyond MLE

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

**MAP:**
$$\text{argmax}_\beta p(\beta \mid x, y) = \text{argmax}_\beta p(\beta)p(y \mid x, \beta)$$

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

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

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

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

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

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

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

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

In practice, use an off-the-shelf sampling library such as PyMC3
Linear regression on COVID-19 data

[Jupyter demo]
Regression on count data

COVID-19 data isn’t arbitrary real number, but integer count in \{0, 1, 2 \ldots\}

What’s a common distribution over count data?

\[ p(\mu)(k) = \exp(-\mu) \frac{\mu^k}{k!} \mid y|x, \beta \sim \text{Poisson}(\exp(\text{link function}(\beta^\top x))) \]
COVID-19 data isn’t arbitrary real number, but integer count in \{0, 1, 2\ldots\}

What’s a common distribution over count data?

Poisson distribution: \( p_\mu(k) = \exp(-\mu) \mu^k / k! \)
Regression on count data

COVID-19 data isn’t 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

COVID-19 data isn’t arbitrary real number, but integer count in \( \{0, 1, 2 \ldots \} \)

What’s a common distribution over count data?

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

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

link function
Regression on count data

COVID-19 data isn’t 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 COVID-19 data

[Jupyter demo]
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
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**

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 regression on COVID-19 data

[Jupyter demo]
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

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

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

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 \( \exp(z_j) / \sum_{j'} \exp(z_{j'}) \)
What other modeling assumptions might be violated for the COVID-19 data?