Theorem: (Bayes)

Given a likelihood $L(\vartheta, x)$ and a prior $P(\vartheta)$ for ϑ. The posterior distribution for ϑ (The conditional distribution of ϑ given the data X) is given by:

$$P(\vartheta | X) = \frac{L(\vartheta, x) \cdot P(\vartheta)}{\int L(\vartheta', x) \cdot P(\vartheta') d\vartheta'}$$

After normalization we write:

$$P(\vartheta | X) \propto L(\vartheta, x) \cdot P(\vartheta)$$

i.e. posterior $\propto$ likelihood $\cdot$ prior. The quantity $P(x) = \int L(\vartheta', x) P(\vartheta') d\vartheta'$ is called the marginal distribution of x. → marginal likelihood

Example: Suppose $X \sim Bin(n, \vartheta)$ and that our prior distribution for ϑ is $Beta(a, b)$

i.e. $P(\vartheta) = \dfrac{\vartheta^{a-1}(1-\vartheta)^{b-1}}{B(a, b)}$, $\quad 0 < \vartheta < 1$

By Baye's Theorem:

$$p(\vartheta|x) \propto \text{likelihood} \times \text{prior}$$
$$\propto \vartheta^{x}(1-\vartheta)^{n-x}\,\vartheta^{a-1}\cdot(1-\vartheta)^{b-1}$$
$$= \vartheta^{a+x-1}(1-\vartheta)^{n-x+b-1}$$

Conjugate priors:

Definition: (Conjugacy) Consider $L(\vartheta, x)$ s.t. $\vartheta \in \Theta$, $x \in \mathcal{X}$. We say that a family of prior distributions $(P_\gamma)_{\gamma \in \Gamma}$ is conjugate if for all $\gamma \in \Gamma$ and $x \in \mathcal{X}$, there exists $\gamma(x) \in \Theta$ s.t. $P_\gamma(\cdot|x) = P_{\gamma(x)}(\cdot)$.

We say that the prior and the posterior are conjugate distributions, and the prior is conjugate prior for the likelihood. In other words we say a conjugate prior is a prior which, when combined with the likelihood, produces a posterior distribution in the same family as the prior.

Proposition: (conjugate priors for exponentials families)

Suppose $L(\vartheta, x) = h(x) \exp\left\{\sum_{i=1}^{k} \eta_i(\theta) T_i(x) - d(\theta)\right\}$

defines a k-parameter exponential family. Then the distributions of the form:

$$P_\gamma(\theta) \propto \exp\left\{\gamma_0 d(\vartheta) + \sum_{i=1}^{k} \gamma_i \eta_i(\theta)\right\}$$

for parameters $\gamma = (\gamma_0, \dots, \gamma_k)$ are a conjugate prior family.

Proof: (HW)

Example: Let $x = (X_1, \dots, X_n)$ be a sample of i.i.d. $Poi(\vartheta)$ r.v. so the joint likelihood is $L(\vartheta, x) \propto \exp(-n\vartheta + T(x) \log \theta)$

where $T(x) = \sum_{i=1}^{n} x_i$. So the natural conjugate prior is of the form:

$$P(\vartheta) \propto \exp(\gamma_0 \vartheta + \gamma_1 \log \vartheta)$$

writing $\beta = -\gamma_0$ and $\alpha = \gamma_1 + 1$ we have

$P(\vartheta) \propto \vartheta^{\alpha - 1} \cdot e^{-\beta\vartheta}$ which is a pdf of $\Gamma(\alpha, \beta)$. We can see that the posterior distribution is $\Gamma(\alpha + T(x), \beta + n)$ so indeed the gamma distribution is a conjugate prior (for the poisson likelihood)

<u>Example</u>: (multinomial distribution and Dirichlet prior)

Consider a multinomial distribution with N trials and K levels with likelihood

$$P(x^{1:K} \mid \vartheta) = \vartheta_1^{x_1} \cdots \vartheta_K^{x_K}$$

Conjugate priors take the form

$$P(\vartheta \mid a) = \vartheta_1^{a_1} \cdots \vartheta_K^{a_K} \quad \sum \vartheta_i = 1 \quad \vartheta_i \geq 0$$

The above defines a prior when $a_1, \ldots, a_K > -1$, so it is more natural to parametrise as

$$P(\vartheta \mid a) \propto \vartheta_1^{a_1 - 1} \cdots \vartheta_K^{a_K - 1} \quad \sum \vartheta_i = 1$$
$$\vartheta_i \geq 0$$

