Machine Learning 10-701 Feb 16, 2015
http://alex.smola.org/teaching/10-701-15 due on Feb 23, 2015
Carnegie Mellon University

Homework 5

START HERE: Instructions

- The homework is due at 9:00am on Feb 23, 2015. Anything that is received after that time will not be considered.
- Answers to every theory questions will be also submitted electronically on Autolab (PDF: Latex or handwritten and scanned). Make sure you prepare the answers to each question separately.
- Collaboration on solving the homework is allowed (after you have thought about the problems on your own). However, when you do collaborate, you should list your collaborators! You might also have gotten some inspiration from resources (books or online etc...). This might be OK only after you have tried to solve the problem, and couldn't. In such a case, you should cite your resources.
- If you do collaborate with someone or use a book or website, you are expected to write up your solution independently. That is, close the book and all of your notes before starting to write up your solution.
- Latex source of this homework: http://alex.smola.org/teaching/10-701-15/homework/hw5_latex.tar

1 Exponential Family [Zhou, Manzil]

In this problem we will review the exponential family, its significance in Bayesian statistics and work out a detailed example for the commonly encountered Multivariate Normal distribution and its conjugate prior Normal Inverse Wishart Distribution.

1.1 Review

Exponential family is a set of probability distributions whose probability density function for $x \in \mathbb{R}^d$ can be expressed in the form:

$$p(x|\theta) = \exp(\langle \phi(x), \theta \rangle - \mathbb{1}^{\mathrm{T}} g(\theta)) \tag{1}$$

where

- $\phi(x)$ is a sufficient statistic of the distribution. For exponential families, the sufficient statistic is a function of the data that fully summarizes the data x within the density function. The sufficient statistic of a set of independent identically distributed data observations is simply the sum of individual sufficient statistics, and encapsulates all the information needed to describe the posterior distribution of the parameters, given the data and hence to derive any desired estimate of the parameters. *We will explore this important property in detail below.*
- θ is called the natural parameter. The set of values of θ for which the function $p(x|\theta)$ is finite is called the natural parameter space. It can be shown that the natural parameter space is always convex (How??).
- $g(\theta)$ is called the log-partition function because it is the logarithm of a normalization factor, without which $p(x|\theta)$ would not be a probability distribution ("partition function" is often used as a synonym of "normalization factor" for historical reasons arising from Statistical Physics).

1.2 Conjugate Priors [10+10+10]

Exponential families are very important in Bayesian statistics. In Bayesian statistics a prior distribution is multiplied by a likelihood function and then normalised to produce a posterior distribution. In the case of

a likelihood which belongs to the exponential family there always exists a conjugate prior, which is also in the exponential family.

Consider the distribution:

$$p(\theta; m_0, \phi_0) = \exp\left(\langle \phi_0, \theta \rangle - \langle m_0, g(\theta) \rangle - h(m_0, \phi_0)\right) \tag{2}$$

where $m_0 > 0$ and $\phi_0 \in \mathbb{R}^d$. These are called hyperparameters (parameters controlling parameters).

Question 1

Show that this distribution, i.e. (2) is a member of the Exponential Family.

Suppose we obtain the independent and identically distributed data $X = (x_1, \ldots, x_n)$, where $x_i \sim p(\cdot|\theta)$, i.e. each single observation follows some distribution from the exponential family.

Question 2

First of all write out the likelihood $p(X|\theta)$. Then use (2) as the prior and derive the posterior $p(\theta|X)$ exactly, i.e. with proper normalization constant.

If you got Question 2 correct (hopefully you did), observe that the posterior has the same form as the prior, thus (2) is a conjugate prior. The difference between the prior, i.e. (2) and your answer to Question 2 lies only in the parameters.

Question 3

Let m_n and ϕ_n be parameters of the posterior $p(\theta|X)$, then show that:

$$\begin{aligned} m_n &= m_0 + n\mathbb{1} \\ \phi_n &= \phi_0 + \sum_{i=1}^{n} \phi(x_i) \end{aligned} \tag{3}$$

We call this update equations.

This shows that the update equations can be written simply in terms of the number of data points and the sufficient statistic of the data. Also, it provides meaning to the hyperparameters. In particular, m_0 corresponds to the effective number of "fake" observations that the prior distribution contributes, and ϕ_0 corresponds to the total amount that these "fake" observations contribute to the sufficient statistic over all observations and "fake" observations.

Various examples of update equations can seen in the conjugate prior page of Wikipedia http://en.wikipedia.org/wiki/Conjugate_prior. Next, we will see the case of Multivariate Normal Distribution in detail.

