MATH 390.03-02 / 650 Fall 2015 Homework #2

Professor Adam Kapelner

Due 4PM in my mail slot, Friday, February 19, 2015

(this document last updated Thursday 11th February, 2016 at 1:10pm)

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, review Math 241 concerning random variables, support, parameter space, PMF's, PDF's, CDF's, Bayes Rule, read about parametric families and maximum likelihood estimators on the Internet, read the preface and ch 1 and 4 of Bolstad and read the preface and Ch1 of 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, preface, chapter 1, 2 and 3.

(a) [easy] Explain Hume's problem of induction with the sun rising every day.

(b) [easy] Explain the "inverse probability problem."

(c) [easy] What is Bayes' billiard table problem?

(d) [difficult] [MA] How did Price use Bayes' idea to prove the existence of the deity?

(e) [easy] Why should Bayes Rule really be called "Laplace's Rule?"

(f) [difficult] Prove the version of Bayes Rule found on page 20. State your assumption(s) explicitly. Reference class notes as well.

(g) [easy] Give two scientific contexts where Laplace used inverse probability theory to solve major problems.

(h) [difficult] [MA] Why did Laplace turn into a frequentist later in life?

(i) [easy] State Laplace's version of Bayes Rule (p31).

(j) [easy] Why was Bayes Rule "damned" (pp36-37)?

(k) [easy] According to Edward Molina, what is the prior (p41)?

(l) [easy] What is the source of the "credibility" metric that insurance companies used in the 1920's?

(m) [easy] Can the principle of inverse probability work without priors? Yes/no

(n) [difficult] In class we discussed the "principle of indifference" which is a term I borrowed from Donald Gillies' Philosophical Theories of Probability. On Wikipedia, it says that Jacob Bernoulli called it the "principle of insufficient reason". McGrayne in her research of original sources comes up with many names throughout history this principle was named. List all of them you can find here.

(o) [easy] Jeffreys seems to be the founding father of modern Bayesian Statistics. But why did the world turn frequentist in the 1920's? (p57)

Problem 2

More about likelihood estimators. These exercises are important because we will soon be doing normal mean estimation using the Bayesian shrinkage estimator.

(a) [easy] Write the PDF of $X \sim \mathcal{N}(\theta, 1^2)$.

(b) [difficult] Find the MLE for θ if $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\theta, 1^2)$.

(c) [difficult] [MA] Find the MLE for θ if $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. Solve the system of equations $\frac{\partial}{\partial \mu}[\ell(\theta)] = 0$ and $\frac{\partial}{\partial \sigma^2}[\ell(\theta)] = 0$ where $\ell(\theta)$ denotes the log likelihood. You can easily find this online. But try to do it yourself.

Problem 3

More about likelihood estimators. These exercises are important because we will soon be doing normal mean estimation using the Bayesian shrinkage estimator.

(a) [easy] Write the PDF of $X \sim \mathcal{N}(\theta, 1^2)$.

(b) [difficult] Find the MLE for θ if $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\theta, 1^2)$.

(c) [difficult] [MA] Find the MLE for θ if $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. Solve the system of equations $\dfrac{\partial}{\partial \mu}[\ell(\theta)] = 0$ and $\dfrac{\partial}{\partial \sigma^2}[\ell(\theta)] = 0$ where $\ell(\theta)$ denotes the log likelihood. You can easily find this online. But try to do it yourself.

Problem 4

This problem is concerned with one of the definitions of probability. In Math 241 I taught that there were two perspectives on probability which loosely yielded four definitions according to Gillies. As a review, we have:

1. **The Objective View** — This is the view that probabilities are properties of the physical world and can only be defined by either

   (a) its **Long Run Frequency** Seeing the same event over and over again and tabulating when the event occurs will create a frequency which will asympotically become the probability or

   (b) its **Propensity** which means deep down inside, the physical object is wired for events in certain proportions

   Thus events that are non physical such as the probability of Donald Trump winning the 2016 election is outside of the purview of probability. The objective view of probability is tied to the frequentist view of statistics. We also have the...

2. **The Epistemic View** — This is the view that probabilities are inherently living inside the minds of human beings who are forced to grapple with uncertainty as they see it. This was Laplace's view — probability is an illusion because we don't have certainty about the universe. The two definitions here are that probability

   (a) is **Logical** which means that given the same information, everyone would come to the same conclusion.

   (b) is **Subjective** which means that given the same information, everyone would *not* come to the same conclusion. Thus probability is defined as the degree of belief of some individual which differs from another individual.

   Thus events that are non physical such as the probability of Donald Trump winning the 2016 election can now be legitimate probabilities as they can be computed.

   We will focus here on the logical definition.

   (a) [easy] In the logical definition we see written "given the same information". What does "information" mean here? X or θ?

   (b) [easy] In the logical definition we see written "would come to the same conclusion". What does "conclusion" mean here? Look at the four parts of the Bayes Rule equation for posterior inference and write the correct one.

   (c) [easy] Given parts (a) and (b), what does everyone need to start with being the same for conclusions to be the same?

