Generative modelling

Intro to ML

Olga Mineeva
omineeva@student.ethz.ch

Conjugate prior

- Definition

- When do we need it?

Conjugate prior

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters	Interpretation of hyperparameters[note 1]	Posterior predictive[note 2]
Bernoulli	p (probability)	Beta	$\alpha,\, \beta$	$\alpha + \sum_{i=1}^{n} x_i,\, \beta + n - \sum_{i=1}^{n} x_i$	$\alpha - 1$ successes, $\beta - 1$ failures[note 1]	$p(\tilde{x} = 1) = \frac{\alpha'}{\alpha' + \beta'}$
Binomial	p (probability)	Beta	$\alpha,\, \beta$	$\alpha + \sum_{i=1}^{n} x_i,\, \beta + \sum_{i=1}^{n} N_i - \sum_{i=1}^{n} x_i$	$\alpha - 1$ successes, $\beta - 1$ failures[note 1]	$\operatorname{BetaBin}(\tilde{x} \mid \alpha', \beta')$ (beta-binomial)
Negative binomial with known failure number, r	p (probability)	Beta	$\alpha,\, \beta$	$\alpha + \sum_{i=1}^{n} x_i,\, \beta + rn$	$\alpha - 1$ total successes, $\beta - 1$ failures[note 1] (i.e., $\frac{\beta - 1}{r}$ experiments, assuming r stays fixed)	
Poisson	λ (rate)	Gamma	$k,\, \theta$	$k + \sum_{i=1}^{n} x_i,\, \frac{\theta}{n\theta + 1}$	k total occurrences in $\frac{1}{\theta}$ intervals	$\operatorname{NB}(\tilde{x} \mid k', \theta')$ (negative binomial)
			$\alpha,\, \beta$[note 3]	$\alpha + \sum_{i=1}^{n} x_i,\, \beta + n$	α total occurrences in β intervals	$\operatorname{NB}\left(\tilde{x} \mid \alpha', \frac{1}{1 + \beta'}\right)$ (negative binomial)
Categorical	$\boldsymbol{p}$ (probability vector), k (number of categories; i.e., size of $\boldsymbol{p}$)	Dirichlet	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha} + (c_1, \ldots, c_k)$, where c_i is the number of observations in category i	$\alpha_i - 1$ occurrences of category i[note 1]	$p(\tilde{x} = i) = \frac{{\alpha_i}'}{\sum_i {\alpha_i}'} = \frac{\alpha_i + c_i}{\sum_i \alpha_i + n}$
Multinomial	$\boldsymbol{p}$ (probability vector), k (number of categories; i.e., size of $\boldsymbol{p}$)	Dirichlet	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha} + \sum_{i=1}^{n} \mathbf{x}_i$	$\alpha_i - 1$ occurrences of category i[note 1]	$\operatorname{DirMult}(\tilde{\mathbf{x}} \mid \boldsymbol{\alpha}')$ (Dirichlet-multinomial)
Hypergeometric with known total population size, N	M (number of target members)	Beta-binomial[4]	$n = N, \alpha,\, \beta$	$\alpha + \sum_{i=1}^{n} x_i,\, \beta + \sum_{i=1}^{n} N_i - \sum_{i=1}^{n} x_i$	$\alpha - 1$ successes, $\beta - 1$ failures[note 1]	
Geometric	p_0 (probability)	Beta	$\alpha,\, \beta$	$\alpha + n,\, \beta + \sum_{i=1}^{n} x_i - n$	$\alpha - 1$ experiments, $\beta - 1$ total failures[note 1]	

https://en.wikipedia.org/wiki/Conjugate_prior

Multinomial distribution

A discrete distribution has a finite set of outcomes $1, ..., m$. It is parameterized by a vector $\boldsymbol{\theta} = (\theta_1, ..., \theta_m), \quad \sum_j \theta_j = 1, \quad P(X = j|\boldsymbol{\theta}) = \theta_j$

Suppose $X_i \sim Discrete(\boldsymbol{\theta})$ for $i = 1, ..., n$ and N_j is the number of times j occurs in $\boldsymbol{X}$

