Date: Thursday, February 28, 2019

Chapter 6

↳ Choosing priors in a "smart" way

The prior distribution provides a way for the modeler to incorporate their knowledge (past experience, etc.) in the statistical model.

Prior specifications are most important / have the biggest impact on our model when we observe few samples of data.

1) Subjective vs Objective Priors

Objective: Let the data speak! Utilize past experiments, data, physical / social laws to construct a prior.

Subjective: Allows the practitioner / modeler to incorporate their own beliefs about the parameters.

On the extreme, an objective prior (the most objective) will not have any preference for values of the parameter
↳ gives the same likelihood to all possibilities.
Leas to a "flat" prior or a $U(a,b)$ where (a,b) is the entire domain of the parameter.

In subjective priors, we place more weight or probability on certain values of the parameter → biases our posterior to give higher weights on the same region.

* Where does data come in?

If it is completely objective, then the data dictates the values of $\mathbb{P}(\theta \mid y)$.

$$\mathbb{P}(\theta \mid y) \propto \underbrace{\pi(\theta)}_{\text{constant over all values}} f(y \mid \theta)$$

If subjective, then $\pi(\theta)$ gives higher (or lower) weights to certain values of θ, which, in turn weights on $f(y|\theta)$ higher (or lower) for those values.

Using data driven methods for validation, one can determine how well the posterior matches the truth (goodness-of-fit). If it does not match, the model should be altered, either through prior choice or through choice of $f(y|\theta)$.

The choice between subjective + objective priors is a bit philosophical → stay principled and think about your data + objective.

Strategic ways of choosing subjective priors:

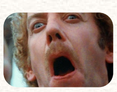
1) **Empirical Bayes**: Choose a prior with hyperparameters α, and estimate α using your data.

Example: $f(x|\theta) = N(\mu, \sigma = 5)$

$$\mu \sim N(\underbrace{\mu_p, \sigma_p^2}_{\text{hyperparameters}})$$

Choices: 1) Hierarchical model – $\mu_p \sim N(0,1)$

$$\sigma_p^2 \sim \chi_1^2$$



Never go full Bayesian ←

2) Scan across a grid of (μ_p, σ_p^2) and check "goodness" of posterior model. This is computationally expensive.

3) Empirical Bayes – take a good guess at μ_p and σ_p^2 using MLE.

a) $\mu_p = \frac{1}{N} \sum_{i=1}^{N} x_i$

b) $\sigma_p^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$

Conjugate families

A conjugate family is a prior-density pair (Prior, data density)
where $p_\beta \cdot f(x|\beta) = p_\beta$. p_β $f(x|\beta$

In other words, the posterior distribution has the same form as the prior distribution.

Same form: Normal prior → Normal posterior
Beta prior → Beta posterior

* why is this nice?

1) Keeping distribution that describes the parameters the same is intuitive. In this way, the data is simply acting to update the shape (mean, std) of the prior.

2) you will know the distributional form (eg Poisson, Normal, Beta, etc) of the posterior. So you only need to estimate $\underbrace{\text{summaries}}_{\text{mean, rate, variance, etc.}}$ of the posterior.

Popular examples:

1) Beta-Binomial model: $X \sim \text{Binomial}(\underset{\text{fixed}}{n}, \underset{\substack{\text{random}\\\text{between } 0,1}}{p})$

$p \sim \underbrace{\text{Beta}(\alpha, \beta)}_{\text{between } 0,1}$

Properties
$\mathbb{E}[p] = \alpha/\beta$; $\text{Var}(p) = \frac{\alpha}{\beta^2}$

Applications:

a) Click-thru rates, conversions

$\#\text{clicks} \sim \text{Binomial}(n, p)$

$\text{rate}(p) \sim \text{Beta}(\alpha, \beta)$

$(X_1, \ldots, X_N)$ Data: # of clicks on several N days of having an ad posted.

$p \mid X_1, \ldots, X_N \sim \text{Beta}(\alpha + \frac{1}{N}\sum_{j=1}^{N} S_j, \beta + \frac{1}{N}\sum_{j=1}^{N} F_j)$

Data gives us $S_1, \ldots, S_N$ and $F_1, \ldots, F_N$ where $S_j = \#$ successes/clicks
$F_j = n - S_j$

2) Dirichlet - Multinomial model:

Dirichlet → multivariate beta distribution
k probability parameters ranging between 0,1

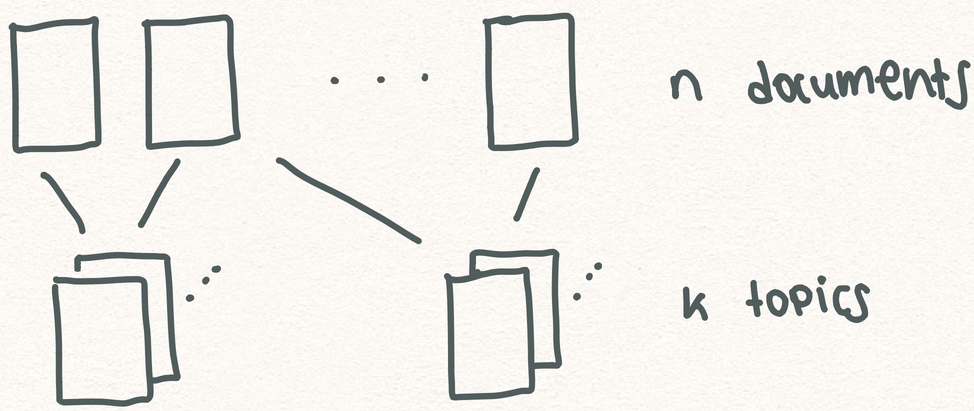
Multinomial → n objects each placed into one of k bins
with probabilities $\pi_1, \dots, \pi_k$.

This is a generalization of the binomial distribution and counts the number of objects in each of the k bins.

Example: Topic modeling in text analysis. Latent Dirichlet allocation is simply an application of this model.

Topic modeling:

Aim: Take n documents of text and bin them into k collections of similar topics.



# of documents per topics $\sim$ multinomial $(n, p_1, \dots, p_k)$

$(p_1, \dots, p_k)^T \sim$ Dirichlet $(\alpha_1, \dots, \alpha_k)$

↓

run Dirichlet-Multinomial model

Output: For each document, D_j, we get a probability of it belonging to each topic.