Announcements

Assignment two due Friday, Feb 10.

Office hour planned for Friday.

Next week:
1) I am away Wednesday through Friday
2) Thursday class will be taken by Kyle
3) No Friday office hour

Plan for today:
Finish intro to Bayesian statistics
Perhaps start on some modeling concepts

Part 02 preview to be posted soon:

Last time

Recall from last time that we were developing Bayesian estimates for the parameters of a normal distribution from data
(In particular, the Maximum a Posteriori (MAP) estimate).

We handled the case of fixed variance, unknown mean

We observed that a particular form of the prior (for the likelihood) lead to an expression of the same form for the posterior.
(This is a conjugate prior).

If you a conjugate prior is acceptable for your problem, it simplifies things.
(Recall Bayesian updating)

Last time

Posterior for estimating the mean given known variance.

$$p\left(\mu \mid \{x_i\}\right) \propto \underbrace{\exp\left(-\frac{\left(\mu_0-\mu\right)^2}{\sigma_0^2}\right)}_{\text{conjugate prior for the likelihood}} \underbrace{\prod_i \exp\left(-\frac{\left(x_i-\mu\right)^2}{\sigma^2}\right)}_{\text{likelihood}}$$

Last time

$$-\log\left(p\left(\mu \mid \{x_i\}\right)\right) = \frac{\left(\mu_0-\mu\right)^2}{\sigma_0^2} + \sum_i \frac{\left(x_i-\mu\right)^2}{\sigma^2}$$

differentiating and setting derivatives to zero gives

$$\frac{\mu_0}{\sigma_0^2} + \frac{1}{\sigma^2}\sum_i x_i = \frac{\mu}{\sigma_0^2} + \frac{N\mu}{\sigma^2}$$

algebra reveals that

$$\mu_{MAP} = \frac{\dfrac{\mu_0}{\sigma_0^2} + \dfrac{N}{\sigma^2}\mu_{ML}}{\dfrac{1}{\sigma_0^2} + \dfrac{N}{\sigma^2}} = \frac{\dfrac{\mu_0}{\sigma_0^2}}{\dfrac{1}{\sigma_0^2} + \dfrac{N}{\sigma^2}} + \frac{\dfrac{N}{\sigma^2}\mu_{ML}}{\dfrac{1}{\sigma_0^2} + \dfrac{N}{\sigma^2}} = \frac{\sigma^2}{\sigma^2+\sigma_0^2 N}\mu_0 + \frac{N\sigma_0^2}{\sigma^2+N\sigma_0^2}\mu_{ML}$$

Example (from Bishop, PRML)

Unknown variance or mean and variance

Similar stories can be told if the mean is known and the variance is not, or both are unknown. We will only set up the problem to have a look at the conjugate priors.

Simplify things by using the inverse of the covariance matrix which is called the precision matrix.

In the univariate case this is simply: $\lambda = \frac{1}{\sigma^2}$

Example (from Bishop, PRML)

Known mean, unknown variance

$$p(\{x_i\}|\lambda) = \prod_{i=1}^{N} \mathbb{N}\left(x_i | \mu, 1/\lambda\right)$$

$$= \prod_{i=1}^{N} \left\{ \left(\frac{\lambda}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\lambda}{2}(x_i-\mu)^2\right) \right\}$$

(u is constant)

$$\propto \lambda^{N/2} \exp\left\{ -\frac{\lambda}{2} \boxed{\sum_i (x_i-\mu)^2} \right\}$$

constant

Example (from Bishop, PRML)

Known mean, unknown variance

$$p(\{x_i\}|\lambda) = \prod_{i=1}^{N} \mathbb{N}\left(x_i | \mu, 1/\lambda\right)$$

$$= \prod_{i=1}^{N} \left\{ \left(\frac{\lambda}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\lambda}{2}(x_i-\mu)^2\right) \right\}$$

(u is constant)

$$\propto \lambda^{N/2} \exp\left\{ -\frac{\lambda}{2} \boxed{\sum_i (x_i-\mu)^2} \right\}$$

constant

Inspection reveals that multiplying this by a gamma distribution

$$\mathrm{Gam}(\lambda|a,b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)$$

yields a posterior of the same form. The normalization constant, $\Gamma(a)$ is the "gamma" function, which extends the concept of factorial to real numbers. $\Gamma(n) = (n-1)!$, for postive integers n. Also $\Gamma(x+1) = x\Gamma(x)$ for postive reals.

