Math 390.4 / 650.3 Spring 2020
Midterm Examination Two

Professor Adam Kapelner

Thursday, May 14, 2020

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:

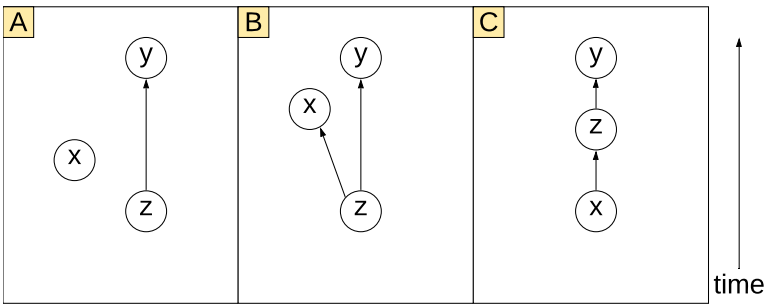
Cheating Using or attempting to use unauthorized assistance, material, or study aids in examinations or other academic work or preventing, or attempting to prevent, another from using authorized assistance, material, or study aids. Example: using an unauthorized cheat sheet in a quiz or exam, altering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

By taking this exam, you acknowledge and agree to uphold this Code of Academic Integrity.

Instructions

This exam has 146 total points, is 120 minutes (variable time per question) and closed-book. You are allowed **two** pages (front and back) of a "cheat sheet", one table of reference and scrap paper and a graphing calculator. Please read the questions carefully. No food is allowed, only drinks.

Problem 1 [10min] Consider the following causal diagrams A, B and C below. In each diagram, all three events x, y, z have other causes that are not displayed.



- [13 pt / 13 pts] Record the letter(s) of all the following that are **true**.

  (a) In diagram A, x and y could be spuriously correlated.

  (b) In diagram A, x and y are correlated.

  (c) In diagram A, z causes y.

  (d) In diagram B, x and y could be spuriously correlated.

  (e) In diagram B, x and y are correlated.

  (f) In diagram A, x causes y.

  (g) In diagram C, x and y could be spuriously correlated.

  (h) In diagram C, x and y are correlated.

  (i) In diagram C, z could cause x.

  (j) In diagram A, x causes z and x causes y.

  (k) Let y be number of car accidents daily in New York City, let z be daily rainfall in New York City and let x be the phase of the moon in degrees. The most likely causal diagram reflecting this situation is A.

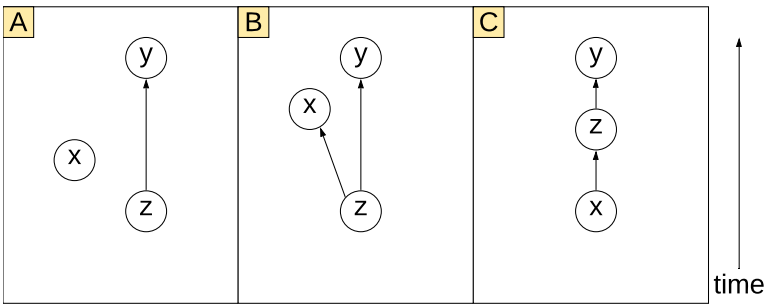
  (l) Let y be number of car accidents daily in New York City, let z be daily rainfall in New York City and let x be the number of people working from home. The most likely causal diagram reflecting this situation is B.

  (m) Let y be number of car accidents daily in New York City, let z be daily rainfall in New York City and let x be the number of people working from home. The most likely causal diagram reflecting this situation is C.

Your answer will consist of a string (e.g. `aebgd`) where the order of the letters does not matter nor does upper / lowercase.

ACDEHKL

Problem 2 [11min] Consider the same causal diagrams A, B and C below. In each diagram, all three events x, y, z have other causes that are not displayed.



