Math 342W / 642 Spring 2024
Midterm Examination Two Solutions

Professor Adam Kapelner

May 14, 2024

Full Name _____

Code of Academic Integrity

Since the college is an academic community, its fundamental purpose is the pursuit of knowledge. Essential to the success of this educational mission is a commitment to the principles of academic integrity. Every member of the college community is responsible for upholding the highest standards of honesty at all times. Students, as members of the community, are also responsible for adhering to the principles and spirit of the following Code of Academic Integrity.

Activities that have the effect or intention of interfering with education, pursuit of knowledge, or fair evaluation of a student's performance are prohibited. Examples of such activities include but are not limited to the following definitions:

Cheating Using or attempting to use unauthorized assistance, material, or study aids in examinations or other academic work or preventing, or attempting to prevent, another from using authorized assistance, material, or study aids. Example: using an unauthorized cheat sheet in a quiz or exam, altering a graded exam and resubmitting it for a better grade, etc.

I acknowledge and agree to uphold this Code of Academic Integrity.

_____ _____

signature date

Instructions

This exam is 110 minutes and closed-book. You are allowed **two** pages (front and back) of a "cheat sheet." You may use a graphing calculator of your choice. Please read the questions carefully. If the question reads "compute," this means the solution will be a number otherwise you can leave the answer in *any* widely accepted mathematical notation which could be resolved to an exact or approximate number with the use of a computer. I advise you to skip problems marked "[Extra Credit]" until you have finished the other questions on the exam, then loop back and plug in all the holes. I also advise you to use pencil. The exam score will be normed to be out of 100 points total plus extra credit if it exists. Partial credit will be granted for incomplete answers on most of the questions. $\boxed{\text{Box}}$ in your final answers. Good luck!

Problem 1 In class, we spoke about probability estimation for a binary phenomenon $\mathcal{Y} = \{0, 1\}$. We modeled each observation as an independent Bernoulli (θ_i) i.e. $Y_i \overset{ind}{\sim} \theta_i^{y_i}(1-\theta_i)^{1-y_i}$ where $\theta_i := \mathbb{E}[Y_i = 1 \mid \boldsymbol{x}_i]$ which for the Bernoulli is synonymous with $\mathbb{P}(Y_i = 1 \mid \boldsymbol{x}_i)$ and it varies with observation based on the features $\boldsymbol{x}_i$ which is a row vector of length $p + 1$ since the first entry is set to be one.

To do so, we used a generalized linear model (GLM) which coerced the linear model $\boldsymbol{x} \cdot \boldsymbol{w}$ into the support of the parameter θ_i, a probability ranging from $[0, 1]$. To do this coercion, we used a link function $\phi(\boldsymbol{x} \cdot \boldsymbol{w})$ which mapped $\boldsymbol{x} \cdot \boldsymbol{w} \in \mathbb{R} \to \mathbb{S}\text{upp}[\theta_i] = [0, 1]$. Any monotonically increasing function with domain $\mathbb{R}$ and range $[0, 1]$ was legal. For example, any CDF of a random variable with support $\mathbb{R}$ fits this definition.

Let's use the link function $\phi(u)$ is the CDF of the standard Gumbel, $F(x) = e^{-e^{-x}}$ where this algorithm is called "cloglog regression" and we'll denote it $\mathcal{A}_{\text{cloglog}}$.

- [4 pt / 4 pts] Write out the objective function to maximize which is the probability of the entire training set $\mathbb{D}$. Since this is a GLM, your answer must include the linear term for the ith observation, $\boldsymbol{x}_i \cdot \boldsymbol{w}$.

$$\mathbb{P}(Y_1, \ldots, Y_n \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \prod_{i=1}^{n} \mathbb{P}(Y_i \mid \boldsymbol{x}_i) = \prod_{i=1}^{n} \left(e^{-e^{-\boldsymbol{x}_i \cdot \boldsymbol{w}}}\right)^{y_i} \left(1 - e^{-e^{-\boldsymbol{x}_i \cdot \boldsymbol{w}}}\right)^{1-y_i}$$