(d) [E.C.] [MA] For some reason in order for the logical definition to work, Keynes required the principle of indifference which is a stronger assumption than everyone begins with equal priors. Why do you think that is?

(e) [easy] The logical definition requires employing the principle of indifference. State the principle of indifference for a discrete set $\Theta_0 = \{\theta_1, \ldots, \theta_m\}$ by giving the prior distribution of θ.

(f) [easy] State the principle of indifference for a continuous set $\Theta_0 = [a,\ b]$ by giving the prior distribution of θ.

(g) [difficult] Show that if $\Theta_0 = \{\theta_1, \ldots\}$ i.e. is countably infinite, the principle of indifference fails. Proof by contradiction is an easy strategy.

(h) [difficult] [MA] Show that if $\Theta_0 = \mathbb{R}$ or some subset of $\mathbb{R}$ of infinite measure, the principle of indifference fails. Proof by contradiction is an easy strategy.

(i) [difficult] State a practical case where you would not want to use the principle of indifference. Think simple. Think coins and coin flips.

(j) [difficult] [MA] How is the objectivist view of probability related to the frequentist view of statistics?

(k) [difficult] [MA] How is the epistemic view of probability related to the Bayesian view of statistics?

Problem 5

More coin flips. The purpose of this exercise is for you to review the methods we did in class concerning the binomial likelihood and Bayesian inference for the parameter $\theta = p$.

(a) [easy] Assume that $\Theta_0 = \{0.1,\ 0.5\}$. Using the principle of indifference, what is the prior?

(b) [harder] Draw the space $\Theta_0 \times \mathcal{X}$ to scale just like we did in class.

(c) [harder] Draw the space for Θ_0 given the data was 0,0,0 to scale just like we did in class.

(d) [easy] Calculate $\mathbb{P}(\theta = 0.1 \mid X_1 = 0,\ X_2 = 0,\ X_3 = 0)$. The picture above should help.

(e) [easy] Assume for the rest of the problem that $\Theta_0 = \{\theta_1, \theta_2, \ldots, \theta_5\}$. Assume the same data as we did in class $x_1 = 0,\ x_2 = 1,\ x_3 = 1$. Write out the likelihood. Your answer should not include any θ_i's but a general free variable θ.

(f) [harder] Right out an expression for the denominator using the sum notation. The denominator here would be $\mathbb{P}(X_1 = 0,\ X_2 = 1,\ X_3 = 1)$.

(g) [harder] Solve for the posterior probability of θ using your answers from the previous questions.

(h) [harder] Solve for the posterior probability of θ using your answers from the previous questions.

(i) [difficult] Write expressions for $\hat{\theta}_{\text{MAP}}$, $\hat{\theta}_{\text{MAE}}$ and $\hat{\theta}_{\text{MMSE}}$, i.e. the three Bayesian point estimators we discussed in class.

(j) [difficult] Find the distribution $\mathbb{P}(X^* \mid X_1 = 0,\ X_2 = 1,\ X_3 = 1)$ where x^* is a new realization from the model $\mathcal{F}$ heretofore never seen. Sum signs are okay.

Problem 6

More about fundamentals of the Bayesian perspective.

(a) [harder] Imagine in class we had a prior on θ in the Bernoulli family of indifferent between $\theta = 0.25$ and $\theta = 0.75$. What is the *prior predictive distribution* i.e. what is the distribution of X_1 not seeing any data just given the prior. Draw the tree out as I did in class and it will be obvious. This is *not* the posterior predictive distribution.

(b) [difficult] Give the formula in general for the prior predictive distribution for both the discrete and continuous case for general $\mathcal{F}$. Call it $\mathbb{P}(X)$.

(c) [easy] Is this the same as the marginal likelihood / prior on the data as we've seen in class? Yes/no. Write a sentence giving more flesh to the definition of that pesky denominator $\mathbb{P}(X)$.

(d) [harder] More about the marginal likelihood... Explain why if the "classic" idea of independence you learned in 241 i.e. $\mathbb{P}(X) := \mathbb{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbb{P}(X_i)$ was applicable in the Bayesian perspective then no learning can be done from experience. Use the expression for the posterior predictive distribution to make a simple one-line statement.

(e) [harder] [MA] We have showed that $X_1, \ldots, X_n$ are not independent. Show here that although they are not independent, they are identically distributed. You will need to use the property that $X_1 \mid \theta, \ldots, X_n \mid \theta \overset{iid}{\sim} p(x)$ which is known as "conditional independence". I know this is splitting hairs, but it's a good property to prove.

(f) [easy] [MA] We showed in class that $X_1, \ldots, X_n$ are not independent. Independence is in some sense "too strong" of an assumption for Bayesian modeling. Instead, we use the weaker assumption called "exchangeability". Look in the book or on Google and give the definition for exchangeability.

(g) [easy] [MA] When we collect data for an experiment are the observations exchangeable? Is this an assumption we make use of all the time?

(h) [easy] [MA] Look up de Finetti's respresentation theorem. State it here.

(i) [easy] [MA] Does exchangeability imply conditional independence for some distribution $\mathbb{P}(X \mid \theta)$? Yes / No.