Consider the following two predictive models fit by OLS:

$$[I] \quad \hat{\boldsymbol{y}}^I = b_0^I + b_1^I x$$
$$[II] \quad \hat{\boldsymbol{y}}^{II} = b_0^{II} + b_1^{II} x + b_2^{II} z$$

- [8 pt / 21 pts] Record the letter(s) of all the following that are **true**.

  (a) In causal diagram A, $b_1^I x \approx b_1^{II} x \approx 0$.

  (b) In causal diagram B, $b_1^I x \approx b_1^{II} x \approx 0$.

  (c) In causal diagram C, $b_1^I x \approx b_1^{II} x \approx 0$.

  (d) In causal diagram B, $b_1^I x \neq b_1^{II} x \approx 0$.

  (e) In causal diagram C, $b_1^I x \neq b_1^{II} x \approx 0$.

  (f) In all causal diagrams, model [II] performs better out of sample.

  The following statements are true about the most correct interpretation of b_1^I for causal diagram B for two observations, O1 and O2.

  (g) O1 and O2 must be sampled in same way as the observations in $\mathbb{D}$ and O2's x measurement is one unit higher than O1's x unit, then the y value of O2 will be on average b_1^I higher than O1 assuming the linear model is true.

  (h) O1 and O2 must be sampled in same way as the observations in $\mathbb{D}$ and O1 and O2 have the same x and then I manipulate the x value of O2 to be one higher than O1, then the y value of O2 will be on average b_1^I higher than O1 assuming the linear model is true.

Your answer will consist of a string (e.g. `aebgd`) where the order of the letters does not matter nor does upper / lowercase.

ADEFG

Problem 3 [11min] In lab 10 we worked with a data from an accounting firm. The tables were bills, payments and discounts. The bills for one customer and the discounts table is below. The `discount_id` column references the `id` column in the discounts table.

```
> billsA[order(discount_id)]
          id   due_date invoice_date    amount discount_id
1: 16294508 2017-06-25   2017-06-25  99476.43     7302585
2: 17129704 2017-07-29   2017-06-29  99475.01     7564949
3: 15247786 2015-09-18   2015-08-19  99484.62     8218876
4: 17096637 2016-02-02   2016-01-03  99668.94     8806662
5: 17222576 2016-09-15   2016-08-16  99476.93     8806662
6: 16797375 2017-01-29   2016-12-30 116487.16     8806662
> discounts[order(id)]
          id num_days pct_off days_until_discount
 1: 6098612       20      NA                   NA
 2: 6386294      120      NA                   NA
 3: 6609438       NA       1                    7
 4: 7197225       60      NA                   NA
 5: 7302585       NA      NA                   NA
 6: 7708050       NA      NA                   NA
 7: 7833213       10      NA                    5
 8: 8295837       14      NA                   10
 9: 8496508       10       2                   15
10: 8583519        0       2                   NA
11: 8610918        1      NA                   NA
12: 8784190       NA      NA                   30
13: 8806662       30      NA                   NA
14: 8951244       NA      NA                   NA
15: 8988984       60      NA                   NA
```

- [12 pt / 33 pts] Each of the following are short for "If we were to do a _____ of billsA and discounts, we would create a table with _____". Record the letter(s) of all the following that are **true**.

  (a) left join / 8 columns.
  (b) left join / 9 columns.
  (c) left join / 6 rows.
  (d) left join / 2 rows.
  (e) right join / 17 rows.
  (f) right join / 15 rows.
  (g) inner join / 2 rows.
  (h) inner join / 4 rows.
  (i) inner join / 21 rows.
  (j) full join / 2 rows.
  (k) full join / 21 rows.
  (l) full join / >21 rows.

Your answer will consist of a string (e.g. `aebgd`) where the order of the letters does not matter nor does upper / lowercase.

