MATH 390.03-02 / 650 Spring 2016 Homework #9

Professor Adam Kapelner

Due *in class*, May 16, 2016

(this document last updated Wednesday 11th May, 2016 at 5:55pm)

Instructions and Philosophy

The path to success in this class is to do many problems. Unlike other courses, exclusively doing reading(s) will not help. Coming to lecture is akin to watching workout videos; thinking about and solving problems on your own is the actual "working out." Feel free to "work out" with others; **I want you to work on this in groups.**

Reading is still *required*. For this homework set, read about ridge regression and Gibbs sampling. Also read ch15 and ch16 in McGrayne.

The problems below are color coded: green problems are considered *easy* and marked "[easy]"; yellow problems are considered *intermediate* and marked "[harder]", red problems are considered *difficult* and marked "[difficult]" and purple problems are extra credit. The *easy* problems are intended to be "giveaways" if you went to class. Do as much as you can of the others; I expect you to at least attempt the *difficult* problems.

Problems marked "[MA]" are for the masters students only (those enrolled in the 650 course). For those in 390, doing these questions will count as extra credit.

This homework is worth 100 points but the point distribution will not be determined until after the due date. See syllabus for the policy on late homework.

Up to 10 points are given as a bonus if the homework is typed using LaTeX. Links to instaling LaTeX and program for compiling LaTeX is found on the syllabus. You are encouraged to use `overleaf.com`. If you are handing in homework this way, read the comments in the code; there are two lines to comment out and you should replace my name with yours and write your section. The easiest way to use overleaf is to copy the raw text from hwxx.tex and preamble.tex into two new overleaf tex files with the same name. If you are asked to make drawings, you can take a picture of your handwritten drawing and insert them as figures or leave space using the "\vspace" command and draw them in after printing or attach them stapled.

The document is available with spaces for you to write your answers. If not using LaTeX, print this document and write in your answers. I do not accept homeworks which are *not* on this printout. Keep this first page printed for your records.

NAME: _____

Problem 1

These are questions about McGrayne's book, chapters 15 and 16.

(a) [easy] During the H-Bomb search in Spain and its coastal regions, RAdm. William Guest was busy sending ships here, there and everywhere even if the ships couldn't see the bottom of the ocean. How did Richardson use those useless searches?

(b) [harder] When the Navy was looking for the *Scorpion* submarine, they used Monte Carlo methods (which we will see in class soon). How does the description of these methods by Richardson (p199) remind you of the "sampling" techniques to approximate integrals we did in class?

(c) [harder] What is a Kalman filter? Read about it online and write a few descriptive sentences.

(d) [harder] Where do frequentist methods practically break down? (end of chapter 15)

(e) [easy] What was the main problem facing Bayesian Statistics in the early 1980's?

(f) [harder] What is the "curse of dimensionality?"

(g) [easy] How did Bayesian Statistics help sociologists?

(h) [easy] How did Gibbs sampling come to be?

(i) [easy] Were the Geman brothers the first to discover the Gibbs sampler?

(j) [easy] Who officially discovered the expectation-maximization (EM) algorithm? And who *really* discovered it?

(k) [harder] How did Bayesians "break" the curse of dimensionality?

(l) [harder] Consider the integrals we use in class to find expectations or to approximate PDF's / PMF's — how can they be replaced?

(m) [easy] What did physicists call "Markov Chain Monte Carlo" (MCMC)? (p222)

(n) [easy] Why is sampling called "Monte Carlo" and who named it that?

(o) [easy] The Metropolis-Hastings (MH) Algorithm is world famous and used in myriad applications. Why didn't Hastings get any credit?

(p) [easy] The combination of Bayesian Statistics + MCMC has been called ... (p224)

(q) [E.C.] p225 talks about Thomas Kuhn's ideas of "paradigm shifts." What is a "paradigm shift" and does Bayesian Statistics + MCMC qualify?

(r) [easy] How did the BUGS software change the world?

(s) [easy] Lindley said that Bayesian Statistics would win out over Frequentist Statistics because it was more logical. What in reality was the reason of its eventual victory?

Problem 2

These are questions about the prior being a mixture distribution.

(a) [easy] Let's say you have a prior distribution $\mathbb{P}(\theta)$ which is a mixture of M distributions that you are mixing. Call them, $\mathbb{P}_1(\theta), \mathbb{P}_2(\theta), \ldots, \mathbb{P}_M(\theta)$ and the mixing proportions are $\rho_1, \rho_2, \ldots, \rho_M$. Write the prior below using the formula we discussed in class.

(b) [easy] Is there a restriction on the mixing proportions, $\rho_1, \rho_2, \ldots, \rho_M$? Discuss.

(c) [harder] Why is a mixture distribution sometimes called a "convex combination?" What's "convex" about it?

