Announcements

Plan for today is to continue with basic Bayesian statistics (nearly done).

Review from last lecture

Estimating the mean of a univariate Gaussian

(Case where variance is known)

$$\mu_{ML} = \frac{1}{N}\sum_i x_i$$

The appropriate conjugate prior for the mean is also Gaussian.

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

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 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

$$
\begin{aligned}
p(\{x_i\} \mid \lambda) &= \prod_{i=1}^{N} \mathbb{N}\left(x_i \mid \mu, 1/\lambda\right) \\
&= \prod_{i=1}^{N} \left\{ \left(\frac{\lambda}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\lambda}{2}(x_i - \mu)^2\right) \right\} \qquad \text{(u is constant)} \\
&\propto \lambda^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_i (x_i - \mu)^2\right\}
\end{aligned}
$$

constant

Inspection reveals that multiplying this by a gamma distribution

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

yields a posterior of the same form.

Example (from Bishop, PRML)

Unknown mean and variance

Indicates optional material

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

(u is variable)

Example (from Bishop, PRML)

Quick review of exponentiation

$$e^{(a+b)} = e^a \cdot e^b$$

$$e^{a \cdot b} = \left(e^a\right)^b$$

$$e^{-a} = 1/e^a$$

Example (from Bishop, PRML)

Quick review of exponentiation

$$\frac{e^a}{e^b} = e^a \bullet e^{-b} = e^{(a-b)}$$

$$e^{\left(\frac{a}{b}\right)} = e^{a \bullet \left(\frac{1}{b}\right)} = \left(e^a\right)^{\left(\frac{1}{b}\right)}$$

Example (from Bishop, PRML)

Unknown mean and variance

Indicates optional material

$$\begin{aligned} p\left(\{x_i\} \mid \lambda\right) &= \prod_{i=1}^{N} \mathbb{N}\left(x_i \mid \mu, 1/\lambda\right) \\ &= \prod_{i=1}^{N} \left\{ \left(\frac{\lambda}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\lambda}{2}\left(x_i - \mu\right)^2\right) \right\} \\ &\propto \lambda^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_i \left(x_i - \mu\right)^2\right\} \\ &= \lambda^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_i x_i^2 + \lambda\mu\sum_i x_i - \frac{N\lambda}{2}\mu^2\right\} \end{aligned}$$

(u is variable)

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

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

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

From the previous slide

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

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

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

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

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

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

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

This is called the Gaussian-Gamma function

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

Unknown mean and variance

To summarize, our conjugate prior is given by

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

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

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 which is not a PDF! (Why?)

Such priors are called improper.

A more interesting example is p(k)=1/k, integral k>0.

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

Bayesian Sequential Update

For independent (conditioned on the model) sequential events

Suppose after N-1 observations we have the posterior $p\left(\theta \middle| D_{1:N-1}\right)$

This is a natural prior to estimate the posterior $p\left(\theta \middle| D_{1:N}\right)$ from one more observation, D_N.

$p\left(\theta \mid D_{1:N}\right) \propto p\left(\theta \middle| D_{1:N-1}\right)p\left(D_N \mid \theta\right)$ (but is it true?)

Bayesian Sequential Update

More formally, for independent sequential events

$$
\begin{aligned}
p(\theta \mid D_{1:N}) &\propto p(\theta)\, p(D_{1:N} \mid \theta) \\
&= p(\theta) \prod_{i=1}^{N} \big(p(D_i \mid \theta) \big) \\
&= \left\{ p(\theta) \prod_{i=1}^{N-1} \big(p(D_i \mid \theta) \big) \right\} p(D_N \mid \theta) \\
&\propto p(\theta \mid D_{1:N-1})\, p(D_N \mid \theta)
\end{aligned}
$$

Prior from first
N-1 observations

Predictive Distribution

Example --- you are tracking points following a curve in time

Question --- where do you think the next point will be?

Your observations up to time N-1 are “training data” (learn the curve).

What is the density for the location of the next point?

Predictive Distribution

Given training observations, X.

What is the density for another observation, x?

If we know the model parameters then we have:

$$p(x \mid \theta)$$

Predictive Distribution

If we know the model parameters then we have:

$$p(x \mid \theta)$$

However, we have been developing methods to compute posterior distributions:

$$p(\theta \mid X)$$

Predictive Distribution

- In the most general case, the 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

Model Selection

- Model selection refers to choosing among different instances within a model class (1) or different model classes (2).
- Examples:
  - The number of clusters (1)
  - The degree of a polynomial to fit a curve to data (1)
  - Polynomials versus other basis functions such as Fourier (2)

Model Selection

- There is no real difference between choosing the parameters of a model, which model class instance, and which model class.
- Difficulties
  - Prior densities are of different dimensionality
  - Constructing priors and likelihoods that properly penalize model complexity
    - Standard penalties include AIC, BIC
      - Cross-validation over novel data is more reliable.