ACEHK

Problem 4 [11min] Here is a dataset about weather:

```
> skim(weather)
-- Data Summary ------------------------
                           Values
Name                       weather
Number of rows             26115
Number of columns          13
_________________________
Column type frequency:
  numeric                  13
_________________________
Group variables            None

-- Variable type: numeric --------------------------------------------------------------------------------
# A tibble: 13 x 11
   skim_variable n_missing complete_rate       mean       sd      p0     p25     p50     p75    p100 hist
 * <chr>             <int>         <dbl>      <dbl>    <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl> <chr>
 1 year                  0         1     2013          0      2013    2013    2013    2013    2013
 2 month                 0         1        6.50       3.44      1       4       7       9      12
 3 day                   0         1       15.7        8.76      1       8      16      23      31
 4 hour                  0         1       11.5        6.91      0       6      11      17      23
 5 temp                  1         1.00    55.3       17.8      10.9    39.9    55.4    70.0   100.
 6 dewp                  1         1.00    41.4       19.4      -9.94   26.1    42.1    57.9    78.1
 7 humid                 1         1.00    62.5       19.4      12.7    47.0    61.8    78.8   100
 8 wind_dir            460         0.982  200.       107.        0     120     220     290     360
 9 wind_speed            4         1.00    10.5        8.54      0       6.90   10.4    13.8  1048.
10 wind_gust         20778         0.204   25.5        5.95     16.1    20.7    24.2    28.8    66.7
11 precip               0         1        0.00447    0.0302    0       0       0       0       1.21
12 pressure          2729         0.896 1018.         7.42    984.    1013.   1018.   1023    1042.
13 visib                0         1        9.26       2.06      0      10      10      10      10
```

- [10 pt / 43 pts] Record the letter(s) of all the following that are **true**.

  (a) If listwise deletion would be used, there would be $n = 5,337$ observations left.

  (b) If `wind_gust` were dropped, there would be less imputations that would have to be made.

  (c) It would be good practice to drop the observations that have missingness in `temp`, `dewp` and `humid`.

  (d) If `wind_gust` were the response variable, then $n = 5,337$.

  (e) If `wind_gust` were the response variable, and listwise deletion was used, then $n \geq 2141$.

  (f) If imputation via `missForest` were used on the `weather` data frame, the resulting data frame would have $n = 26,115$ observations.

  If each of these columns contained measurements that were features in a model and we were to both impute and record missingness, the resulting $\boldsymbol{X}$ matrix in $\mathbb{D}$ would have ___ columns:

  (g) 13

  (h) 16

  (i) 20

  (j) 26

Your answer will consist of a string (e.g. `aebgd`) where the order of the letters does not matter nor does upper / lowercase.

BCDEFI

Problem 5 [11min] Let $p+1 = 501$ of features that are all standardized and $n > p+1$. Let $f(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta}$, the first entry of $\boldsymbol{x}$ is one always, let δ be a non-trivial amount of homoskedastic noise and $\boldsymbol{\beta} = [\beta_0 \ \beta_1 \ \ldots \ \beta_{500}]^\top = [-5 \ 5 \ 4 \ 3 \ 2 \ 1 \ \mathbf{0}_{495}]^\top$. Consider the usual definition of SSE as the sum of squared error, $\mathcal{H} = \left\{\boldsymbol{w}^\top \boldsymbol{x} \ : \ \boldsymbol{w} \in \mathbb{R}^{p+1}\right\}$ and all the following algorithms that produce an estimate of $\boldsymbol{\beta}$ via:

$$
\begin{aligned}
\mathcal{A}_1 \ &: \ \boldsymbol{b}^{\mathcal{A}_1} = \underset{\boldsymbol{w}\in\mathbb{R}^{p+1}}{\arg\min} \ \{SSE\} \\
\mathcal{A}_2 \ &: \ \boldsymbol{b}^{\mathcal{A}_2} = \underset{\boldsymbol{w}\in\mathbb{R}^{p+1}}{\arg\min} \ \left\{SSE + \sum_{j=1}^{p+1} |w_j|\right\} \\
\mathcal{A}_3 \ &: \ \boldsymbol{b}^{\mathcal{A}_3} = \underset{\boldsymbol{w}\in\mathbb{R}^{p+1}}{\arg\min} \ \left\{SSE + \sum_{j=1}^{p+1} w_j^2\right\} \\
\mathcal{A}_4 \ &: \ \boldsymbol{b}^{\mathcal{A}_4} = \underset{\boldsymbol{w}\in\mathbb{R}^{p+1}}{\arg\min} \ \left\{SSE + \frac{1}{2}\sum_{j=1}^{p+1} |w_j| + \frac{1}{2}\sum_{j=1}^{p+1} w_j^2\right\}
\end{aligned}
$$

