Announcements

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

Estimating the mean of a univariate Gaussian

\[ \mu_{ML} = \frac{1}{N} \sum_{i} x_i \]

The appropriate conjugate prior for the mean is also Gaussian.

\[ \mu_{MAP} = \frac{\sigma^2 + \frac{1}{2} \sum_{i} (x_i - \mu)^2}{\sigma^2 + \frac{1}{2} \sum_{i} (x_i - \mu)^2} \]

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

\[ p(x_i | \lambda, \mu) = \prod_{i}^N \left( \frac{1}{\Gamma(a)} \right) \left( \frac{1}{2} \right)^a \exp \left( - \frac{1}{2} \sum (x_i - \mu)^2 \right) \]

\( (a \text{ is constant}) \)

Inspection reveals that multiplying this by a gamma distribution

\[ \text{Gam}(\lambda, a, b) = \left( \frac{1}{\Gamma(a)} \right) \lambda^{-a} \exp(-\lambda b) \]

yields a posterior of the same form.
Unknown mean and variance

Example (from Bishop, PRML)

\( p(x|\mu, \sigma^2) = \prod p(x_i|\mu, \sigma^2) \) (\( u \) is variable)

Indicates optional material

Quick review of exponentiation

\[ e^{a+b} = e^a \cdot e^b \]
\[ e^{ab} = (e^a)^b \]
\[ e^{-a} = \frac{1}{e^a} \]

Unknown mean and variance

Example (from Bishop, PRML)

\[ p(x|\mu, \sigma^2) = \prod p(x_i|\mu, \sigma^2) \]
\[ = \prod \left( \frac{1}{\sqrt{2\pi}\sigma} \exp \left( -\frac{1}{2}(x_i - \mu)^2 \right) \right) \]
\[ = \lambda \sqrt{\frac{\sigma}{2\pi}} \exp \left( -\frac{1}{2} \sum (x_i - \mu)^2 \right) \]
\[ = \lambda \sqrt{\frac{\sigma}{2\pi}} \exp \left( -\frac{1}{2} \sum x_i^2 + \mu \right) \] (\( u \) is variable)
We now manipulate the formula to a more standard form.

\[
p((x_i|n, \lambda)) = \lambda^n \exp\left\{ \frac{1}{2} \sum c_i + n \lambda c_i \right\} \\
= \lambda^n \exp\left\{ \frac{1}{2} \lambda \right\} \exp\left\{ n \lambda c_i \right\} \\
= \lambda^n \exp\left\{ \frac{1}{2} \lambda \right\} \exp(C\lambda \mu - D\lambda)
\]

From the previous slide

\[
p((x_i|n, \lambda)) = \lambda^n \exp\left\{ \frac{1}{2} \lambda \right\} \exp(C\lambda \mu - D\lambda)
\]

So a conjugate prior of the form

\[
p(\lambda, \mu) = \lambda^n \exp\left\{ \frac{1}{2} \lambda \right\} \exp(C\lambda \mu - D\lambda)
\]

will do (recall that \( \exp(a)\exp(b) = \exp(a+b) \)).
From the previous slide

\[ p(\alpha, \lambda) = \exp \left( \frac{-\lambda}{2} \left( x - \frac{1}{\beta} \right)^2 \right) \frac{\lambda^{\alpha-1}}{\Gamma(\alpha)} \exp \left( -\frac{\lambda}{2\beta} \right) \]

= \text{Gam}(\lambda; a, b)

where \( \bar{\mu} = \frac{1}{\beta} \) and \( a = 1 + \frac{\lambda}{2} \) and \( b = \frac{\lambda}{2\beta} \)

Recall that \( \text{Gam}(\lambda; a, b) = \lambda^{a-1} \exp(-\lambda b) \)

\[ p(\mu, \lambda) = p(\mu | \lambda)p(\lambda) = \text{Gam}(\lambda; a, b) \]

This is called the Gaussian-Gamma function

*According to Bishop, this is how the gamma parameter, \( \alpha \), 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.

**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 \( p(k) = k^{-1} \), integral \( k \geq 0 \).

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

**Unknown mean and variance**

To summarize, our conjugate prior is given by

\[ p(\mu, \lambda) = p(\mu | \lambda)p(\lambda) = \text{Gam}(\lambda; a, b) \]

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

**Bayesian Sequential Update**

For independent (conditioned on the model) sequential events

Suppose after \( N-1 \) observations we have the posterior \( p(\theta | D_{1..N-1}) \)

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

\[ p(\theta | D_{1..N}) = \frac{p(\theta | D_{1..N-1}) p(\theta | D_N)}{p(D_N | D_{1..N-1})} \] (but is it true?)
Bayesian Sequential Update

More formally, for independent sequential events

\[ p(\theta | D_1) = p(\theta) \cdot p(D_1 | \theta) \]

\[ = p(\theta) \prod_i p(D_i | \theta) \]

\[ = \left[ p(\theta) \prod_i p(D_i | \theta) \right] \cdot p(D_1 | \theta) \]

\[ = p(\theta) \cdot p(D_1 | \theta) \]

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:

\[ p(x | \theta) \]

Predictive Distribution

If we know the model parameters then we have:

\[ p(x | \theta) \]

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

\[ p(\theta | X) \]
Predictive Distribution

- In the most general case, the predictive distribution marginalizes over uncertain model parameters

\[ p(x|X) = \int p(x|\theta)p(\theta|X)d\theta \]

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)

- 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.