“Inspection”

$$p(\{x_i\}|\lambda) \propto \lambda^{N/2} \exp\left\{ -\frac{\lambda}{2} \sum_i (x_i-\mu)^2 \right\} = \lambda^{N/2} \exp\left\{ -\frac{\lambda}{2} K \right\}$$

$$\mathrm{Gam}(\lambda|a,b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda) \propto \lambda^{a-1} \exp(-b\lambda)$$

$$p(\{x_i\}|\lambda)\mathrm{Gam}(\lambda|a,b) \propto \lambda^{N/2} \lambda^{a-1} \exp\left\{ -\frac{\lambda}{2} K \right\} \exp\{-b\lambda\}$$

$$= \lambda^{((N/2)+a-1)} \exp\left\{ -\lambda\left(\left(K/2\right)+b\right) \right\}$$

Gamma distribution illustrated (*)

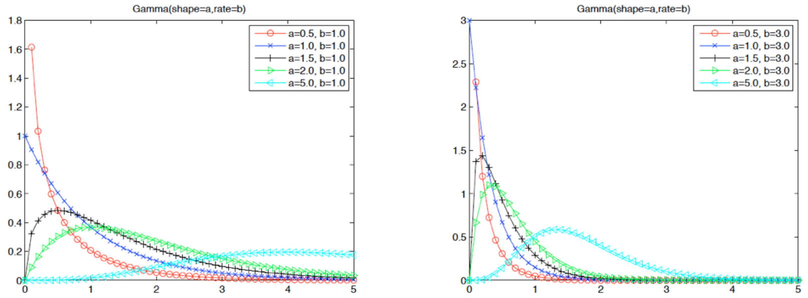


Figure 1: Some $Ga(a,b)$ distributions. If $a<1$, the peak is at 0. As we increase b, we squeeze everything leftwards and upwards. Figures generated by `gammaDistPlot2`.

* From an on-line note by Kevin Murphy (*www.cs.ubc.ca/~murphyk/Teaching/CS340-Fall07/reading/NG.pdf*)

Example (from Bishop, PRML)

Unknown mean and variance

$$p(u,\lambda)=p(u\,|\,\lambda)p(\lambda)$$
$$=N\left(u|u_o,(\beta\lambda)^{-1}\right)Gam(\lambda\,|\,a,b)$$

Here a,b,β are constants. This is the normal-gamma (Gaussian-gamma) distribution.

(Derivation follows for completeness)

◆◆

Example (from Bishop, PRML)

Unknown mean and variance

Indicates optional material

$$p(\{x_i\}\,|\,\lambda)=\prod_{i=1}^{N}\mathbb{N}\left(x_i\,|\,\mu,{}^{1}\!/\!_{\lambda}\right)$$
$$=\prod_{i=1}^{N}\left\{\left(\frac{\lambda}{2\pi}\right)^{\frac{1}{2}}\exp\left(-\frac{\lambda}{2}(x_i-\mu)^2\right)\right\}$$
$$\propto\lambda^{N/2}\exp\left\{-\frac{\lambda}{2}\sum_i(x_i-\mu)^2\right\}$$
$$=\lambda^{N/2}\exp\left\{-\frac{\lambda}{2}\sum_i x_i^2+\lambda\mu\sum_i x_i-\frac{N\lambda}{2}\mu^2\right\}$$

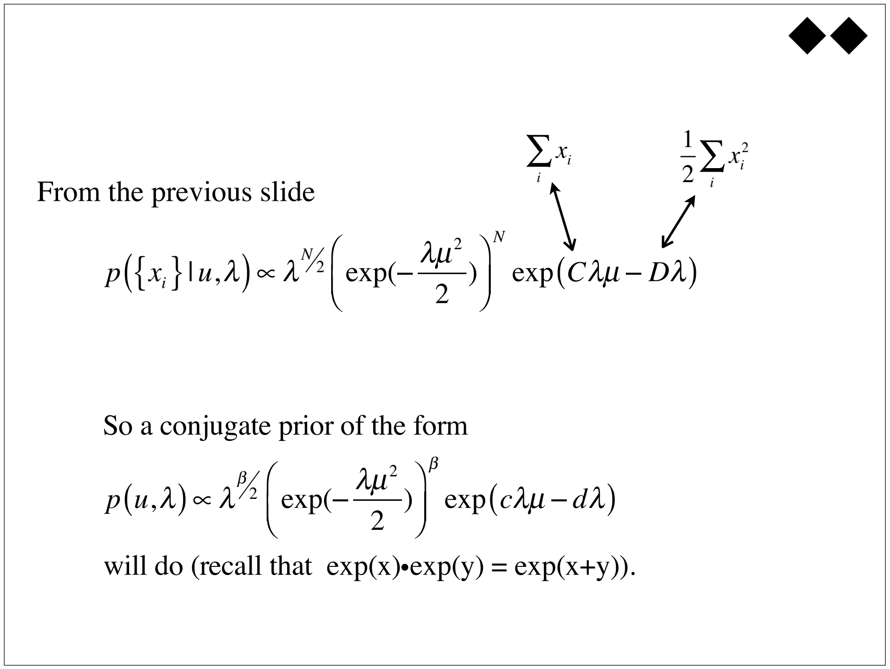
(u is variable)