- [1 pt / 5 pts] Our algorithm $\mathcal{A}_{\text{cloglog}}$ involves running this optimization problem in the computer: $\boldsymbol{b} := \arg\max_{\boldsymbol{w}} \{\text{your answer from the previous problem}\}$. What is the dimension of the vector $\boldsymbol{b}$? $p + 1$

- [3 pt / 8 pts] Given $\boldsymbol{b}$ and a new observation $\boldsymbol{x}_\star$, write the explicit functional form of $g(\boldsymbol{x}_\star)$, an expression that computes $\hat{p}_\star$, the probability estimate that the measured phenomenon will be one.

$$\hat{p}_\star = e^{-e^{-\boldsymbol{x}_\star \cdot \boldsymbol{b}}}$$

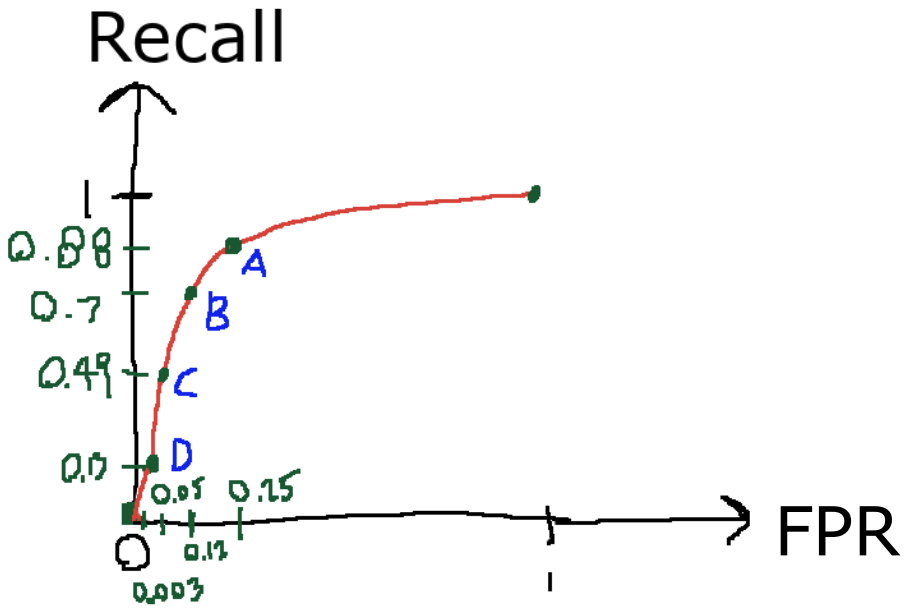
We now examine performance of this model out of sample using different $\hat{p}$ thresholds. Here are four of the possible classifiers' oos confusion tables labeled A, B, C, D:

A	B	C	D
y_test_hat y_test 0 1 0 2866 938 1 144 1052	y_test_hat y_test 0 1 0 3357 447 1 360 836	y_test_hat y_test 0 1 0 3601 203 1 611 585	y_test_hat y_test 0 1 0 3794 10 1 1043 153
ME = (144+938) / 5000 = 0.216 Recall = 1052 / 1196 = 0.88 FPR = 938 / (938 + 2866) = 0.25 FDR/FOR = 9.85	ME = (447+360) / 5000 = 0.161 Recall = 836 / 1196 = 0.70 FPR = 447 / (447 + 3357) = 0.12 FDR/FOR = 3.60	ME = (203+611) / 5000 = 0.163 Recall = 585 / 1196 = 0.49 FPR = 203 / (203 + 3601) = 0.05 FDR/FOR = 1.78	ME = (10+1043) / 5000 = 0.211 Recall = 153 / 1196 = 0.13 FPR = 10 / (10 + 3794) = 0.003 FDR/FOR = 3.51

- [2 pt / 10 pts] With the information provided, is it possible to compute the $\hat{p}$ threshold values for the four classifiers above? Circle one: Yes / No

- [1 pt / 11 pts] Which model has the lowest oos misclassification error? Circle one: A / B / C / D

- [8 pt / 19 pts] Draw the ROC curve below to the best of your ability. Label the axes. Your curve should have 6 include distinct points whose values are marked on the axes. Four of these points should additionally be labeled as A, B, C, D corresponding to the four models displayed on the previous page.



