STAT 422/722 Spring 2016 Homework #3

Professor Adam Kapelner

Optionally Due *4th floor JMHH* Monday, February 27 4PM

(this document last updated Wednesday 22nd February, 2017 at 3:32pm)

Instructions and Philosophy

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 green 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 purple problems.

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 but this homework does not count toward your average. Links to instaling LaTeX and the programs for compiling LaTeX is written about in the syllabus. You are encouraged to use `overleaf.com`. If you are handing in homework this way, (1) upload `hwxx.tex` and `preamble.tex` from the correct github folder, (2) read the comments in the code as there is *one line to comment out*, (3) you should replace my name with your name and (4) your section. 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, you must print this document and write in your answers. You must print after downloading and opening in Adobe reader (not from Google Chrome viewer). I do not accept homeworks not on the correctly paginated printout of this document. Write your name and section below (A or B).

You may collaborate, but hand in your own copy with your own wording. See the syllabus for more information.

NAME: _____ COURSE (422 or 722): _____

SECTION ("A" for Tuesday or "B" for Wednesday): _____

Problem 1

We will be investigating missing data.

(a) [easy] Give an example of a selection model.

(b) [easy] Give the same example, but change one thing about the missingness that would render it into a pattern mixture model.

(c) [easy] If $p = 10$ in your dataset but only $p = 5$ features have missingness and you assume a pattern mixture model for all missingness, how many features would your design matrix have after augmentation *and* imputation?

(d) [easy] Explain two reasons why listwise deletion would not be recommended.

(e) [difficult] Besides imputation, what else could you do?

(f) [difficult] Explain why trying to impute in the pattern mixture case can still bear fruit.

(g) [easy] Give examples of measurement in real-life that exhibits MCAR, MAR and NMAR.

(h) [difficult] If MCAR is the result of random holes, why can't it be imputed?

(i) [difficult] Explain the procedure to impute for the new $\boldsymbol{x}^*$ records.

(j) [difficult] Explain the procedure in the `R` package `missForest` and why is it a good procedure to use in practice?

Problem 2

We will looking at using oos methods to do model selection.

(a) [difficult] Demonstrate on a dataset of your choice, the complexity-fit tradeoff a la slide 12 of Lecture 5. Explain what you did.

(b) [easy] Give an example of a stationary model and a non-stationary model.

(c) [difficult] JMP practice: use JMP to do 5-fold CV on the white wine data. This is a lot of work! You have to create the folds manually. Report how you made the folds and your oos metrics.

(d) [easy] Give an example of a stationary model and a non-stationary model.

(e) [easy] Slide 24 of Lecture 5 — what precisely is not legal about this procedure?

(f) [difficult] JMP practice: duplicate the exercise and demonstrate model C is the "best".

(g) [harder] In three splits, explain exactly how $\hat{f}$ (the model that is shipped for production) is ultimately created.

(h) [difficult] Explain the main disadvantage of LOOCV.

(i) [harder] Given model candidates $M_1, M_2, \ldots, M_m$, you can find the best one using oos validation as M^*. What main issue was ignored here?

(j) [difficult] Provide as many ways as you can to expand the predictor set and derive new features beyond the strategies discussed in class.

Problem 3

We will reviewing stepwise regression.

(a) [harder] Why does stepwise logistic regression take so long?

(b) [harder] Why would running stepwise regression on the white wine data as-is be of little value?

(c) [difficult] JMP practice: Use the baseball dataset to fit stepwise on a highly expanded menu of derived predictors of your choice. Did you *truly* beat the fit of a simple linear model?

(d) [E.C.] Derive the general AIC formula from scratch (not AICc).

(e) [difficult] Given your previous answer, derive AIC for the linear model.

(f) [difficult] How much more likelihood would you need to justify adding one more predictor if your beginning likelihood was 1 in 100?

(g) [easy] What does the AICc metric attempt to correct in the AIC metrc?

(h) [easy] What does the AICc metric attempt to correct in the AIC metrc?

(i) [harder] Why is stepwise based on a p-value threshold for an individual predictor not a wise choice in general?

Problem 4

We will build some decision trees.

(a) [easy] When is binning a good idea to do non-parametric regression (if there is enough data)?

(b) [easy] When (and why) does binning break down?

(c) [difficult] What would be the problem with allowing for all bins and all interactions (if of course the computer can crunch the numbers) in a forward stepwise procedure?

(d) [difficult] Given the data below fit a tree that minimizes SSE. No stopping rule. Do it by hand for the practice.

x	y
1	6.12
2	6.02
3	6.25
4	5.95
5	6.09
6	-21.34
7	-20.85
8	-21.03
9	-3.87
10	9.67
11	9.63
12	9.01

(e) [easy] Why do you suppose the model in the previous question is overfit?

(f) [easy] What is its depth? How many leaves? How many inner nodes?

(g) [easy] Fit the same data using the stopping rule of minimum nodesize of 5.

(h) [easy] How would you determine overfittedness to these two trees?

(i) [easy] Fit the best tree you can find for the white wine data with 10 total nodes using a stepwise procedure. Which variables were split on?

Problem 5

We will be investigating Random Forests (RF).

(a) [difficult] Explain why bagging reduces error.

(b) [difficult] Explain why sampling prediction reduces error.

(c) [easy] Fit an RF to the white wine data. Fit a linear model to the white wine data. Why did the RF do better?

(d) [difficult] Report RF's oos $RMSE$ - in-sample $RMSE$. Why is this a large number?

(e) [easy] Why does RF's in-sample R^2 lie?

(f) [E.C.] How would you show the effect of alcohol on quality from the RF model?

(g) [easy] Fit an RF to the baseball data. Fit a linear model to the baseball data. Why didn't the RF do better?

(h) [difficult] Fit an RF to the churn dataset (the fixed dataset in the Lec 6 folder). Build an ROC curve and eyeball the AUC.

(i) [difficult] Why are more trees in the RF better?

(j) [harder] Why would the RF do worse if you use all variables in each tree?

(k) [harder] Why would the RF do worse if you use only one variable in each tree?