◆◆

$$p(\{x_i\}\,|\,u,\lambda)\propto\lambda^{N/2}\exp\left\{-\frac{\lambda}{2}\sum_i x_i^2+\lambda\mu\sum_i x_i-\frac{N\lambda}{2}\mu^2\right\}$$

$$=\lambda^{N/2}\left(\exp(-\frac{\lambda\mu^2}{2})\right)^N\exp\left\{\lambda\mu\sum_i x_i-\frac{\lambda}{2}\sum_i x_i^2\right\}$$

$$=\lambda^{N/2}\left(\exp(-\frac{\lambda\mu^2}{2})\right)^N\exp\left(C\lambda\mu-D\lambda\right)$$



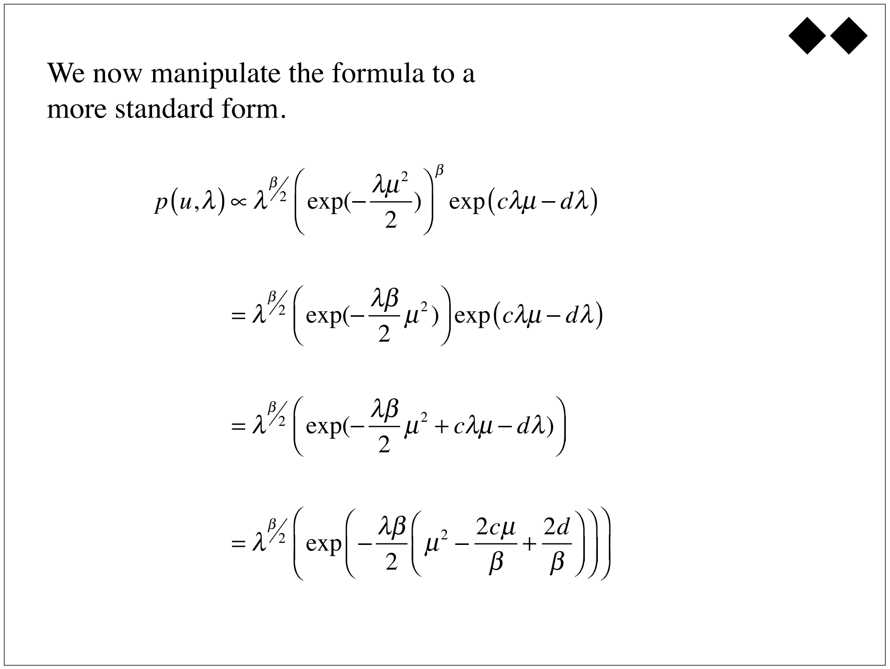
From the previous slide

$$p(\{x_i\}|u,\lambda) \propto \lambda^{N/2}\left(\exp(-\frac{\lambda\mu^2}{2})\right)^N \exp(\overset{\sum_i x_i}{C}\lambda\mu - \overset{\frac{1}{2}\sum_i x_i^2}{D}\lambda)$$

So a conjugate prior of the form

$$p(u,\lambda) \propto \lambda^{\beta/2}\left(\exp(-\frac{\lambda\mu^2}{2})\right)^\beta \exp(c\lambda\mu - d\lambda)$$

will do (recall that exp(x)•exp(y) = exp(x+y)).



