STAT 422/722 Spring 2016 Homework #3 Limited Solutions

Professor Adam Kapelner

Sunday 26th February, 2017

Problem 1

We will be investigating missing data.

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

Let's go back to the banking example of predicting mortgage default. People fill out mortgage applications. Each person gets a secretary randomly. One secretary does not print out page 7 for the mortgage application. All the information elicited on page 7 goes missing. However, this type of missingness should not affect the response in its own right.

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

Instead of the secretary not printing out page 7, some individuals deliberately do not fill out page 7 because they are trying to pull a fast one.

(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?

$10 + 5 = 15$

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

(a) You would likely be removing observations non-randomly. The historical dataframe would no long represent the population you likely would want to predict for in the future. Hence, your model will likely have higher generalization error.

(b) By removing samples, your model estimation error increases again leading to higher generalization error.

(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.

MCAR Random digital corruption in a database caused by hard drive failure where each observation only has one feature.

MAR Random digital corruption in an image files' pixels. The pixel values can be imputed from the surrounding pixels' color values.

NMAR Information revealing square footage of an apartment going missing on the MLSI listing.

(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.

Stationary A model of predicting voltage in an electric circuit based on inputs.
Non-Stationary Predicting apartment prices in Queens.

(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.

Duplicate question

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

Snooped the test set multiple times.

(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.

Regardless of the observations in the three splits, the entire dataset is used to produce $\hat{f}$ on the model that was found to be best in the validation set.

(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?

How did you obtain the candidates and how do you know they're any good?

(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?

Logistic regression itself is purely numerical which means there are iterations to converge on a local minimum of the log-likelihood function. Now we have to do that procedure for all possible models iteratively.

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

There are only 11 features. Here, we can actually find the best subset model since all $2^{11} = 2048$ possible models can be checked explicitly and the best one selected. Thus, stepwise would be suboptimal.

(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?

It attempts to make it more stable for low sample sizes.

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

(1) because out-of-sample validation is better at assessing overfitting and (2) AICc is proven to reflect overfitting under some assumptions while the p value threshold I do not believe has been proven to reflect overfitting.

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)?

Binning allows for fitting non-linear and interacting functions flexibly.

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

When there are many features.

(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?

No stopping rule will split all the way down to one observation per leaf. Since leaf assignment is $\hat{\boldsymbol{y}} = \bar{y} = y$, there will be 100% R^2. Since $\mathcal{E}$ is always assumed in any real-life problem, there is definite overfitting.

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

5 / 12 / 11

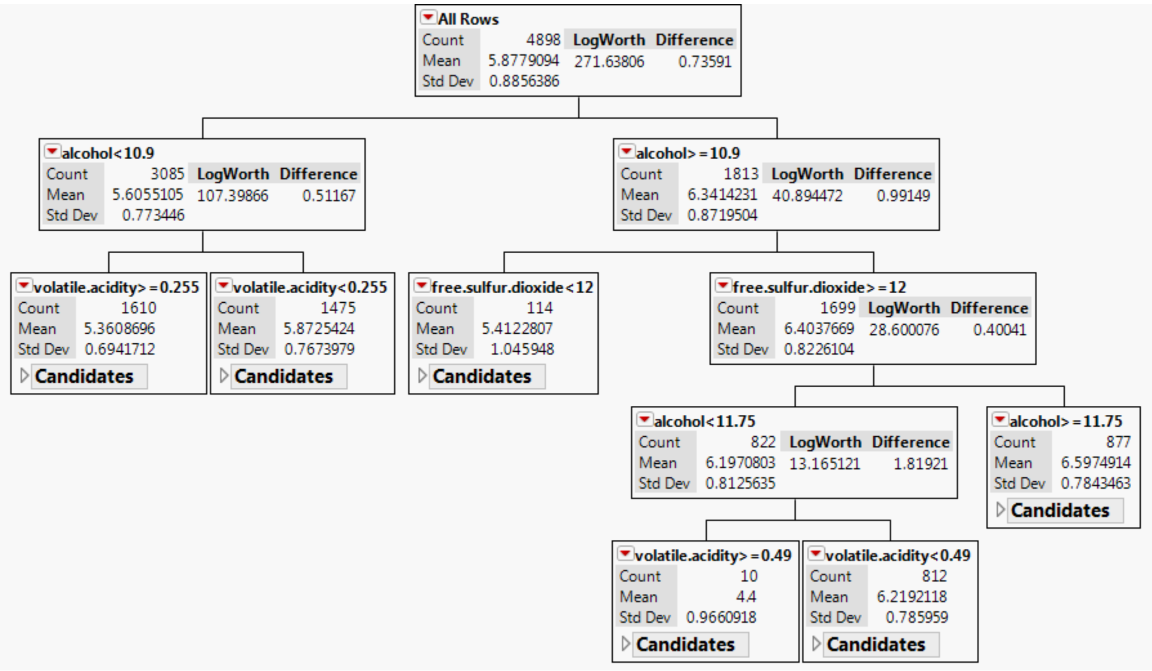
(g) [easy] Fit the same data using the stopping rule of minimum nodesize of 5.

There is one split at $x \leq 6$.

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

Keep a hold out set (test set) or use K-fold CV.

(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?



All Rows

		LogWorth	Difference
Count	4898		
Mean	5.8779094	271.63806	0.73591
Std Dev	0.8856386		

alcohol<10.9

		LogWorth	Difference
Count	3085		
Mean	5.6055105	107.39866	0.51167
Std Dev	0.773446		

alcohol>=10.9

		LogWorth	Difference
Count	1813		
Mean	6.3414231	40.894472	0.99149
Std Dev	0.8719504		

volatile.acidity>=0.255

Count	1610
Mean	5.3608696
Std Dev	0.6941712

Candidates

volatile.acidity<0.255

Count	1475
Mean	5.8725424
Std Dev	0.7673979

Candidates

free.sulfur.dioxide<12

Count	114
Mean	5.4122807
Std Dev	1.045948

Candidates

free.sulfur.dioxide>=12

		LogWorth	Difference
Count	1699		
Mean	6.4037669	28.600076	0.40041
Std Dev	0.8226104		

alcohol<11.75

		LogWorth	Difference
Count	822		
Mean	6.1970803	13.165121	1.81921
Std Dev	0.8125635		

alcohol>=11.75

Count	877
Mean	6.5974914
Std Dev	0.7843463

Candidates

volatile.acidity>=0.49

Count	10
Mean	4.4
Std Dev	0.9660918

Candidates

volatile.acidity<0.49

Count	812
Mean	6.2192118
Std Dev	0.785959

Candidates

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?

RF can "find" non-linearities and interactions and the linear model cannot.

(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?

The trees are overfit by design. Hence, each observation when dropped down the trees will be in-bag 2/3 of the time (and hence overfit).

(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?

It seems that (remarkably) the model truly is linear.

(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?

The trees would be too correlated and the total variance of the model will increase (i.e. the generalization error).

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

The trees would be too weak at fitting the function. Here, you are forcing an additive model by precluding RF's ability to find interactions.