- [2 pt / 21 pts] Is the oos AUC is definitively greater than 0.5 for this model? Circle one: Yes / No

- [2 pt / 23 pts] If false positives were much more costly than false negatives, which one of the four models is best to employ? Circle one: A / B / C / D

- [2 pt / 25 pts] If false negatives were much more costly than false positives, which one of the four models is best to employ? Circle one: A / B / C / D

- [3 pt / 28 pts] If you required a symmetric-cost classifier, which one out of the four models is best to employ? Circle one: A / B / C / D
The term "symmetric-cost classifier" is ambiguous. It could mean (a) minimize ME, (b) equalize FPR and FNR and (c) equalize FDR and FOR giving the answers B, A and C respectively.

Problem 2 In the lab we analyzed three tables: bills and bill payments which have 226,434 rows and 194,850 rows. Here are a subset of interest of the bills table followed by a subset of interest of the bill payments table:

```
         id   due_date invoice_date tot_amount customer_id discount_id
      <num>     <IDat>       <IDat>      <num>       <int>       <num>
 1: 5000000 2016-07-31   2016-06-16   99480.18    12867871     7397895
 2: 5693147 2017-05-11   2017-04-11   99528.76    12871311     7397895
 3: 6098612 2016-01-15   2016-01-04   99477.35    13135347     7397895
 4: 6386294 2016-12-30   2016-12-30   99479.31    12867871     7397895
 5: 6609438 2017-05-07   2017-04-07   99477.20    12867871     7397895
 6: 6791759 2016-12-16   2016-11-16   99477.17    13479284     7397895
 7: 6945910 2017-06-08   2017-06-08   99477.32    12855932     7397895
 8: 7079442 2017-05-12   2017-04-12   99477.19    13135347     7397895
 9: 7197225 2016-05-24   2016-04-24   99485.43    12867871     7397895
10: 7302585 2017-03-28   2017-02-26   99477.47    12855932     7397895

          id paid_amount transaction_date bill_id
       <num>       <num>           <IDat>   <num>
 1: 13780480    99150.43       2016-07-01 5000000
 2: 13654517    99220.42       2016-08-08 5693147
 3:  5000000    99148.07       2016-08-03 6098612
 4: 13845921    99154.67       2016-07-15 6386294
 5:  5693147    99148.07       2016-08-03 6609438
 6:  6945910    99152.35       2016-08-19 7302585
 7:  7197225    99151.47       2016-08-26 7397895
 8:  7079442    99148.43       2016-08-19 7484907
 9:  7302585    99158.61       2016-09-02 7564949
10:  7397895    99148.07       2016-11-04 7890372
```

- [2 pt / 30 pts] If we were to do a left join where the left table was the subset of interest of bills and the right table was the subset of interest of bill payments, what would be the number of rows in the final joined table? 10

- [2 pt / 32 pts] Given the join in the previous question, what would be the number of columns in the final joined table? 9

- [2 pt / 34 pts] Given the join in the previous question, what would be the number of rows in the final joined table after listwise deletion of missingness? 6

- [3 pt / 37 pts] Consider the table of the subset of interest of bill payments which is in "wide" format. If this table was converted from wide to long format where the metric variables were all variables (except `id`), how many rows would the final long table have? 30

