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

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

Example:

- $\ell(x, y; \theta) = (y - \theta^\top x)^2$ (least squares regression)
- Other examples?

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

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

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

$$\begin{aligned} \ell(x, y; \theta) &= -y \log \sigma(\theta^\top x) - (1 - y) \log(1 - \sigma(\theta^\top x)) \\ &= \log(1 + \exp((-1)^y \theta^\top x)) \end{aligned}$$

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

$$\begin{aligned} \ell(x, y; \theta) &= -y \log \sigma(\theta^\top x) - (1 - y)\log(1 - \sigma(\theta^\top x)) \\ &= \log(1 + \exp((-1)^y \theta^\top x)) \end{aligned}$$

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

- Where does logistic loss come from?
- How to generalize (e.g. to counting; $y \in \{0, 1, 2, \ldots\}$)

Linear Regression, Bayesian Interpretation

Consider linear Gaussian model: $y^{(i)} \mid x^{(i)}, \beta \sim N(\beta^\top x^{(i)}, 1)$

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

Linear Regression, Bayesian Interpretation

Consider linear Gaussian model: $y^{(i)} \mid x^{(i)}, \beta \sim N(\beta^\top x^{(i)}, 1)$

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

Maximium likelihood estimate (MLE):

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

Linear Regression, Bayesian Interpretation

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):

$$\begin{aligned} \operatorname{argmax}_\beta p(y^{(1:n)} \mid x^{(1:n)}, \beta) &= \operatorname{argmin}_\beta - \log p(y^{(1:n)} \mid x^{(1:n)}, \beta) \\ &= \operatorname{argmin}_\beta \sum_{i=1}^n (y^{(i)} - \beta^\top x^{(i)})^2/2 + \log(\sqrt{2\pi}) \end{aligned}$$

Linear Regression, Bayesian Interpretation

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):

$$\begin{aligned}
\operatorname{argmax}_\beta p(y^{(1:n)} \mid x^{(1:n)}, \beta) &= \operatorname{argmin}_\beta -\log p(y^{(1:n)} \mid x^{(1:n)}, \beta) \\
&= \operatorname{argmin}_\beta \sum_{i=1}^{n} (y^{(i)} - \beta^\top x^{(i)})^2/2 + \log(\sqrt{2\pi}) \\
&= \operatorname{argmin}_\beta \sum_{i=1}^{n} (y^{(i)} - \beta^\top x^{(i)})^2
\end{aligned}$$

Linear Regression, Bayesian Interpretation

Consider linear Gaussian model: $y^{(i)} \mid x^{(i)}, \beta \sim N(\beta^\top x^{(i)}, 1)$

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

Maximium likelihood estimate (MLE):

$$\begin{aligned}
\operatorname{argmax}_\beta p(y^{(1:n)} \mid x^{(1:n)}, \beta) &= \operatorname{argmin}_\beta -\log p(y^{(1:n)} \mid x^{(1:n)}, \beta) \\
&= \operatorname{argmin}_\beta \sum_{i=1}^{n} (y^{(i)} - \beta^\top x^{(i)})^2/2 + \log(\sqrt{2\pi}) \\
&= \operatorname{argmin}_\beta \sum_{i=1}^{n} (y^{(i)} - \beta^\top x^{(i)})^2
\end{aligned}$$

Least squares regression $\leftrightarrow$ MLE under Gaussian likelihood!

Beyond MLE

Recall different estimates of β: MLE, MAP, full posterior distribution

Beyond MLE

Recall different estimates of β: MLE, MAP, full posterior distribution

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

Beyond MLE

Recall different estimates of β: MLE, MAP, full posterior distribution

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

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

Beyond MLE

Recall different estimates of β: MLE, MAP, full posterior distribution

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

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

$$\begin{aligned}\beta_{MAP} &= \operatorname{argmin}_\beta - \log p(\beta) - \log p(y^{(1:n)} \mid x^{(1:n)}, \beta) \\ &= \operatorname{argmin}_\beta \|\beta\|_2^2/\lambda^2 + \sum_{i=1}^{n} (y^{(i)} - \beta^\top x^{(i)})^2\end{aligned}$$

Beyond MLE

Recall different estimates of β: MLE, MAP, full posterior distribution

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

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

$$\begin{aligned}\beta_{MAP} &= \mathrm{argmin}_\beta -\log p(\beta) - \log p(y^{(1:n)} \mid x^{(1:n)}, \beta) \\ &= \mathrm{argmin}_\beta\, \|\beta\|_2^2/\lambda^2 + \sum_{i=1}^{n} (y^{(i)} - \beta^\top x^{(i)})^2\end{aligned}$$

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

Sampling from the posterior

Suppose we want full posterior over β. Proportional to:

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

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

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

In practice, use an off-the-shelf sampling library such as PyMC3

Linear regression on wind turbine data

[Jupyter demo]

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?

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

Regression on count data

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

What's a common distribution over count data?

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

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

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

$$y \mid x, \beta \sim \text{Poisson}(\underbrace{\exp}_{\text{link function}}(\beta^\top x))$$

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

$$y \mid x, \beta \sim \text{Poisson}(\underbrace{\exp}_{\text{link function}}(\beta^\top x))$$

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**

Pitfalls of Bayes

Peril of Bayesian thinking: at the mercy of your model

Poisson distribution too narrow, leads to overconfident posterior

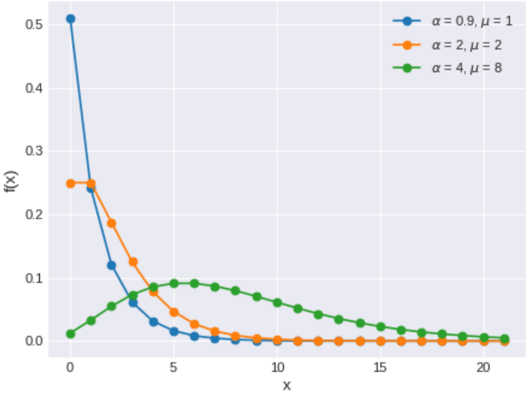
Common issue (esp. with count data): **overdispersion**

Typical fix: negative binomial distribution

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

Mean μ, overdispersion α (variance $\mu \cdot (1 + \mu/\alpha)$)

Negative binomial plots



[Credit: PyMC3 docs]

Negative binomial regression on turbine 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:

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

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:

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

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!

Regression	Inverse link function	Link function	Likelihood
Linear	identity	identity	Gaussian
Logistic	sigmoid	logit	Bernoulli
Poisson	exponential	log	Poisson
Negative binomial	exponential	log	Negative binomial

Discussion: modeling assumptions

What other modeling assumptions might be violated for the turbine data?