- [12 pt / 55 pts] Record the letter(s) of all the following that are **true**.

  (a) $\boldsymbol{b}^{\mathcal{A}_1} = \left(\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{I}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y}$.

  (b) $\boldsymbol{b}^{\mathcal{A}_2} = \left(\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{I}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y}$.

  (c) $\boldsymbol{b}^{\mathcal{A}_3} = \left(\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{I}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y}$.

  (d) The OLS algorithm will fail in this setting.

  (e) $\mathcal{A}_4$ is elastic net regression with $\alpha = \frac{1}{2}$.

  (f) Out of sample, $\mathcal{A}_2, \mathcal{A}_3$ and $\mathcal{A}_4$ will have higher predictive performance than $\mathcal{A}_1$.

  (g) $\mathcal{A}_2$ likely produces an estimate of $\boldsymbol{\beta}$ with more zero entries than $\mathcal{A}_4$.

  (h) $\mathcal{A}_3$ likely produces an estimate of $\boldsymbol{\beta}$ with more zero entries than $\mathcal{A}_4$.

  (i) $\mathcal{A}_2$ is the most straightforward algorithm that could be used to locate variables that affect the response.

  (j) $\left|\left|\left[b_6^{\mathcal{A}_1}, b_7^{\mathcal{A}_1}, \ldots, b_{500}^{\mathcal{A}_1}\right]^\top\right|\right| < \left|\left|\left[b_6^{\mathcal{A}_2}, b_7^{\mathcal{A}_2}, \ldots, b_{500}^{\mathcal{A}_2}\right]^\top\right|\right|$

  (k) $\left|\left|\left[b_6^{\mathcal{A}_2}, b_7^{\mathcal{A}_2}, \ldots, b_{500}^{\mathcal{A}_2}\right]^\top\right|\right| < \left|\left|\left[b_6^{\mathcal{A}_3}, b_7^{\mathcal{A}_3}, \ldots, b_{500}^{\mathcal{A}_3}\right]^\top\right|\right|$

  (l) $\left|\left|\left[b_6^{\mathcal{A}_4}, b_7^{\mathcal{A}_4}, \ldots, b_{500}^{\mathcal{A}_4}\right]^\top\right|\right| < \left|\left|\left[b_6^{\mathcal{A}_1}, b_7^{\mathcal{A}_1}, \ldots, b_{500}^{\mathcal{A}_1}\right]^\top\right|\right|$

Your answer will consist of a string (e.g. `aebgd`) where the order of the letters does not matter nor does upper / lowercase.

CEFGIKL

Problem 6 [11min] Assume the response is a real number. In class we studied the following three decompositions of MSE in the modeling context where δ was realized from a mean-centered r.v. with variance σ^2 independent of the value of $\boldsymbol{x}$:

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

- [17 pt / 72 pts] Record the letter(s) of all the following that are **true**.

(a) In I and II, the new observation $\boldsymbol{x}_*$ was considered drawn from a random universe.

(b) In I and II, the training set $\mathbb{D}$ was considered drawn from a random universe.

(c) In II and III, the training set $\mathbb{D}$ was considered drawn from a random universe.

(d) In II, the responses $\boldsymbol{y}$ were considered drawn from a random universe.

(e) In III, the responses $\boldsymbol{y}$ were considered drawn from a random universe.

(f) In II, the design matrix $\boldsymbol{X}$ was considered drawn from a random universe.

