STA 414/2104

Statistical Methods for Machine Learning and Data Mining

Radford M. Neal, University of Toronto, 2012

Week 4

Analytically-Tractable Bayesian Models

Conjugate Prior Distributions

For most Bayesian inference problems, the integrals needed to do inference and prediction are not analytically tractable — hence the need for numerical quadrature, Monte Carlo methods, or various approximations.

Most of the exceptions involve *conjugate priors*, which combine nicely with the likelihood to give a posterior distribution of the same form. Examples:

1) Independent observations from a finite set, with Beta / Dirichlet priors.

2) Independent observations of Gaussian variables with Gaussian prior for the mean, and either known variance or inverse-Gamma prior for the variance.

3) Linear regression with Gaussian prior for the regression coefficients, and Gaussian noise, with known variance or inverse-Gamma prior for the variance.

It's nice when a tractable model and prior are appropriate for the problem. Unfortunately, people are tempted to use such models and priors even when they aren't appropriate.

Independent Binary Observations with Beta Prior

We observe binary (0/1) variables $Y_1, Y_2, \ldots, Y_n$.

We model these as being *independent*, and *identically distributed*, with

$$P(Y_i = y \,|\, \theta) \;=\; \left\{ \begin{array}{ll} \theta & \text{if } y = 1 \\ 1-\theta & \text{if } y = 0 \end{array} \right\} \;=\; \theta^y \, (1-\theta)^{1-y}$$

Let's suppose that our prior distribution for θ is Beta(a,b), with a and b being known positive reals. With this prior, the probability density over $(0,1)$ of θ is:

$$P(\theta) \;=\; \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, \theta^{a-1}(1-\theta)^{b-1}$$

Here, the Gamma function, $\Gamma(c)$, is defined to be $\int_0^\infty x^{c-1} \exp(-x)\, dx$. Note that $\Gamma(c) \;=\; (c-1)!$ when c is an integer.

When $a = b = 1$ the prior is uniform over $(0,1)$.

The prior mean of θ is $a \,/\, (a+b)$. Big a and b give smaller prior variance.

Posterior Distribution with Beta Prior

With this Beta prior, the posterior distribution is also Beta:

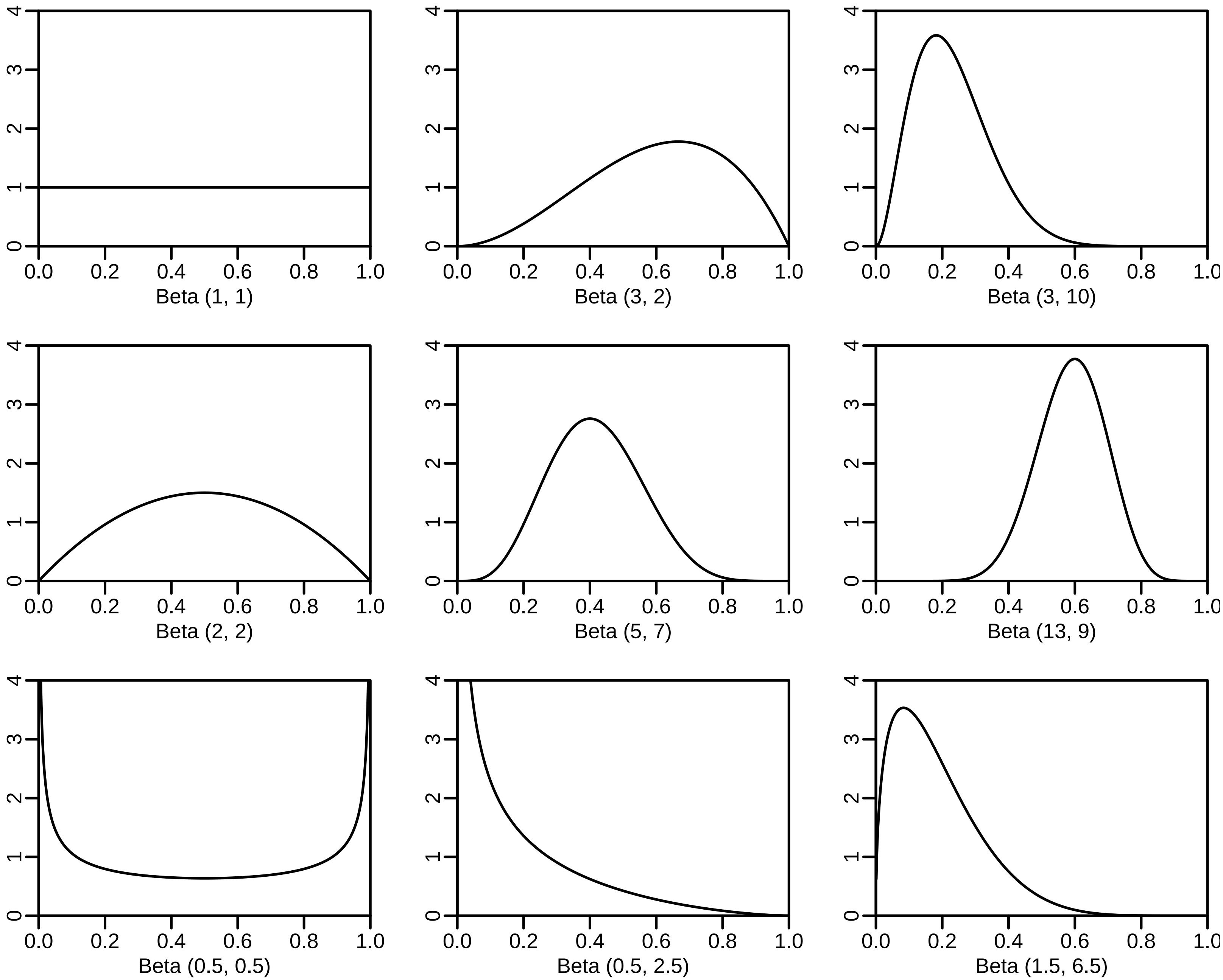
$$
\begin{aligned}
& P(\theta \,|\, Y_1 = y_1,\, Y_2 = y_2,\, \ldots,\, Y_n = y_n) \\
& \quad \propto \quad P(\theta) \prod_{i=1}^{n} P(Y_i = y_i \,|\, \theta) \\
& \quad \propto \quad \theta^{a-1}\,(1-\theta)^{b-1} \prod_{i=1}^{n} \theta^{y_i}\,(1-\theta)^{1-y_i} \\
& \quad \propto \quad \theta^{\Sigma y_i + a - 1}\,(1-\theta)^{n - \Sigma y_i + b - 1}
\end{aligned}
$$

So the posterior distribution is Beta $(\sum y_i + a,\ n - \sum y_i + b)$.

One way this is sometimes visualized is as the prior being equivalent to a fictitious observations with $Y = 1$ and b fictitious observations with $Y = 0$.

Note that all that is used from the data is $\sum y_i$, which is a *minimal sufficient statistic*, whose values are in one-to-one correspondence with possible likelihood functions (ignoring constant factors).

Examples of Beta Priors and Posteriors



Predictive Distribution from Beta Posterior

From the Beta$\left(\sum y_i + a,\ n - \sum y_i + b\right)$ posterior distribution, we can make a probabilistic prediction for the next observation:

$$
\begin{aligned}
&P(Y_{n+1} = 1 \,|\, Y_1 = y_1,\ Y_2 = y_2,\ \ldots,\ Y_n = y_n) \\
&\qquad = \int_0^1 P(Y_{n+1} = 1 \,|\, \theta)\, P(\theta \,|\, Y_1 = y_1,\ Y_2 = y_2,\ \ldots,\ Y_n = y_n)\, d\theta \\
&\qquad = \int_0^1 \theta\, P(\theta \,|\, Y_1 = y_1,\ Y_2 = y_2,\ \ldots,\ Y_n = y_n)\, d\theta \\
&\qquad = \int_0^1 \theta\, \frac{\Gamma(n+a+b)}{\Gamma(\Sigma y_i + a)\Gamma(n - \Sigma y_i + b)}\, \theta^{\Sigma y_i + a - 1}\, (1-\theta)^{n - \Sigma y_i + b - 1}\, d\theta \\
&\qquad = \frac{\Gamma(n+a+b)}{\Gamma(\Sigma y_i + a)\Gamma(n - \Sigma y_i + b)}\, \frac{\Gamma(1 + \Sigma y_i + a)\Gamma(n - \Sigma y_i + b)}{\Gamma(1 + n + a + b)} \\
&\qquad = \frac{\sum y_i + a}{n + a + b}
\end{aligned}
$$

This uses the fact that $c\Gamma(c) = \Gamma(1+c)$.

Generalizing to More Than Two Values

For i.i.d. observations with a finite number, K, of possible values, with $K > 2$, the conjugate prior for the probabilities $\theta_1, \ldots, \theta_K$ is the Dirichlet distribution, with the following density on the simplex where all $\theta_k > 0$ and $\sum \theta_k = 1$:

$$P(\theta_1, \ldots, \theta_K) \;\; = \;\; \frac{\Gamma(\Sigma_k \alpha_k)}{\Pi_k \, \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

The parameters $\alpha_1, \ldots, \alpha_K$ can be any positive reals.

The posterior distribution after observing n items, with m_1 having value 1, m_2 having value 2, etc. is Dirichlet with parameters $\alpha_1 + m_1, \ldots, \alpha_K + m_K$.

The predictive distribution for item $n+1$ is

$$P(Y_{n+1} = k \,|\, Y_1 = y_1, \ldots, Y_K = y_k) \;\; = \;\; \frac{m_k + \alpha_k}{n + \Sigma \alpha_k}$$

Independent Observations from a Gaussian Distribution

We observe real variables Y_1, Y_2, $\ldots$, Y_n.

We model these as being independent, all from some Gaussian distribution with unknown mean, μ, and known variance, σ^2.

The conjugate prior for μ is Gaussian with some mean μ_0 and variance σ_0^2.

Rather than talk about the variance, it is more convenient to talk about the *precision*, equal to the reciprocal of the variance. A data point has precision $\tau = 1/\sigma^2$ and the prior has precision $\tau_0 = 1/\sigma_0^2$.

The posterior distribution for μ is also Gaussian, with precision $\tau_n = \tau_0 + n\tau$, and with mean

$$\mu_n = \frac{\tau_0\mu_0 + n\tau\overline{y}}{\tau_0 + n\tau}$$

where $\overline{y}$ is the sample mean of the observatons $y_1, \ldots, y_n$.

The predictive distribution for Y_{n+1} is Gaussian with mean μ_n and variance $(1/\tau_n) + \sigma^2$.

If we let σ_0 go to infinity — an example of an *improper* prior — the posterior mean, μ_n, will equal the sample mean, $\overline{y}$.

Gaussian with Unknown Variance

What if both the mean and the variance (precision) of the Gaussian distribution for $Y_1, \ldots, Y_n$ are unknown?

There is still a conjugate prior, but in it, μ and τ are dependent:

$$\begin{aligned} \tau &\sim \text{Gamma}(a, b) \\ \mu \,|\, \tau &\sim N(\mu_0, c/\tau) \end{aligned}$$

for some positive constants a, b, and c.

It's hard to imagine circumstances where our prior information about μ and τ would have a dependence of this sort. But unfortunately, people use this conjugate prior anyway, because it's convenient.

Bayesian Linear Basis Function Models

A Bayesian Linear Basis Function Model

Let's set up a Bayesian linear basis function model by giving β a Gaussian prior:

$$\begin{aligned} y_i \,|\, x_i,\, \beta &\sim N(\phi(x_i)^T\beta,\, \sigma^2) \\ \beta &\sim N(m_0, S_0) \end{aligned}$$

This Gaussian prior will turn out to be conjugate.

For the moment, we regard σ^2, m_0, and S_0 as known.