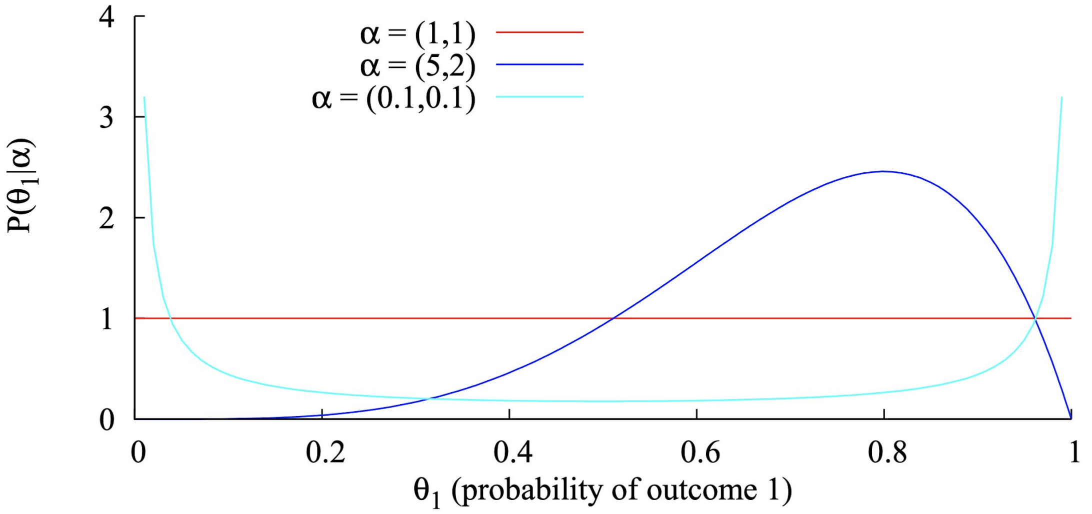
Then $\boldsymbol{N}|n, \boldsymbol{\theta} \sim Multi(\boldsymbol{\theta}, n)$

$$\boxed{P(\boldsymbol{N}|n, \boldsymbol{\theta}) = \frac{n!}{\prod_{j=1}^m N_j!} \prod_{j=1}^m \theta_j^{N_j}}$$

Dirichlet distribution

$$\boxed{P(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{j=1}^{m} \alpha_j)}{\prod_{j=1}^{m} \Gamma(\alpha_j)} \prod_{k=1}^{m} \theta_k^{\alpha_k - 1}}$$

$$\alpha = (\alpha_1, ..., \alpha_m), \quad where \quad \alpha_j > 0$$



https://plot.ly/~david_avakian/14.embed

Inference for θ with Dirichlet priors

By Bayes Rule, posterior is:

$$
\begin{aligned}
P(\boldsymbol{\theta}|\boldsymbol{X}) &\propto P(\boldsymbol{X}|\boldsymbol{\theta})P(\boldsymbol{\theta}) \\
&\propto \Big(\prod_{j=1}^{m} \theta_j^{N_j}\Big)\Big(\prod_{j=1}^{m} \theta_j^{\alpha_j - 1}\Big) \\
&= \prod_{j=1}^{m} \theta_j^{N_j + \alpha_j - 1}, \quad so
\end{aligned}
$$

$$P(\boldsymbol{\theta}|\boldsymbol{X}) = Dir(\boldsymbol{N} + \boldsymbol{\alpha})$$

Inference for θ with Dirichlet priors

By Bayes Rule, posterior is:

$$P(\boldsymbol{\theta}|\boldsymbol{X}) = Dir(\boldsymbol{N} + \boldsymbol{\alpha})$$

So if prior is Dirichlet with parameters $\boldsymbol{\alpha}$,
posterior is Dirichlet with parameters $\boldsymbol{N} + \boldsymbol{\alpha}$
⇒ can regard Dirichlet parameters $\boldsymbol{\alpha}$ as "pseudo-counts" from "pseudo-data"

Normalising constant?

Point estimates from Bayesian posterior

- MLE

$$\boldsymbol{\theta}_j^* = \frac{N_j}{n}$$

- MAP

$$\boldsymbol{\theta}_j^* = \frac{N_j + \alpha_j - 1}{n + \sum_{j'=1}^{m}(\alpha_{j'} - 1)}$$