(g) In III, the design matrix $\boldsymbol{X}$ was considered drawn from a random universe.

In general...

(h) the $\mathbb{B}\text{ias}\left[G(\boldsymbol{x}_*)\right]$ term is low for OLS.

(i) the $\mathbb{B}\text{ias}\left[G(\boldsymbol{x}_*)\right]$ term is low for CART.

(j) the $\mathbb{B}\text{ias}\left[G(\boldsymbol{x}_*)\right]$ term is low for a bag of CART models.

(k) the $\mathbb{B}\text{ias}\left[G(\boldsymbol{x}_*)\right]$ term is low for RF.

(l) the $\mathbb{V}\text{ar}\left[G(\boldsymbol{x}_*)\right]$ term is low for OLS.

(m) the $\mathbb{V}\text{ar}\left[G(\boldsymbol{x}_*)\right]$ term is low for CART.

~~(n) the $\mathbb{V}\text{ar}\left[G(\boldsymbol{x}_*)\right]$ term is low for bag of CART models.~~

~~(o) the $\mathbb{V}\text{ar}\left[G(\boldsymbol{x}_*)\right]$ term is low for RF.~~

(p) the $\mathbb{V}\text{ar}\left[G(\boldsymbol{x}_*)\right]$ term is lower for CART than it is for a bag of CART models.

(q) the $\mathbb{V}\text{ar}\left[G(\boldsymbol{x}_*)\right]$ term is lower for RF than it is for a bag of CART models.

Your answer will consist of a string (e.g. `aebgd`) where the order of the letters does not matter nor does upper / lowercase.

CDEGIJKLQ

Problem 7 [11min] Consider the following RF model output:

```
YARF v1.0 for classification
Missing data feature ON.
500 trees, training data n = 2000 and p = 99
Model construction completed within 0.08 minutes.
OOB results on all observations as a confusion matrix:
            predicted 0 predicted 1 model errors
actual 0       1425.000      77.000        0.051
actual 1        192.000     306.000        0.386
use errors        0.119       0.201        0.134
```

- [20 pt / 92 pts] Record the letter(s) of all the following that are **true**.

(a) $\mathcal{Y} = \{0, 1\}$.

(b) The features that are important in this model are simple to ascertain.

(c) The best estimate of this model's accuracy in-sample is less than 86.6%.

(d) The best estimate of this model's accuracy in-sample is more than 86.6%.

(e) The best estimate of this model's accuracy oos is 86.6%.

(f) This RF model will be more accurate oos if it was rebuilt with $M > 500$.

(g) This RF model will be less accurate oos if it was rebuilt with $M > 500$.

(h) This RF model will be more accurate if $p_{\text{try}} = 99$.

(i) When predicting in the future, if there were 100 samples where $y = 1$, this model would predict incorrectly 38.6% of the time.

(j) When predicting in the future, if there were 100 samples where $y = 1$, this model would predict incorrectly 20.1% of the time.

(k) When predicting in the future, if there were 100 samples where $\hat{y} = 1$, this model would predict incorrectly 38.6% of the time.

(l) When predicting in the future, if there were 100 samples where $\hat{y} = 1$, this model would predict incorrectly 20.1% of the time.

(m) The oos precision of this model is 79.9%.

(n) The oos recall of this model is 79.9%.

(o) The oos FDR estimate of this model is 11.9%.

(p) The oos FOR estimate of this model is 11.9%.

(q) If the cost of the false positive is greater than the cost of a false negative, this would be a good model to use in the real world.

(r) The AUC for this model can be calculated given the output above.

(s) The OOB results were calculated on about 2/3 of the total number of observations, $n = 2000$.

(t) The predictions from each of the $M = 500$ constituent trees in this model are positive correlated.

Your answer will consist of a string (e.g. `aebgd`) where the order of the letters does not matter nor does upper / lowercase.

