Math 342W / 650 Fall 2021
Midterm Examination Two

Professor Adam Kapelner

Thursday, May 13, 2021

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:

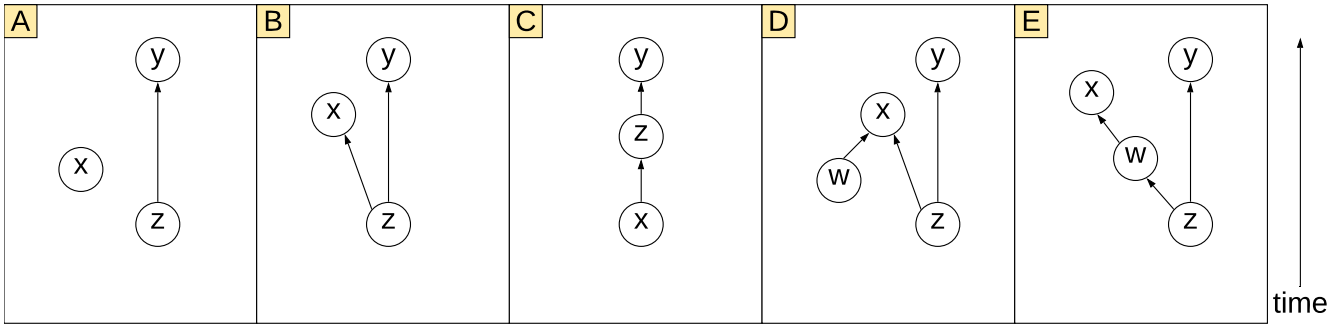
Cheating Using or attempting to use unauthorized assistance, material, or study aids in examinations or other academic work or preventing, or attempting to prevent, another from using authorized assistance, material, or study aids. Example: using an unauthorized cheat sheet in a quiz or exam, altering 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 is 100 minutes (variable time per question) and closed-book. You are allowed **two** 8.5 $\times$ 11" pages (front and back) of a "cheat sheet", blank scrap paper and a graphing calculator. Please read the questions carefully. No food is allowed, only drinks.

Problem 1 [11min] (and 11min will have elapsed) Consider the following causal diagrams where all events have other causes that are not displayed. It is assumed that the timing of events is known and to scale. Assume you have a training set $\mathbb{D}$ with a large sample size n with columns x, y, z (and w if included).



- [11 pt / 11 pts] Record the letter(s) of all the following that are **true** in general. At least one will be true.

  (a) Assume diagram A is the data generating process that $\mathbb{D}$ is sampled from. When running the OLS model y $\sim$ x you find a strong correlation. In out-of-sample data there would likely be a near-zero correlation betweeen y and x.

  (b) In diagram A, z is a lurking variable when analyzing the causal effect of x on y.

  (c) In diagram B, z is a lurking variable when analyzing the causal effect of x on y.

  (d) In diagram D, w is a lurking variable when analyzing the causal effect of x on y.

  (e) In diagram E, w is a lurking variable when analyzing the causal effect of x on y.

  (f) In diagram C, x causes y (according to our in-class definition of causality).

  (g) In diagram B, an OLS model of y $\sim$ x demonstrates that x and y are correlated but this correlation disappears if you run an OLS model of y $\sim$ x + z

  (h) In diagram D, an OLS model of y $\sim$ x demonstrates that x and y are correlated but this correlation disappears if you run an OLS model of y $\sim$ x + w

  (i) In diagram D, x can be used to predict y better than g_0 (out of sample)

  (j) In diagram D, w can be used to predict y better than g_0 (out of sample)

  (k) In diagram E, an OLS model of y $\sim$ x demonstrates that x and y are correlated but this correlation disappears if you run an OLS model of y $\sim$ x + w

Your answer will consist of a lowercase string (e.g. `aebgd`) where the order of the letters does not matter.