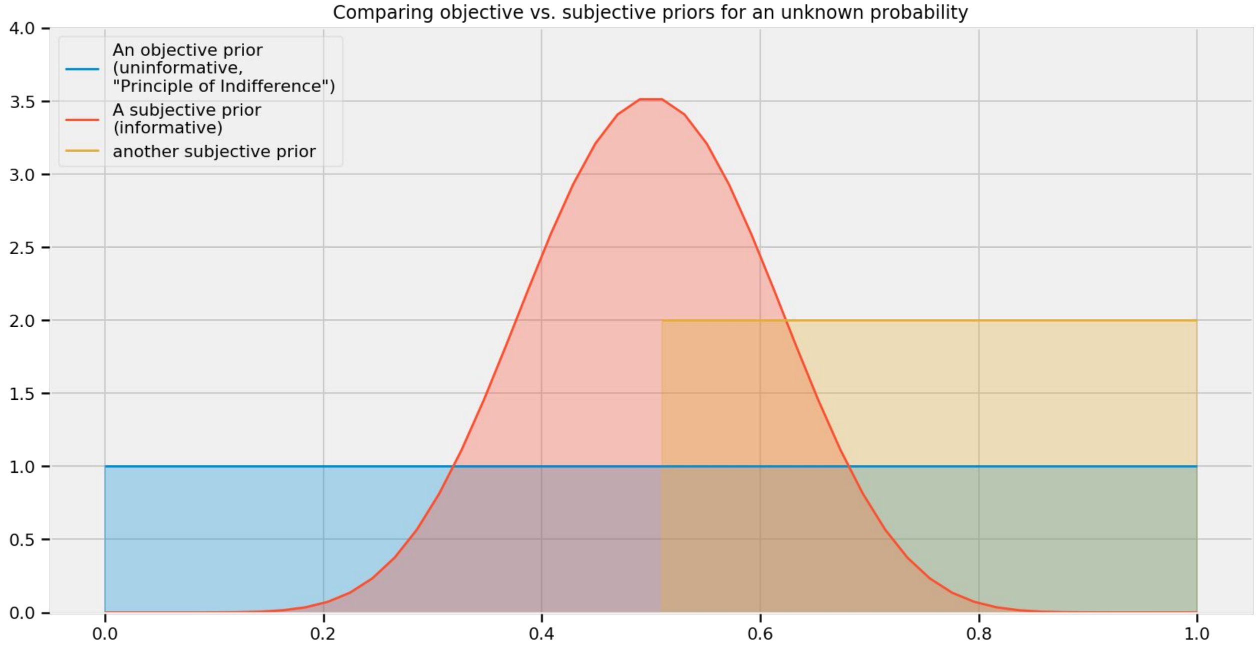
Connection between them https://wiseodd.github.io/techblog/2017/01/01/mle-vs-map/

Regularisation

- Degeneracy in GMM or GBC

- Wishart distribution – is a family of probability distributions for symmetric positive definite matrices

Can be priors harmful?



https://github.com/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers

EM for the Mixture of Distributions

$$x_i \in \{1, 2, 3\}, where \quad i = 1, ..., N$$

$$p(x) = \gamma p_1(x) + (1 - \gamma) p_2(x)$$

$$p_1(x) = \begin{cases} \alpha, & if\ x = 1 \\ 1 - \alpha, & if\ x = 2 \\ 0, & if\ x = 3 \end{cases} \qquad p_2(x) = \begin{cases} 0, & if\ x = 1 \\ 1 - \beta, & if\ x = 2 \\ \beta, & if\ x = 3 \end{cases}$$

$k_1 = 30, k_2 = 20, k_3 = 60$ – observations

$\alpha_0 = \beta_0 = \gamma_0 = \frac{1}{2}$ – starting point for EM

To do:

1. Write joint distribution over observed and latent variables governed by parameters $\theta = (\alpha, \beta, \gamma)$.

2. E step. Evaluate the responsibilities using the current parameter values.

3. M step. Re-estimate the parameters using the current responsibilities.

4. Using given numbers calculate E and M steps until convergence.

Solution

$$p(X,Z) = \gamma^{k_1}\alpha^{k_1}(1-\gamma)^{k_3}\beta^{k_3}\prod_{i=1}^{k_2}(\gamma(1-\alpha))^{[z_i=1]}\,((1-\gamma)(1-\beta))^{[z_i=2]}$$

$$p(z_i = 1|X,\theta) = \frac{\gamma(1-\alpha)}{1-\beta+\gamma(\beta-\alpha)} = \sigma \qquad p(z_i = 2|X,\theta) = \frac{(1-\gamma)(1-\beta)}{1-\beta+\gamma(\beta-\alpha)}$$

$$\alpha = \frac{k_1}{k_1+k_2\sigma} \qquad \beta = \frac{k_3}{k_3+k_2\sigma} \qquad \gamma = \frac{k_1+k_2\sigma}{k_1+k_2+k_3}$$

$$\alpha = \frac{3}{4}, \beta = \frac{6}{7}, \gamma = \frac{4}{11}$$