ADEFILMPT

Problem 8 [11min] Let $\mathcal{Y} = \{0, 1\}$ and employ the following algorithm on training data $\mathbb{D}$:

$$\mathcal{A} \quad : \quad \boldsymbol{b} = \underset{\boldsymbol{w} \in \mathbb{R}^{p+1}}{\arg\min} \left\{ \prod_{i=1}^{n} \left(\frac{1}{1 + e^{-\boldsymbol{w}^\top \boldsymbol{x}_i}} \right)^{y_i} \left(\frac{1}{1 + e^{\boldsymbol{w}^\top \boldsymbol{x}_i}} \right)^{1-y_i} \right\}$$

- [13 pt / 105 pts] Record the letter(s) of all the following that are **true**.

  (a) This algorithm has a closed-form solution for $\boldsymbol{b}$ as a function of $\boldsymbol{X}$ and $\boldsymbol{y}$.

  (b) R^2 is an interpretable metric in this model.

  (c) This algorithm has hyperparameter(s) that need to be specified before running.

  (d) This algorithm is called "logistic regression".

  (e) The $\boldsymbol{b}$ that is returned by this algorithm may have numerical error.

  (f) This algorithm makes statistical assumptions about the training data.

  (g) The $\boldsymbol{b}$ returned by this algorithm allows us to compute probability estimates of $\mathbb{P}(Y_* = 1 \mid \boldsymbol{x}_*) = \boldsymbol{b}\boldsymbol{x}_*$.

  (h) If $\boldsymbol{b}\boldsymbol{x}_{*_1} > \boldsymbol{b}\boldsymbol{x}_{*_2}$ that implies that the model's estimate for the probability of the response being one for observation $\boldsymbol{x}_{*_1}$ is higher than for observations $\boldsymbol{x}_{*_2}$.

  Consider the following output of $\mathcal{A}$ on the training data $\mathbb{D}$:

```
> b
      (Intercept)          x1          x2
     -10.0175573   0.5971474   0.3090818
```

  (i) If $x_* = [0\ 0]$ then the probability estimate of $Y_* = 1$ will be very low.

  (j) Using the above output, one can compute the Brier score for training data $\mathbb{D}$.

  (k) Using the above output, one can compute an ROC curve for this model.

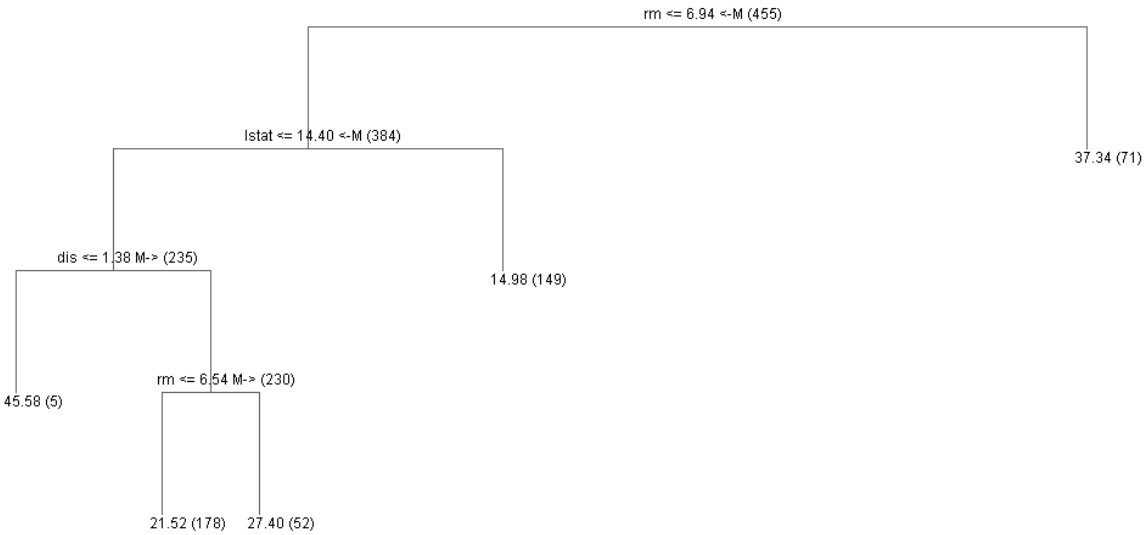
  (l) $g_0(\boldsymbol{x}) = 0$ for this model.