ACEFGIK

Problem 2 [7min] (and 18min will have elapsed) Consider the `iris` data frame, a famous dataset of four measurements on $n = 150$ iris flowers where the response is `Species`, a categorical variable with three levels: versicolor, viriginica and setosa (where each species has 50 observations). Below is a sample:

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1          5.5         4.2          1.4         0.2     setosa
2          7.6         3.0          6.6         2.1  virginica
3          7.2         3.2          6.0         1.8  virginica
4          5.6         3.0          4.1         1.3 versicolor
5          5.2         4.1          1.5         0.1     setosa
...
```

And also consider the `common_name` data frame which provides the common names:

```
Species      English_Name
aphylla      table iris
setosa       bristle-pointed iris
reichenbachii rock iris
versicolor   blue flag iris
flavescens   lemonyellow iris
```

- [8 pt / 19 pts] Record the letter(s) of all the following that are **true** in general. At least one will be true.

  Joining the two data frames on the `Species` column via a ...

  (a) ... left join (where `iris` was on the left) would yield a new data frame with 150 rows.

  (b) ... right join (where `iris` was on the right) would yield a new data frame with 150 rows.

  (c) ... left join (where `common_name` was on the left) would yield a new data frame with 150 rows.

  (d) ... right join (where `common_name` was on the right) would yield a new data frame with 150 rows.

  (e) ... inner join would yield a new data frame with 150 rows.

  (f) ... full join would yield a new data frame with 150 rows.

  (g) ... inner join would yield a new data frame with 100 rows.

  (h) ... full join would yield a new data frame with 100 rows.

Your answer will consist of a lowercase string (e.g. `aebgd`) where the order of the letters does not matter.

ABG

Problem 3 [6min] (and 24min will have elapsed) Consider the `iris` data frame, a famous dataset of four measurements on $n = 150$ iris flowers where the response is `Species`, a categorical variable with three levels: versicolor, viriginica and setosa (where each species has 50 observations). Consider the subset of the observations for only y = setosa. Below is a sample:

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1          5.2         4.1          1.5         0.1
2          5.4         3.4          1.7         0.2
3          5.3         3.7          1.5         0.2
4          4.5         2.3          1.3         0.3
5          4.8         3.0          1.4         0.3
6  ...
```

- [7 pt / 26 pts] Record the letter(s) of all the following that are **true** in general. At least one will be true.

  If you were to convert this dataset from wide format to long format ...

  (a) ... there would be two columns

  (b) ... there would be four columns

  (c) ... all columns would have continuous features

  (d) ... all columns would have categorical features

  (e) ... there would be 50 rows

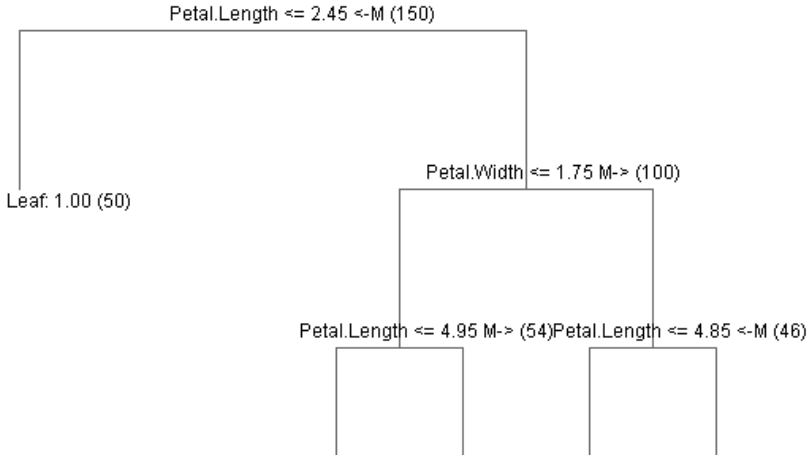
  (f) ... there would be 200 rows

  (g) ... the resulting data frame would be easier to manipulate using a visualization library such as `ggplot2`

Your answer will consist of a lowercase string (e.g. `aebgd`) where the order of the letters does not matter.

AFG

Problem 4 [9min] (and 33min will have elapsed) Consider the `iris` data frame, a famous dataset of four measurements on $n = 150$ iris flowers where the response is `Species`, a categorical variable with three levels: versicolor, viriginica and setosa (where each species has 50 observations). We fit a CART model with `Nodesize` = 1 to this dataset. The result is a model with 17 internal nodes and 9 leaf nodes with $\hat{y}$ values of 1 (= setosa), 2 (= virginica) and 3 (= versicolor). Below is an abridged illustration of the tree. Ignore the "M→" and "←M" notation. Numbers in parentheses indicate number of observation in a node. The left direction means the split condition is true.