- [2 pt / 39 pts] Consider the operation in the previous question. Which table has more entry values? Circle one:

  the original wide table / the transformed long table

Problem 3 Consider the following two meta-algorithms: bagging and boosting both with a large number of constituent base learners denoted M. Assume the algorithms to generate the "base learners" in both bagging and boosting are the same algorithm denoted $\mathcal{A}$. Let $\mathbb{D}_1 := \langle \boldsymbol{X}_1, \boldsymbol{y}_1 \rangle, \mathbb{D}_2 := \langle \boldsymbol{X}_2, \boldsymbol{y}_2 \rangle, \ldots, \mathbb{D}_M := \langle \boldsymbol{X}_M, \boldsymbol{y}_M \rangle$ denote the dataset used by each of the base learners. Consider a modeling scenario where $\mathcal{Y} = \mathbb{R}$ and f is known to be highly non-linear with interactions among the p covariates.

- [20 pt / 59 pts] Circle the letters of all the following that are **true**.

[a] In bagging, the higher the M, the better the oos performance (in general)

In boosting, the higher the M, the better the oos performance (in general)

[c] In bagging, the order of fitting the M base learners does not matter

In boosting, the order of fitting the M base learners does not matter

In bagging at iteration t, $\mathbb{D}_t$ is updated with the results of $g_1, g_2, \ldots, g_{t-1}$

[f] In boosting at iteration t, $\mathbb{D}_t$ is updated with the results of $g_1, g_2, \ldots, g_{t-1}$

In bagging, the M base learners are fit on the same dataset i.e.
$\mathbb{D}_1 = \mathbb{D}_2 = \ldots = \mathbb{D}_M$

In boosting, the M base learners are fit on the the same dataset i.e.
$\mathbb{D}_1 = \mathbb{D}_2 = \ldots = \mathbb{D}_M$

In bagging, the M base learners are fit on the the same input matrices i.e.
$\boldsymbol{X}_1 = \boldsymbol{X}_2 = \ldots = \boldsymbol{X}_M$

[j] In boosting, the M base learners are fit on the the same input matrices i.e.
$\boldsymbol{X}_1 = \boldsymbol{X}_2 = \ldots = \boldsymbol{X}_M$

In bagging, the M base learners are fit on the the same response vectors i.e.
$\boldsymbol{y}_1 = \boldsymbol{y}_2 = \ldots = \boldsymbol{y}_M$

In boosting, the M base learners are fit on the the same response vectors i.e.
$\boldsymbol{y}_1 = \boldsymbol{y}_2 = \ldots = \boldsymbol{y}_M$

In bagging, if $\mathcal{A}$ = OLS, oos performance is likely to be high

In boosting, if $\mathcal{A}$ = OLS, oos performance is likely to be high

In bagging, if $\mathcal{A}$ = CART with N_0 large, oos performance is likely to be high

[p] In boosting, if $\mathcal{A}$ = CART with N_0 large, oos performance is likely to be high

[q] In bagging, if $\mathcal{A}$ = CART with N_0 small, oos performance is likely to be high

In boosting, if $\mathcal{A}$ = CART with N_0 small, oos performance is likely to be high

Bagging is in general a better meta-algorithm than boosting as measured by oos performance

Boosting is in general a better meta-algorithm than bagging as measured by oos performance

Problem 4 Consider a subset of the `boston housing` data frame, which has $n = 500$ observations and $p_{raw} = 13$ numeric measurements. The response variable is `medv` with an average value of 22.53 (measured in 1000 USD). We wish to fit a forward stepwise OLS model to this data beginning with the intercept and with a pool consisting of all first-order interactions (i.e. the `R` formula `y ~ .*.`)