(d) [easy] Explain how you can use a mixture distribution of betas to approximate any continuous distribution (within reason) with support $[0, 1]$.

(e) [harder] Rederive from the class notes that the posterior is also a mixture distribution. What distributions does it mix? What are the new mixing proportions $(\rho'_1, \rho'_2, \ldots, \rho'_M)$?

(f) [difficult] Derive the posterior predictive distribution.

Problem 3

If the prior is a mixture distribution of conjugate priors, then the math is convenient. We explore a case like this here.

(a) [easy] Let's say you have a sample of heights from people in the U.S. but you don't know if the people are males or females. According to wikipedia, the average male in America is 5' 9.5" or 69.5" and the average female is 5' 4" or 64". Let's design a mixture prior. Assume equal mixing. What is $\mu_{0,M}$ for the males and what is $\mu_{0,F}$ for females? Use inches as the unit to keep things simple.

(b) [easy] If there is no reason to suspect that the sample you have is male or female, what are the mixing proportions, ρ_M and ρ_F?

(c) [easy] Use θ for the mean height in your sample. Assume the standard deviation is $\sigma = 2.8''$. What is the likelihood model?

(d) [harder] According to Wikipedia, there were 895 males and 980 females? Using that information and your answer to (a) and (b) and the information given in (c), construct $\mathbb{P}(\theta)$. This is review for the final as a question similar to this one was on the midterm.

Hint: $\tau_M^2 = \sigma^2/895$ and $\tau_F^2 = \sigma^2/980$.

(e) [difficult] Derive the posterior distribution, $\mathbb{P}(\theta \mid X, \sigma^2 = 2.8"^2)$.

We need to compute $\rho'_M = \rho_M \frac{\mathbb{P}_M(X)}{\mathbb{P}_M(X)+\mathbb{P}_F(X)}$ where:

$$
\begin{aligned}
\mathbb{P}_M(X) &:= \mathbb{P}_M(X_1, \ldots, X_n) \\
&= \int_{\mathbb{R}} \mathbb{P}(X_1, \ldots, X_n \mid \theta)\, \mathbb{P}_M(\theta)\, d\theta \\
&= \int_{\mathbb{R}} \left(\prod_{i=1}^{n} \mathcal{N}(\theta,\, \sigma^2) \right) \mathcal{N}(\mu_M,\, \tau_M^2)\, d\theta \\
&= \int_{\mathbb{R}} \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (x_i - \theta)^2 \right) \right) \frac{1}{\sqrt{2\pi\tau_M^2}} \exp\left(-\frac{1}{2\tau_M^2} (\theta - \mu_M)^2 \right) d\theta \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \int_{\mathbb{R}} \exp\left(-\frac{1}{2\sigma^2} \underbrace{\sum_{i=1}^{n} (x_i - \theta)^2}_{\text{Lec 13 page 5}} \right) \mathcal{N}(\mu_M,\, \tau_M^2)\, d\theta \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \int_{\mathbb{R}} \exp\left(-\frac{1}{2\sigma^2} \left((n-1)s^2 + n(\bar{x} - \theta)^2 \right) \right) \mathcal{N}(\mu_M,\, \tau_M^2)\, d\theta \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-\frac{1}{2\sigma^2}\left((n-1)s^2\right)} \int_{\mathbb{R}} \exp\left(-\frac{1}{2\sigma^2/n} (\bar{x} - \theta)^2 \right) \mathcal{N}(\mu_M,\, \tau_M^2)\, d\theta \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-\frac{1}{2\sigma^2}\left((n-1)s^2\right)} \sqrt{2\pi\sigma^2/n} \times \\
&\quad \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left(-\frac{1}{2\sigma^2/n} (\bar{x} - \theta)^2 \right) \mathcal{N}(\mu_M,\, \tau_M^2)\, d\theta
\end{aligned}
$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\left((n-1)s^2\right)}\sqrt{2\pi\sigma^2/n}\underbrace{\int_{\mathbb{R}}\mathcal{N}\left(\bar{x},\,\sigma^2/n\right)\mathcal{N}\left(\mu_M,\,\tau_M^2\right)d\theta}_{\text{See p9 Lec 12}}$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\left((n-1)s^2\right)}\sqrt{2\pi\sigma^2/n}\,\underbrace{\mathcal{N}\left(\mu_M,\,\sigma^2/n+\tau_M^2\right)}_{\text{free variable: }\bar{x}}$$