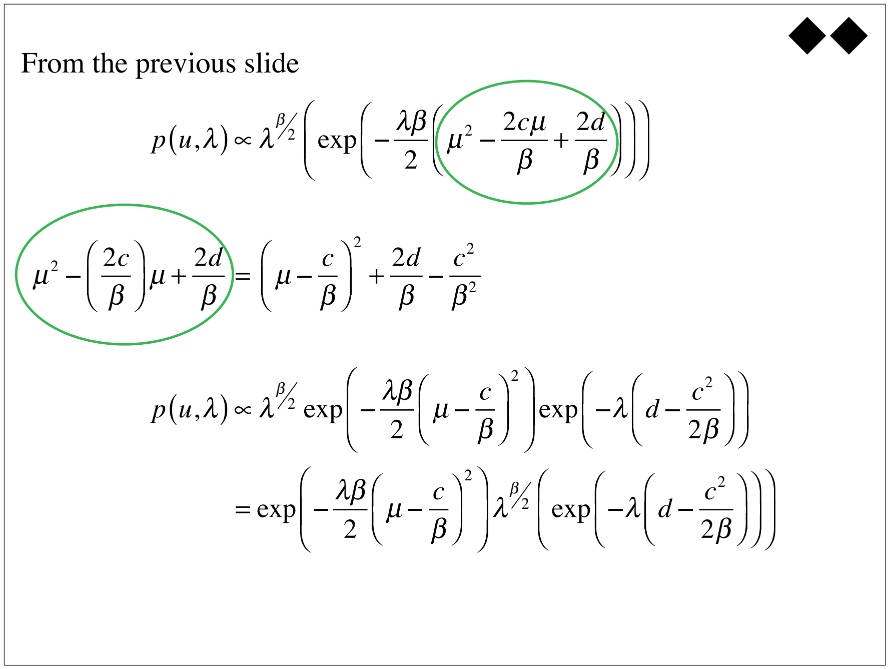
We now manipulate the formula to a more standard form.

$$p(u,\lambda) \propto \lambda^{\beta/2}\left(\exp(-\frac{\lambda\mu^2}{2})\right)^\beta \exp(c\lambda\mu - d\lambda)$$

$$= \lambda^{\beta/2}\left(\exp(-\frac{\lambda\beta}{2}\mu^2)\right)\exp(c\lambda\mu - d\lambda)$$

$$= \lambda^{\beta/2}\left(\exp(-\frac{\lambda\beta}{2}\mu^2 + c\lambda\mu - d\lambda)\right)$$

$$= \lambda^{\beta/2}\left(\exp\left(-\frac{\lambda\beta}{2}\left(\mu^2 - \frac{2c\mu}{\beta} + \frac{2d}{\beta}\right)\right)\right)$$



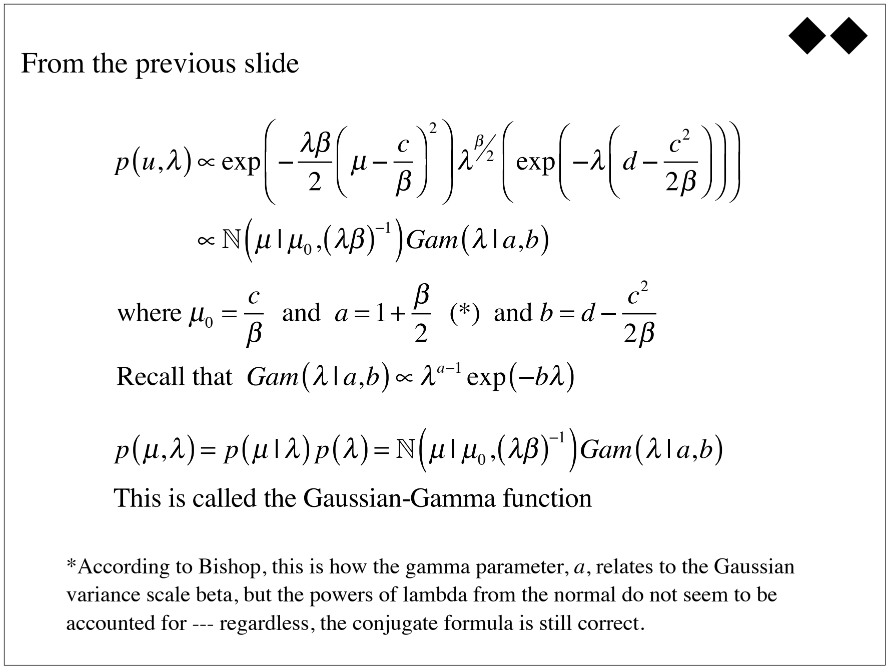
From the previous slide

$$p(u,\lambda) \propto \lambda^{\beta/2}\left(\exp\left(-\frac{\lambda\beta}{2}\left(\mu^2 - \frac{2c\mu}{\beta} + \frac{2d}{\beta}\right)\right)\right)$$

$$\mu^2 - \left(\frac{2c}{\beta}\right)\mu + \frac{2d}{\beta} = \left(\mu - \frac{c}{\beta}\right)^2 + \frac{2d}{\beta} - \frac{c^2}{\beta^2}$$

$$p(u,\lambda) \propto \lambda^{\beta/2}\exp\left(-\frac{\lambda\beta}{2}\left(\mu - \frac{c}{\beta}\right)^2\right)\exp\left(-\lambda\left(d - \frac{c^2}{2\beta}\right)\right)$$

$$= \exp\left(-\frac{\lambda\beta}{2}\left(\mu - \frac{c}{\beta}\right)^2\right)\lambda^{\beta/2}\left(\exp\left(-\lambda\left(d - \frac{c^2}{2\beta}\right)\right)\right)$$