We split the dataset into a training set of size 300, a select set of size 100 and a test set of size 100. We do not randomize the order of the dataset when doing so. We then use nested resampling which rotates the select set and the test set. Assume we have sufficient resources to run the stepwise until its completion with all first-order interactions for every iteration (in both the inner and outer loops). We report some RMSE metrics below for the first training-select split of the first inner loop:

```
> Dtrain_1_1 = MASS::Boston[1:300, ]
> Dselect_1_1 = MASS::Boston[301:400, ]
> mod_1_1_0 = lm(medv ~ 1, Dtrain_1_1)
> mod_1_1_M = lm(medv ~ . * ., Dtrain_1_1)
> round(summary(mod_1_1_0)$sigma, 2)
[1] 8.89
> round(summary(mod_1_1_M)$sigma, 2)
[1] 2.16
> yhat_1_1_0 = predict(mod_1_1_0, Dselect_1_1)
> sqrt(mean((yhat_1_1_0 - Dselect_1_1$medv)^2))
[1] 10.21488
> yhat_1_1_M = predict(mod_1_1_M, Dselect_1_1)
> sqrt(mean((yhat_1_1_M - Dselect_1_1$medv)^2))
[1] 59.01323
```

- [2 pt / 61 pts] How many different training-select splits are in one inner loop?
  $K_{inner} = 4$ since we are dividing 400 observations into a set of 300 and 100.

- [2 pt / 63 pts] How many different training-test splits are in one outer loop?
  $K_{outer} = 5$ since we are dividing 500 observations into a set of 400 and 100.

- [3 pt / 66 pts] How many iterations of the greedy stepwise algorithm are done for the first training-select split of the first inner loop of the first outer loop?

  We begin with the intercept (so that doesn't count as an "iteration". Then, there's p linear terms, p squared terms and $\binom{p}{2}$ interactions of two different variables, i.e., $p + p + \binom{p}{2} = 26 + \binom{13}{2} = 104$. Note: addition of the intercept is also accepted as it is ambiguous if it an "iteration" or not. So 105 is the other acceptable answer.

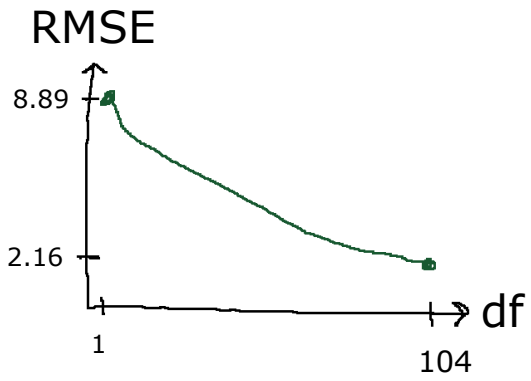
- [3 pt / 69 pts] How many iterations of the greedy stepwise algorithm are done for the entire nested resampling procedure across all folds?

  $\left(p + p + \binom{p}{2}\right) K_{inner} K_{outer} = 104(4)(5) = 2080$

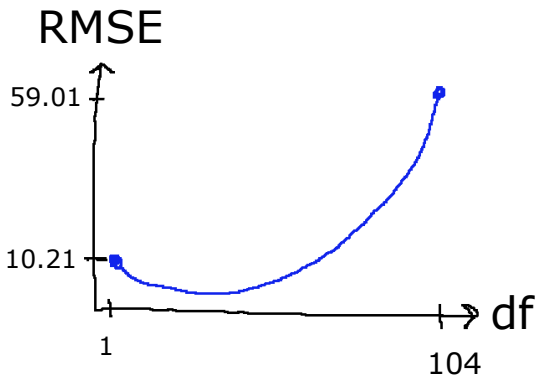
- [3 pt / 72 pts] When building the final model (after validation is completed), how many iterations of the greedy stepwise algorithm are done across all folds?

  $\left(p + p + \binom{p}{2}\right) K_{outer} = 104(5) = 520$

- [5 pt / 77 pts] For the first training-select split of the first inner loop, draw the in-sample RMSE plot as a function degrees of freedom (draw it as best as possible given the information you have). Label your axes and mark critical points numerically.



- [5 pt / 82 pts] For the first training-select split of the first inner loop, draw the oosRMSE plot as a function degrees of freedom (draw it as best as possible given the information you have). Label your axes and mark critical points numerically.