1.3 Multivariate Normal Distribution [10+20+20]

The Multivariate Normal $\mathcal{N}(\mu, \Sigma)$ is a distribution that is encountered very often. The distribution is given by:

$$p(x|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\right) \tag{4}$$

where $\mu \in \mathbb{R}^d$ and $\Sigma \succ 0$ is a symmetric positive definite $d \times d$ matrix. We claim that it belongs to the Exponential Family.

Question 4

Identify the natural parameters θ in terms of μ and Σ. Also derive the sufficient statistics $\phi(x)$ and log partition function $g(\theta)$ in terms of μ and Σ. *Hint: Design a two dimensional $g(\theta)$, where first dimension is $\frac{1}{2}\mu^{\mathrm{T}}\Sigma^{-1}\mu$.*

The conjugate prior for Multivariate Normal Distribution can be parametrized as the Normal Inverse Wishart Distribution $\mathcal{NIW}(\mu_0, \kappa_0, \Sigma_0, \nu_0)$. The distribution is given by:

$$\begin{aligned} p(\mu, \Sigma;\ \mu_0, \kappa_0, \Sigma_0, \nu_0) &= \mathcal{N}(\mu|\mu_0, \Sigma/\kappa_0)\mathcal{W}^{-1}(\Sigma|\Sigma_0, \nu_0) \\ &= \frac{\kappa_0^{\frac{d}{2}} |\Sigma_0|^{\frac{\nu_0}{2}} |\Sigma|^{-\frac{\nu_0+d+2}{2}}}{2^{\frac{(\nu_0+1)d}{2}} \pi^{\frac{d}{2}} \Gamma_d(\frac{\nu_0}{2})} e^{-\frac{\kappa_0}{2}(\mu-\mu_0)^{\mathrm{T}}\Sigma^{-1}(\mu-\mu_0) - \frac{1}{2}\operatorname{tr}(\Sigma_0\Sigma^{-1})} \end{aligned} \tag{5}$$

where $\kappa_0, \nu_0 > 0$, $\mu_0 \in \mathbb{R}^d$ and $\Sigma_0 \succ 0$ is a symmetric positive definite $d \times d$ matrix., and $\Gamma_d(\cdot)$ is the multivariate gamma function.

Question 5

Notice that Normal Inverse Wishart Distribution will fit into the form of (2). Find the mapping between $(\mu_0, \kappa_0, \Sigma_0, \nu_0)$ and (m_0, ϕ_0) and the function $h(m_0, \phi_0)$ in terms of $(\mu_0, \kappa_0, \Sigma_0, \nu_0)$. *Hint: m_0 and $g(\theta)$ is two dimensional.*

Equipped with these, results we move on to tackle the problem of finding posterior for μ, Σ. One can follow brute force approach to find it be using (4) and (5), but things can get really messy. We will adopt a more elegant and easy approach exploiting the fact that these distribution belong to the exponential family.

Question 6

Using the update equations described in Question 3 and your answers to Question 4 and 5, directly write down the posterior for $p(\mu, \Sigma|X)$. (Just providing appropriate update equations would suffice.)

This is one of the remarkable cases where working out the general case saves you effort than working out the special case! The algebra can become very complicated, e.g. see http://www.cs.ubc.ca/~murphyk/Papers/bayesGauss.pdf where they have explicitly done complicated math! We hope that after solving this homework, you can take advantage of this neat short-cut :)

1.4 Posterior Predictive - Bonus [10+10]

Another quantify, which is often of interest in Bayesian Statistics, is the posterior predictive. The posterior predictive distribution is the distribution of unobserved observations (prediction) conditional on the observed data. Specifically, it is computed by marginalising over the parameters, using the posterior distribution:

$$p(\tilde{x}|X) = \int p(\tilde{x}|\theta)\, p(\theta|\mathbf{X})\, \mathrm{d}\theta \tag{6}$$

The posterior predictive distribution for a distribution in exponential family has a rather nice form.

Question 7

Show that the posterior predictive for the distribution (1) with prior (2) is given by:

$$p(\tilde{x}|X) = \exp\left\{h\left(m_n + 1, \phi_n + \phi(\tilde{x})\right) - h\left(m_n, \phi_n\right)\right\} \tag{7}$$

The result of previous problem can be specialized for the Multivariate Normal case.

Question 8

Find the predictive posterior for the case of Multivariate Normal Distribution with Normal Inverse Wishart Distribution, having parameters as described in 1.3 by using Question 7. *Hint: The matrix determinant lemma might come handy* http://en.wikipedia.org/wiki/Matrix_determinant_lemma.

For this problem set, feel free to refer to any Wikipedia page (e.g. http://en.wikipedia.org/wiki/Exponential_family or the other distribution pages). The formulation here maybe slightly different from that in Wikipedia.