  (m) $g_0(\boldsymbol{x}) = \bar{y}$ for this model.

Your answer will consist of a string (e.g. `aebgd`) where the order of the letters does not matter nor does upper / lowercase.

DEFHIJKM

Problem 9 [11min] Consider the boston housing data that has response as the median home value and $n = 506$ and $p = 13$. We fit a regression tree model fit setting $N_0 = 200$. Here is the tree visualized:



- [12 pt / 117 pts] Record the letter(s) of all the following that are **true**.

  (a) This tree model can be written as a linear model.

  (b) This tree model has 5 nodes.

  (c) This tree model has 5 root nodes.

  (d) This tree model splits $\mathcal{X}$ into rectangular prisms (i.e. dimension 3).

  (e) This tree model is overfit.

  (f) If you label the leaves 1,2 ..., L from left to right in the above, and you must split further on one of them, then splitting on leaf #1 is most likely to yield a better tree model.

  (g) An OLS model will be more accurate out of sample than this model.

  (h) The most important feature in the data for predicting the response is `lstat`.

  (i) The model of the form $\hat{\boldsymbol{y}} = b_0 + b_1\boldsymbol{u}$ where is $\boldsymbol{u}$ restricted totr be one of the features in this dataset has highest R^2 when $\boldsymbol{u}$ is the variable `rm`.

  (j) The model of the form $\hat{\boldsymbol{y}} = b_0 + b_1\boldsymbol{u}$ where $\boldsymbol{u}$ is restricted to be dummy has highest R^2 when $\boldsymbol{u}$ is the computed via the function $\mathbb{1}_{rm \leq 6.94}$.

  (k) If `rm` $= 7$, the tree would predict 37.34 for y regardless of the values of the other 12 features.

  (l) If `rm` $= 6$, the tree would predict <37.34 for y regardless of the values of the other 12 features.

Your answer will consist of a string (e.g. `aebgd`) where the order of the letters does not matter nor does upper / lowercase.

ADGIJK

Problem 10 [11min] Let $p = 1$, $n > 100 + 1$, $d \in \{10, 11, \ldots, 100\}$, $\mathcal{A} =$ OLS and consider the following different complexity sets that are used within the algorithm:

$$\begin{aligned}
\mathcal{H}_1 &= \{w_0 + w_1 x_1 \ : \ w_0, w_1 \in \mathbb{R}\} \\
\mathcal{H}_2 &= \{w_0 + w_1 \ln(x_1) \ : \ w_0, w_1 \in \mathbb{R}\} \\
\mathcal{H}_3 &= \left\{w_0 + w_1 x_1 + w_2 x^2 \ : \ w_0, w_1, w_2 \in \mathbb{R}\right\} \\
\mathcal{H}_4 &= \left\{w_0 + w_1 x_1 + w_2 x^2 + \ldots + w_d x^d \ : \ w_0, w_1, \ldots, w_d \in \mathbb{R}\right\}
\end{aligned}$$

- [17 pt / 134 pts] Record the letter(s) of all the following that are **true**.

(a) The lowest RMSE will be from $\mathcal{H}_4$.

(b) The lowest oos RMSE will be from $\mathcal{H}_4$.

(c) The degrees of freedom in $\mathcal{H}_4$ equals d.

(d) b_1 is the same in both $\mathcal{H}_1$ and $\mathcal{H}_3$.

(e) If the response was logged, b_1 would be the same in both $\mathcal{H}_1$ and $\mathcal{H}_2$.