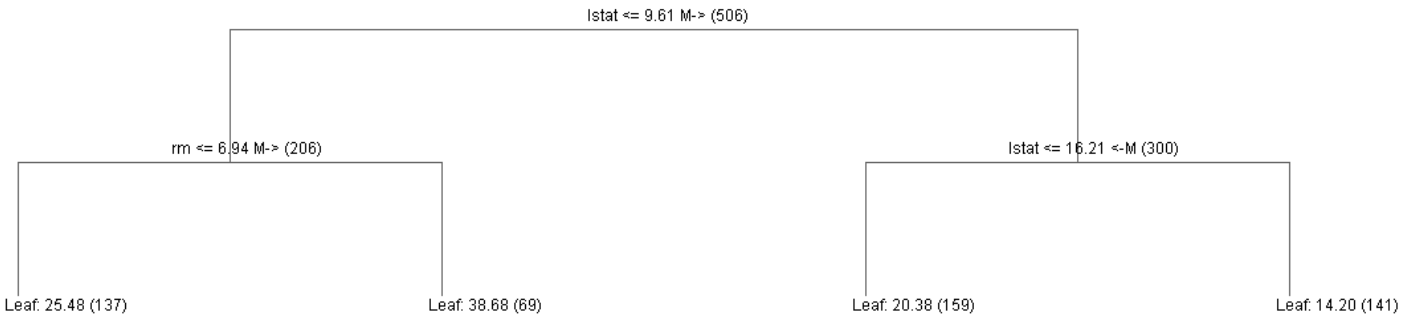


We now fit a regression tree to the full boston housing dataset. The full tree model is visualized on the top of the next page.

- [1 pt / 83 pts] How many nodes does this tree model have? 7

- [2 pt / 85 pts] What are the two most important variables (defined as variables that can decrease in-sample SSE)? lstat, rm

- [2 pt / 87 pts] What is the most precise statement you can make about the value of N_0? $N_0 \geq 159$ and $N_0 < 206$



- [2 pt / 89 pts] If the same algorithm that produced the tree visualized above was implemented as base learner for bagging with M sufficiently large, would the the oos error of the bagged model be lower than the oos error of the single tree above in all likelihood? Circle one: Yes

- [3 pt / 92 pts] If the same algorithm that produced the tree visualized above was implemented as base learner for random forest with M sufficiently large, what would likely need to be changed about the base learner's algorithm to improve performance?

  Set N_0 to a small value

Problem 5 Assume $g(\boldsymbol{x}) = \mathcal{A}(\mathbb{D}, \mathcal{H})$ is a model for a real-valued response. In class we studied the following three decompositions of MSE in the modeling context where δ was realized from a mean-centered r.v. Δ with variance σ^2 independent of the value of $\boldsymbol{x}$:

$$
\begin{aligned}
[I] \quad & MSE(\boldsymbol{x}_*) = \sigma^2 + (f(\boldsymbol{x}_*) - g(\boldsymbol{x}_*))^2 \\
[II] \quad & MSE(\boldsymbol{x}_*) = \sigma^2 + \mathbb{B}\text{ias}\left[g(\boldsymbol{x}_*)\right]^2 + \mathbb{V}\text{ar}\left[g(\boldsymbol{x}_*)\right] \\
[III] \quad & MSE = \sigma^2 + \mathbb{E}\left[\mathbb{B}\text{ias}\left[g(\boldsymbol{x}_*)\right]^2\right] + \mathbb{E}\left[\mathbb{V}\text{ar}\left[g(\boldsymbol{x}_*)\right]\right]
\end{aligned}
$$

- [2 pt / 94 pts] In [I], the MSE is taken as an expectation over which random variable(s)? Δ_*

- [3 pt / 97 pts] In [II], the MSE is taken as an expectation over which random variable(s)? Δ_* and $(\Delta_1, \ldots, \Delta_n$ or $\boldsymbol{Y})$

- [3 pt / 100 pts] In [III], the MSE is taken as an expectation over which random variable(s)? Δ_* and $(\Delta_1, \ldots, \Delta_n$ or $\boldsymbol{Y})$ and $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n, \boldsymbol{X}_*$