This is called the Dirichlet distribution denoted $\text{Dirichlet}(a_1, \ldots a_K)$

if $(x_1, \dots, x_n)$ are i.i.d samples from $p(\cdot \mid \vartheta)$ where for $i=1,\dots,n$, $x_i = (x_i^1, \dots, x_i^k)$

$$p(x_{1:N} \mid \vartheta) = \vartheta_1^{\sum x_j^1} \dots \vartheta_k^{\sum x_j^k}$$

It can be easily seen that the posterior is also Dirichlet

$$p(\vartheta \mid x_{1:n}, a) = \text{Dirichlet}\left(\sum x_j^1 + a_1 - 1, \dots, \sum x_j^k + a_k - 1\right)$$

Improper Priors:

Remark: we do not require that the prior to be a "real" probability distribution for the posterior to exist and be well-defined

Definition: we say that a pdf P is an improper prior if it has infinite mass:

$$\int_{\Theta} p(\vartheta)\, d\vartheta = \infty \quad p(\vartheta) \geq 0 \ \forall \vartheta \in \Theta.$$

A posterior distribution can be defined as long as $\widehat{p(\vartheta \mid x)}$

$$\int_{\Theta} f(x;\theta)\, p(\theta)\, d\theta < \infty$$

Example:

(1) Likelihood $X|\theta \sim N(\theta, 1)$ and prior $p(\theta) = 1 \ \forall \theta \in \mathbb{R}$. In this case $\log p(\theta|x) = -\frac{1}{2}(x-\theta)^2 +$ constant, i.e posterior distribution is $N(x, 1)$

(2) Likelihood $X|p \sim Bin(n, p)$ $p(p) = (p(1-p))^{-1}$

The posterior is $p(p|x) \propto p^{x-1}(1-p)^{n-x-1}$ which is improper iff $x=0$ or $x=n$; so the posterior is not always well defined.

Predictive distributions:

Briefly, we discuss on how we can make predictions for new data points

Definition: if $X_1, \dots, X_n, X_{n+1}$ are i.i.d. observations from the distribution $f(x, \theta)$ with prior $p(\theta)$

then the posterior distribution is

$$f(x_{n+1} \mid x_n) = \int_{\Theta} f(x_{n+1}, \theta) \cdot p(\theta \mid x)\, dx$$

where $x = (x_1, \ldots, x_n)$

<u>Note</u>: Predictive distribution describes the distribution of a new observation given all the observations we've already made.

<u>Example</u>: Poison likelihood, Gamma prior

Suppose $Y \sim Poi(\theta)$ and that our prior for θ is a $\Gamma(\alpha, \beta)$ distribution.

The marginal likelihood for this model is

$$m(y) = \int_0^\infty e^{-\lambda} \frac{\lambda^y}{y!} \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} \cdot e^{-\beta\lambda}\, d\lambda.$$

On the other hand, we can use that

$$p(\theta \mid y) = \frac{f(y, \theta) \cdot p(\theta)}{m(y)}, \text{ so}$$

$$m(y) = \frac{f(y,\vartheta)\cdot p(\vartheta)}{p(\vartheta \mid y)}.$$

we have seen previously that in this setting the posterior is $\pi(\vartheta \mid y) \sim \Gamma(\alpha+y,\ \beta+1)$

Hence $$m(y) = \frac{\left(\frac{e^{-\vartheta}\cdot \vartheta^{y}}{y!}\right)\left(\frac{\beta^{\alpha}\cdot e^{-\beta\vartheta}\cdot \vartheta^{\alpha-1}}{\Gamma(\alpha)}\right)}{\left(\frac{(\beta+1)^{\alpha+y}\cdot \vartheta^{\alpha+y-1}\, e^{-(\beta+1)\vartheta}}{\Gamma(\alpha+y)}\right)}$$

$$= \frac{\Gamma(\alpha+y)}{\Gamma(\alpha)\, y!}\left(\frac{\beta}{\beta+1}\right)^{\alpha}\left(\frac{1}{\beta+1}\right)^{y}$$

which is pmf of a neg Bin(α, β)

Thus we have shown that the densities/mass(es) of the Poisson, Gamma and negative binomial distribution are related by

$$P_{NegBin}(y;\alpha,\beta) = \int_0^\infty P_{Po}(y;\vartheta)\cdot P_{\Gamma}(\vartheta;\alpha,\beta)\, d\vartheta$$

Hence, the predictive distribution has pmf

$$P(y_{n+1} \mid y_n) = \int_0^\infty P_{Po}(y_{n+1}; \theta)\, P_{\Gamma}(\theta; a + \sum y_i, \beta + n)\, d\theta$$

$$= P_{NegBin}(y_{n+1}; a + \sum y_i, b + n)$$

So is negative Binomial distribution with parameters $a + \sum_{i=1}^{n} y_i$ and $\beta + n$.