Note how everything before the normal density is a constant which cancels out when we compute the posterior weight:

$$\begin{aligned}
\rho'_M &= \rho_M\frac{\mathbb{P}_M(X)}{\mathbb{P}_M(X)+\mathbb{P}_F(X)} \\
&= \frac{1}{2}\frac{\mathcal{N}\left(\mu_M,\,\sigma^2/n+\tau_M^2\right)}{\mathcal{N}\left(\mu_M,\,\sigma^2/n+\tau_M^2\right)+\mathcal{N}\left(\mu_F,\,\sigma^2/n+\tau_F^2\right)} \\
&= \frac{1}{2}\frac{\frac{1}{\sqrt{2\pi\left(\sigma^2/n+\tau_M^2\right)}}e^{-\frac{1}{2\left(\sigma^2/n+\tau_M^2\right)}(\bar{x}-\mu_M)^2}}{\frac{1}{\sqrt{2\pi\left(\sigma^2/n+\tau_M^2\right)}}e^{-\frac{1}{2\left(\sigma^2/n+\tau_M^2\right)}(\bar{x}-\mu_M)^2}+\frac{1}{\sqrt{2\pi\left(\sigma^2/n+\tau_F^2\right)}}e^{-\frac{1}{2\left(\sigma^2/n+\tau_F^2\right)}(\bar{x}-\mu_F)^2}}
\end{aligned}$$

and ρ'_F can be calculated similarly. Your answer for (f) should show that $\rho'_F \approx 100\%$ and $\rho'_M \approx 0\%$.

(f) [easy] You get a sample of $n = 10$ where the heights are:

$X = \{$`62.8, 60.2, 58.0, 62.7, 66.4, 58.6, 64.0, 59.8, 66.7, 62.0`$\}$.

Given your answer to (e), what is the posterior distribution $\mathbb{P}(\theta \mid X, \sigma^2 = 2.8"^2)$ now?

(g) [easy] What is the posterior mixture proportion ρ'_M? Does this make sense given the data?

(h) [difficult] [M.A.] Your answer to (g) was a "best guess" but it was not Bayesian. Let's put a prior on $\rho \sim \text{Beta}(\alpha, \beta)$. This is now known as a "hierarchical Bayesian model" since there's both a prior on the parameters and a prior on the prior called a "hyperprior" with "hyperhyperparameters." Write out the posterior as best as you could below.

Problem 4

We practice using the Newton-Raphson algorithm here using the beta distribution.

(a) [easy] If $\theta \sim \text{Beta}(10, 30)$ then what is the prior expectation and the prior mode? Use the formulas we gave in class before the first midterm.

(b) [easy] State the Newton-Raphson algorithm below.

(c) [difficult] We will make sure this formula for the mode of a beta distribution is correct by using Newton-Raphson. Begin at the expectation and iterate twice. By what percentage are you off by?

(d) [easy] [M.A.] State the generalized Newton-Raphson algorithm (for arbitrary dimension) below.

(e) [easy] [M.A.] In lecture 20 we discussed the likelihood mixture model

$$X_1, \ldots, X_n \overset{exch}{\sim} \rho\, \mathcal{N}\left(\theta_1,\, \sigma_1^2\right) + (1-\rho)\, \mathcal{N}\left(\theta_2,\, \sigma_2^2\right) \text{ where } \boldsymbol{\theta} := \left[\theta_1,\ \sigma_1^2,\ \theta_2,\ \sigma_2^2,\ \rho\right].$$

Explain why it would be a disaster to find $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ via the generalized Newton-Raphson algorithm.

Problem 5

We practice using the E-M algorithm here using an example similar to the lecture.

(a) [easy] What does "data augmentation" mean?

(b) [easy] Imagine we have the likelihood model found in 4(e). If we knew the membership of each X_i denoted by I_i where if $I_i = 1$ it means the ith observation belongs to the first distribution i.e. $\mathcal{N}(\theta_1, \sigma_1^2)$ and if $I_i = 0$ it means it did not belong to the first distribution, it belongs to the second distribution i.e. $\mathcal{N}(\theta_2, \sigma_2^2)$. What is the likelihood now?

(c) [easy] Consider the situation where we are once again trying to estimate mean heights. Here, we are trying to estimate the mean height of males which is denoted θ_1, the mean height of females denoted θ_2 and the variances denoted by σ_1^2 and σ_2^2. If $n = 10$ and you know that the first sample of $n_1 = 5$ heights comes from males: $71.6, 67.0, 70.0, 66.5, 72.2$ (in inches) and the second sample of heights $n_2 = 5$ comes from females: $63.1, 64.3, 59.2, 60.0, 68.3$ (in inches), what are your best estimates of θ_1, θ_2, σ_1^2, σ_2^2 and ρ? This is marked easy for a reason.

(d) [harder] An 11th data point comes in and its height is $x_{11} = 68.1$" but you do not know if it's from a male or a female. Could this new data point help with estimating θ_1 even though you do not know it's a measurement from a male?

(e) [easy] What is the E-M algorthim generally speaking? And is it Bayesian in its general form?

(f) [difficult] Implement the E-M algorithm here by starting with the values from (c) and do two iterations. What are your new values for θ_1 and θ_2?

(g) [difficult] What is the probability the 11th measurement comes from a male?