Often, we will let $m_0 = 0$ and let S_0 be diagonal, so that the β_j are independent. We might let β_0 have a large variance, and all the other β_j have the same variance.

The symbol y will sometime denote a single, generic response value, and other times denote the vector $[y_1, \ldots, y_n]^T$ of responses for training cases. We use Φ for the matrix of basis function values for the n training cases.

Multivariate Gaussian Model with Multivariate Gaussian Prior

To warm up... Suppose we model an observed vector b as having a multivariate Gaussian distribution with known covariance matrix B and unknown mean x. We give x a multivariate Gaussian prior with known covariance matrix A and known mean a.

The posterior distribution of x will be Gaussian, since the product of the prior density and the likelihood is proportional to the exponential of a quadratic function of x:

$$\text{Prior} \times \text{Likelihood} \quad \propto \quad \exp(-(x-a)^T A^{-1}(x-a)/2)\ \exp(-(b-x)^T B^{-1}(b-x)/2)$$

The log posterior density is this quadratic function ($\cdots$ is parts not involving x):

$$\begin{aligned} & -\tfrac{1}{2}\Big[(x-a)^T A^{-1}(x-a) \;+\; (b-x)^T B^{-1}(b-x)\Big] \;+\; \cdots \\ &= \; -\tfrac{1}{2}\Big[x^T(A^{-1}+B^{-1})x \;-\; 2x^T(A^{-1}a+B^{-1}b)\Big] \;+\; \cdots \\ &= \; -\tfrac{1}{2}\Big[(x-c)^T(A^{-1}+B^{-1})(x-c)\Big] \;+\; \cdots \end{aligned}$$

where $c = (A^{-1}+B^{-1})^{-1}\,(A^{-1}a+B^{-1}b)$. This is the density for a Gaussian distribution with mean c and variance $(A^{-1}+B^{-1})^{-1}$.

Posterior for Linear Basis Function Model

Both the log prior and the log likelihood are quadratic functions of β. The log likelihood for β is

$$-\frac{1}{2}\Big[(y-\Phi\beta)^T(\sigma^2 I)^{-1}(y-\Phi\beta)\Big] \;+\; \cdots \;\;=\;\; -\frac{1}{2}\frac{1}{\sigma^2}\Big[\beta^T\Phi^T\Phi\beta \;-\; 2\beta^T\Phi^T y\Big] \;+\; \cdots$$

which is the same quadratic function of β as for a Gaussian log density with covariance $\sigma^2(\Phi^T\Phi)^{-1}$ and mean $(\Phi^T\Phi)^{-1}\Phi^T y$.

This combines with the prior for β in the same way on the previous slide, with the result that the posterior distribution for β is Gaussian with covariance

$$S_n \;=\; \Big[S_0^{-1} \,+\, (\sigma^2(\Phi^T\Phi)^{-1})^{-1}\Big]^{-1} \;=\; \Big[S_0^{-1} \,+\, (1/\sigma^2)\Phi^T\Phi\Big]^{-1}$$

and mean

$$\begin{aligned} m_n &= (S_n^{-1})^{-1}\Big[S_0^{-1}m_0 \,+\, (1/\sigma^2)\Phi^T\Phi(\Phi^T\Phi)^{-1}\Phi^T y\Big] \\ &= S_n\Big[S_0^{-1}m_0 \,+\, (1/\sigma^2)\Phi^T y\Big] \end{aligned}$$

Predictive Distribution for a Test Case

We can write the response, y, for some new case with inputs x as

$$y \;=\; \phi(x)^T\beta + e$$

where the "noise" e has the $N(0, \sigma^2)$ distribution, independently of β.

Since the posterior distribution for β is $N(m_n,\, S_n)$, the posterior distribution for $\phi(x)^T\beta$ will be $N(\phi(x)^T m_n,\, \phi(x)^T S_n \phi(x))$.

Hence the predictive distribution for y will be $N(\phi(x)^T m_n,\, \phi(x)^T S_n \phi(x) + \sigma^2)$.