Midterm – Minute Questions and Responses **Stat 495**

Questions are organized by topic, starting from chapter 1 topics through to chapter 8 (with chapter 16 before chapter 8). Questions on non-statistical topics are at the end.

Ch. 1-5 - *Example of how a theory question would be presented.*

As I stated in class (and below in the general section), I've designed the midterm so it doesn't test your understanding of math/calculus. So, you should (for the exam) be able to explain things like (take chapter 3 for context on Bayesian statistics) how the posterior is related to the prior, and the idea of updating the prior with data, but won't have to derive results. You may be asked to apply a result that is provided to you. For example, in class, we derived the Gamma-Poisson posterior, and so, I could state that result (or another result), give you a prior and some data, and you'd need to get the posterior from that. **E.G.** Problem states result from class, or perhaps a result you would have seen somewhere else (like homework). Then asks if X is sampled from a Poisson distribution with parameter lambda and the prior on lambda is a Gamma(3,4), and you have observed 4 values of X to be: 8, 11, 5, and 9, what is the posterior for lambda?

Ch. 5 – What should we know about the exponential family of distributions?

I'm unsure if this was the exact question intended, but figure it's asking about the material in chapter 5 and what should be focused on.

I would make sure you understand:

- why statisticians like working with the exponential family of distributions
- how to identify if a distribution is in the family (note, the result would be given to you to apply)
- some examples of distributions in the family, and
- general properties of the multinomial and multivariate normal (bivariate), but don't memorize formulas.

7.3 and Ch. 16 – Can you give a review of cross-validation?

The discussion in the CASI notes for 7.3 and Ch. 16 is pretty good for this. Look at the filled in notes that are posted. Just note that if you are doing a search to find lambda, the software is doing a CV at each potential lambda to estimate the MSE, and then after that, is choosing the lambda that had the lowest MSE resulting from the CV. So, what is described in the filled in handout is what would happen at each lambda, and then you'd compare across all lambdas to find the "best" one.

7.3 and Ch. 16 - Do you have suggestions about understanding ridge and lasso output in RStudio? What are all the pieces to look for?

For this, I would suggest looking over your homework 3 or the homework 3 solution. Telling these models apart is important, and that's the alpha option set in them.

I would make sure you can:

- recognize when CV is being used and when it is not to pick lambda
- understand when predictions are being obtained
- understand when MSE is being calculated
- understand when coefficients are being printed to the screen, and
- recognize when code is counting coefficients that are 0 or non-zero for you. (Some of you counted this manually for the homework – you can have R do it for you.)

Your homework 3 had examples of these and so the solution does as well in case you had issues with it.

Ch. 16 – What do we need to know from the Tibshirani 1996 paper?

This is a great question! The work with the paper was designed to give you insights into LASSO. So, some of this will just be a general list of what to know about LASSO.

I would make sure you:

- understand what the LASSO penalty is and how it differs from ridge
- can read a path diagram
- can explain why LASSO can set coefficients to 0 but ridge really can't (again, tiny chance it could)
- know that shrinkage concepts aren't new, and that LASSO was building from the non-negative garrote (but not details of its formulas, etc.)
- have a sense of when LASSO outperforms some of the other methods presented (e.g. If you have a model to fit and only 3 predictors, are you going to run LASSO? When would you?)

7.3 and Ch. 16 and Ch. 8 – **Can we review what we should know about assessing model performance?**

We've seen assessment of model performance in two broad types of models: logistic regression and other models with a continuous response.

For logistic, at the beginning of the semester, we saw confusion matrices, concordance, and even some pseudo-R^2 statistics. Of those, I would want you to be comfortable with the confusion matrices. More exists here that we didn't cover, too, such as ROC curves.

For all the other models, we've been focused on using the MSE to compare models. There are other ways to compare models, such as AIC and BIC, but as long as you can think about the MSE for the exam, that's good. Obviously, concepts you've seen before such as residual SE and R^2 are relevant as well.

Ch. 8 – Can we please see another example of Poisson regression?

One challenge with the Poisson is that if you have a large number of 0 response values, the Poisson isn't appropriate anymore, and so, you may see something called the zero-inflated Poisson, which attempts to deal with the 0s. So, I had to look a little bit to find other examples, while trying not to run into a zero-inflated model.

For examples that seem appropriate, I went to a textbook written by some colleagues of mine at St. Olaf. They have a textbook they use that goes "Beyond MLR", with a chapter on Poisson regression, including the zero-inflated Poisson model. They give you three case studies you could look at (the third is a zero-inflated model). It's a free online text, and here is the link to the Poisson regression chapter:

https://bookdown.org/roback/bookdown-BeyondMLR/ch-poissonreg.html

Ch. 8 – What should we know about regression trees?

Again, this is a bit broad, but I would make sure you:

- understand when you would use a regression tree versus a classification tree
- can explain a bit about the hypercubes and recursive partitioning
- understand that there are a variety of different criteria that govern tree growth, with some examples
- understand the software has different criteria available to determine the best splits
- understand how predictions are obtained at the end of the tree
- understand a bit about concerns about tree instability (will motivate future methods)
- know how to read a tree and obtain a prediction if given an observation

Ch. 8 - Can you explain more about the components of a regression / classification tree and their significance?

I'm assuming this is based on the output? Basic output gives you a summary of the tree including number of splits and cp values. You should know predict can be used to get predicted values and then you can do the usual MSE and misclassification error, etc. as appropriate. You can get way more information about the entire process as well – it will save what splits it considered so you can see why it did what it did. This is usually more than you need for using a tree though.

General

Will exam questions include definitions/descriptions of terms/techniques or just application?

I think the best answer to this is just looking over the practice questions. There are some that ask you to describe, and others are application. The main difference is you won't be coding during the exam – I'm giving you all the output, so you should make sure you can understand what output comes for each of the main techniques we covered: ridge, lasso, glms, trees, and how to know which is which.

How will theoretical concepts be tested on the midterm?

You've already been tested on the aspects of computations, etc. in Stat 370. The idea on the midterm is you may see concepts (because they are foundational), as well as some of the specifics/extensions CASI went into. Additionally, a result may be supplied to you for you to apply, but you won't be doing major derivations, etc. again. You don't need to memorize formulas to apply here. I will provide necessary equations. The goal is not to test your calculus or re-hash derivations you've already seen. I would expect more short-answer type questions related to the theory, or problems like the example in the first question/response on this sheet.

What is the split between material in CASI that's review and new material on the midterm?

The ridge, lasso, GLM, and tree material is the bulk of it. I can't give a precise breakdown because I haven't written the exam yet. Much of the theoretical material has already been tested so, again, I'd want to just make sure some of the foundational concepts are sound.

How hard is the midterm expected to be?

I'm not sure how to answer this question, without a point of reference. It's a midterm for a 400-level class. I expect at least one of you will obtain a near 100%, while others may encounter some challenges. Exam scores are typically left-skewed but medians can range from say, maybe 80 up to say, 86 or so, thinking of past exams?

The midterm is an opportunity to show your knowledge of the new methods we've learned and your foundational understanding of concepts.

What are my best resources for studying for the exam?

Homework questions and the supplied practice questions.