Stan by Example

Bernoulli example: bernoulli.stan

- Assume independent observations of Bernoulli random variable
- `data <- list(N = 5, y = c(0, 0, 1, 1, 1))`
- Exercise: write down the log joint distribution as an R function. Hint:

  ```
  lp <- function(theta, data) {
    lp <- ...
    return(lp)   # lp should be a single value
  }
  ```

- Evaluate: `lp(0.3, data)`

Bernoulli example

- Joint model:

$$p(\theta, y) = p(\theta) * \prod_{n=1}^{N} p(y_n|\theta)$$
$$= 1 * \prod_{n=1}^{N} \theta^{y_n}(1-\theta)^{1-y_n}$$

- Log joint:

$$\log p(\theta, y) = \log(1) + \sum_{n=1}^{N} \left(y_n * \log(\theta) + (1-y_n) * \log(1-\theta)\right)$$
$$= \sum_{n=1}^{N} y_n * \log(\theta) + (N - \sum_{n=1}^{N} y_n) * \log(1-\theta)$$

Bernoulli example

```
lp <- function(theta, data) {
  lp <- 0
  for (n in 1:length(data$y))
    lp <- lp + (data$y[n] * log(theta)
                + (1 - data$y[n]) * log(1 - theta))
  return(lp)
}

lp(0.3, data)
lp(0.6, data)
```

Direct translation of model

- `bernoulli_1.stan`
- Recall, Stan specifies joint distribution
- `increment_log_prob()`
- Run model, compare results to `bernoulli.stan`

Bernoulli to Binomial

- Observations from independent Bernoulli random variable
- Drill
  - load `bernoulli_large.data.R`
  - fit using `bernoulli.stan`
  - save result as `bernoulli`
- If $n = \sum_{i=1}^{N} y_n$, then
  $\log p(\theta, y) = n * \log(\theta) + (N - n) * \log(1 - \theta)$
- Compare to `binomial.stan`

Simple linear regression example

- Data: N, y, x
- Generate

```
> a <- 10
> b <- -5
> err_sd <- 10

> data <- list()
> data$N <- 50
> data$x <- rnorm(data$N, 0, 20)
> data$y <- (a + data$x * b) + rnorm(data$N, 0, err_sd)
> plot(data$x, data$y)
```

Simple linear regression

- Joint model:

$$\begin{aligned} a &\sim \dots \\ b &\sim \dots \\ \text{err\_sd} &\sim \dots \\ y &\sim \text{Normal}(a + b * x, \text{err\_sd}) \end{aligned}$$

- Or

$$\begin{aligned} p(a, b, \text{err\_sd}, x, y) = &\frac{1}{\text{err\_sd}\sqrt{2\pi}} \prod_{n=1}^{N} \exp - \frac{(y_n - (a + b * x_n))^2}{2 \times \text{err\_sd}^2} \\ &* p(a) * p(b) * p(\text{err\_sd}) \end{aligned}$$

Drills

- Give 0 data. What do you get?

```
data <- list()
data$N <- 0
data$x <- numeric(0)
data$y <- numeric(0)
```

- How can you fix this?
- List of available distributions in Stan?

Naive Bayes Model, estimation

- Data:
  - V words in volcabulary
  - K topics
  - M documents, each is assigned to one of K topics
  - Each document m has N_m words: $w_{m,1}, \ldots, w_{m,N_m}$
  - z_m topic for document m
- Parameters:
  - θ, a K-simplex, topic prevalance
  - ϕ_k, for each topic, a V-simplex represeting distribution of words

- Model:

$$
\begin{aligned}
\theta &\sim \text{Dirichlet}(1, \ldots, 1) \\
\phi_k &\sim \text{Dirichlet}(0.1, \ldots, 0.1) \\
z_m &\sim \text{Categorical}(\theta) \\
w_{m,n} &\sim \text{Categorical}(\phi_{z_m})
\end{aligned}
$$

- `naive-bayes.stan`, `naive-bayes.data.R`

Naive Bayes, unsupervised

- Data:
  - Same as before, but now we don't know: z_m topic for document m
- Parameters:
  - Same as before.
- Model:

- Marginalize out latent z_m

$$
\begin{aligned}
&\log p(w_{m,1}, \ldots, w_{m,N_m} | \theta, \phi) \\
&= \log \sum_{k=1}^{K} \left(\mathsf{Categorical}(k|\theta) \times \prod_{n=1}^{N_m} \mathsf{Categorical}(w_{m,n}|\phi_k) \right) \\
&= \log \sum_{k=1}^{K} \exp(\log \mathsf{Categorical}(k|\theta) \\
&\quad + \sum_{n=1}^{N_m} \log \mathsf{Categorical}(w_{m,n}|\phi_k))
\end{aligned}
$$

- Fit: `naive-bayes-unsup.stan` `naive-bayes-unsup.data.R`