(f) The design matrix for $\mathcal{H}_2$ looks like $\boldsymbol{X} = \left[\mathbf{1}_n, [\ln(x_1) \ \ln(x_2) \ \ldots \ \ln(x_n)]^\top\right]$.

(g) The design matrix for $\mathcal{H}_3$ looks like $\boldsymbol{X} = \left[\mathbf{1}_n, [x_1^2 \ x_2^2 \ \ldots \ x_n^2]^\top\right]$.

(h) In $\mathcal{H}_4$, if $d = 100$, then $R^2 = 100\%$.

(i) In $\mathcal{H}_4$, if d was set to be large, predictions for $\boldsymbol{x}_* \in \mathcal{X}$ will likely be more accurate than $\boldsymbol{x}_* \notin \mathcal{X}$.

(j) In $\mathcal{H}_4$, if d was set to be large, predictions for $\boldsymbol{x}_* \notin \mathcal{X}$ will likely be worse than if d were set to be small.

(k) In $\mathcal{H}_4$, if d was varied across its range specified in the problem, then you can trace out the in-sample vs. oos complexity curve and locate the optimal model as a function of the hyperparameter d.

(l) Assuming you do the procedure in the previous question and locate the optimal model given by $d = d^\star$, the oos s_e from that procedure for the optimal model would be biased upwards.

(m) The oos s_e from that procedure for the optimal model would be biased downwards.

When comparing two mutually observed observations (A) and (B) sampled in the same fashion as those in $\mathbb{D}$, when the x measurement of (A) is _____ than the x measurement of (B) then (A) is predicted to have a _____ from the (B)'s assuming the model is true. The blanks would be filled in by...

(n) In $\mathcal{H}_1$, "one unit larger" / "response y that differs by b_1".

(o) In $\mathcal{H}_2$, "one unit larger" / "response y that differs by b_1".

(p) In $\mathcal{H}_2$, "10% larger" / "response y that differs by $0.1b_1$".

(q) In $\mathcal{H}_2$, "10% larger" / "response y 10% larger".

Your answer will consist of a string (e.g. `aebgd`) where the order of the letters does not matter nor does upper / lowercase.

AFIJKMNP

Problem 11 This question is about the concept of model validation and the strategy we discussed in class. Let's say we divide scramble the rows of $\mathbb{D}$ then create a partition

$$\mathbb{D} = \begin{bmatrix} \mathbb{D}_{\text{train}} \\ \hline \mathbb{D}_{\text{select}} \\ \hline \mathbb{D}_{\text{test}} \end{bmatrix}$$

in a 8:1:1 ratio train : select : test (in number of rows). We wish to select the best model g_{m_*} out of M candidate models $g_1, g_2, \ldots, g_M$ using the model selection procedure from class.

- [12 pt / 146 pts] Record the letter(s) of all the following that are **true**.

  (a) The three sets above could be permuted and the model selection procedure would still be valid.

  (b) The three sets above could be permuted and the model selection procedure would yield the same g_{m_*}.

  (c) The test set has 10% of the total dataset.

  (d) The select set has 11.11% of the total dataset.

  (e) When selecting models, the training data has 80% of the data used for exclusively the selection step of the model selection procedure.

  (f) If we were to use only inner CV, then there would be a total of $K = 10$ folds.

  (g) If we were to use only outer CV, then there would be a total of $K = 10$ folds.

  (h) If we were to not use any CV, then we would get an estimate of the future performance of one specific g_{m_*}.

  (i) If we were to use only inner CV, then we would get an estimate of the future performance of one specific g_{m_*}.

  (j) If we were to use only outer CV, then we would get an estimate of the future performance of one specific g_{m_*}.

  (k) If we were to use both inner and outer CV, then we would get an estimate of the future performance of one specific g_{m_*}.

  (l) If $\mathbb{D}_{\text{train}}$ was reduced in size, then we increase misspecification error among all M models during the model selection procedure.

Your answer will consist of a string (e.g. `aebgd`) where the order of the letters does not matter nor does upper / lowercase.

ACGHI