From the previous slide

$$p(u,\lambda) \propto \exp\left(-\frac{\lambda\beta}{2}\left(\mu - \frac{c}{\beta}\right)^2\right)\lambda^{\beta/2}\left(\exp\left(-\lambda\left(d - \frac{c^2}{2\beta}\right)\right)\right)$$

$$\propto \mathbb{N}\left(\mu \,|\, \mu_0, (\lambda\beta)^{-1}\right) Gam(\lambda \,|\, a, b)$$

where $\mu_0 = \frac{c}{\beta}$ and $a = 1 + \frac{\beta}{2}$ (*) and $b = d - \frac{c^2}{2\beta}$

Recall that $Gam(\lambda \,|\, a, b) \propto \lambda^{a-1}\exp(-b\lambda)$

$$p(\mu,\lambda) = p(\mu \,|\, \lambda)\, p(\lambda) = \mathbb{N}\left(\mu \,|\, \mu_0, (\lambda\beta)^{-1}\right) Gam(\lambda \,|\, a, b)$$

This is called the Gaussian-Gamma function

*According to Bishop, this is how the gamma parameter, a, relates to the Gaussian variance scale beta, but the powers of lambda from the normal do not seem to be accounted for --- regardless, the conjugate formula is still correct.

Beta (and Dirichlet) distributions

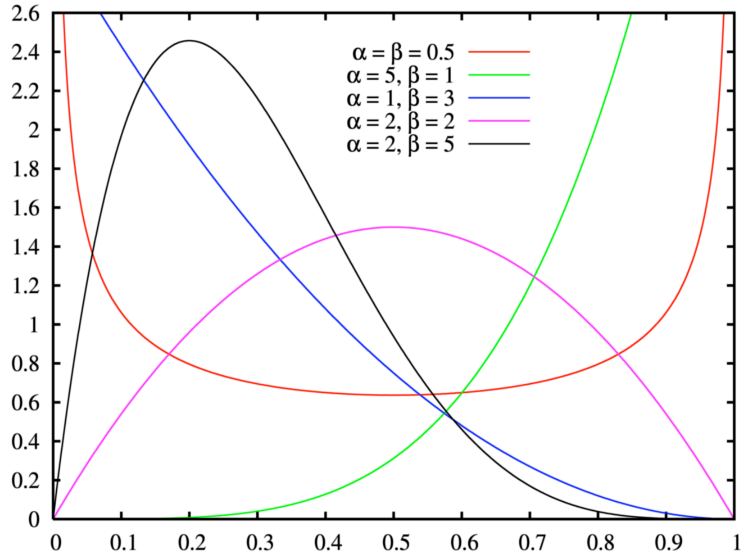
Beta (binary case)

Conjugate prior for the Bernoulli and binomial distributions

Dirichlet (multi-outcome case)

Conjugate priors for the multi outcome Bernoulli and multinomial distributions

$$Beta(u \mid a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} u^{a-1} (1-u)^{b-1}$$



$$Beta(u \mid a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} u^{a-1} (1-u)^{b-1}$$

$$Bern(x \mid \mu) = \mu^{x} (1-\mu)^{(1-x)}$$

(You should be able to tell the rest of the story ...)

More on priors

If we leave off the prior, then we are completely ignorant.

Note that the prior might be the uniform distribution over all numbers

This is not a PDF!

Such priors are called improper.

A more interesting example is p(k)=1/k.

Everything can work out fine if the posterior is a PDF.

Bayesian Sequential Update

- For independent sequential events

$$p\left(\theta \mid D_{1:N}\right) = \left\{ p\left(\theta\right) \prod_{i=1}^{N-1} \left(p\left(D_i \mid \theta\right) \right) \right\} p\left(D_N \mid \theta\right)$$

New prior

Already introduced with the example for the Bayesian estimate of the mean

Predictive Distribution

- The general predictive distribution marginalizes over uncertain model parameters

$$p\left(x \mid X\right) = \int p(x \mid \theta) p(\theta \mid X) d\theta$$

Test data

Training data

Bayesian statistics summary

- Bayesian statistical models
  - We prefer generative models for likelihood (and prior)
  - Conjugate priors are useful
  - Bayesian updating for independent sequence of data
- Inference uses Bayes rule to “invert” the forward model
- Predictive distribution
  - Marginalizes out uncertainty about models
- Related topics coming soon
  - Model selection
  - Decision making