- [14 pt / 40 pts] Record the letter(s) of all the following that are **true** in general. At least one will be true.
  - (a) This is a regression tree model
  - (b) This model can be written as a linear model of the form $g(\boldsymbol{x}) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4$ where x_1, x_2, x_3, x_4 are the four measurements `Sepal.Length`, `Sepal.Width`, `Petal.Length` and `Petal.Width`
  - (c) This model is overfit
  - (d) If the stump is considered depth zero, then this tree has a depth of 3
  - (e) For all $\mathbb{D}$, a binary split on `Petal.Length` is the best overall split to reduce heterogeneity in the response
  - (f) For future data, this model will likely predict setosa correctly with high probability
  - (g) If `Petal.Length` > 2.45, there exists a second binary split that is able to isolate all virginica observations in one leaf and all versicolor observations in the other leaf
  - (h) If `Petal.Length` > 4.95, $\hat{y}$ will always be the same

Your answer will consist of a lowercase string (e.g. `aebgd`) where the order of the letters does not matter.

BCEF

Problem 5 [6min] (and 39min will have elapsed) Consider the following data frame displayed in random order, where missingess is visualized in red. The last column is the response y and there is no missingness in the response. The feature that has the most missingness is x_{34} and the observation with the most missingness is row #10.

- [7 pt / 47 pts] Record the letter(s) of all the following that are **true** in general. At least one will be true.

  (a) Dropping all observations with missingness (listwise deletion) will seriously impact future performance of any model fit with the data

  (b) The dataset exhibits one MCAR missingness mechanism and this mechanism is the same for all entries

  (c) The dataset may exhibit a different independent NMAR missingness mechanism for each feature

  (d) When building a predictive model, the most prudent thing to do is to drop the feature x_{34} and to drop observation #10

  (e) Imputing missing values using the columns' sample averages will result in a data frame with no missingness

  (f) The recommended procedure is to impute with the `missForest` and then use only the imputed matrix $\boldsymbol{X}$ to construct your model ignoring the original $\boldsymbol{X}$ matrix

Your answer will consist of a lowercase string (e.g. `aebgd`) where the order of the letters does not matter.

ACDE

Problem 6 [6min] (and 45min will have elapsed) Consider the following data frame, where missingness is visualized in red. The last column is the response y and there is no missingness in the response. The feature that has the most missingness is x_{34} and the observation with the most missingness is row #10.

Now assume that x_{34} and observation #10 were dropped and the remaining missing values were imputed via the `missForest` algorithm.

- [6 pt / 53 pts] Record the letter(s) of all the following that are **true** in general. At least one will be true.

  (a) An OLS model cannot be fit on the imputed data frame

  (b) A ridge model with $\lambda = 10^{-6}$ cannot be fit on the imputed data frame

  (c) A lasso model with $\lambda = 10^{-6}$ cannot be fit on the imputed data frame

  (d) An elastic net model with $\lambda = 10^{-6}$ and $\alpha = 0.1$ cannot be fit on the imputed data frame

  (e) After using the model selection procedure to select λ for the lasso based on oos s_e, the number of nonzero linear coefficients in the resulting model will likely be small relative to p

  (f) After using the model selection procedure to select λ for the ridge based on oos s_e, the number of nonzero linear coefficients in the resulting model will likely be small relative to p

Your answer will consist of a lowercase string (e.g. `aebgd`) where the order of the letters does not matter.

AE

Problem 7 [14min] (and 59min will have elapsed) Consider the `diamonds` data frame, which has $n = 53940$ diamonds with 9 measurements each and the response variable `price` measured in USD. We wish to fit many models to this dataset: (1) OLS with the formula $y \sim .$ (2) OLS with the formula $y \sim . * .$ (3) OLS with the formula $y \sim . * . * .$ (4) a regression tree with `nodesize` $= 10$ (5) a Random Forest (RF) with 500 trees. Consider two model selection procedures:

A. The three-split dataset model selection procedure where the original dataset is sampled into a distinct training set of size $n = 3,000$, a distinct select set of $n = 3,000$ and a distinct test set of $n = 3,000$.

B. The *nested resampling procedure* for model selection where $K_{select} = 3$ and $K_{test} = 5$ on a subset of $n = 9,000$ observations from the original data frame.

- [16 pt / 69 pts] Record the letter(s) of all the following that are **true** in general. At least one will be true.

  (a) In procedure [A], the training set is fit to each of the five models above
  (b) In procedure [A], the the select set is fit to each of the five models above
  (c) In procedure [A], predicting on the training set is "out of sample"
  (d) In procedure [A], predicting on the select set is "out of sample"
  (e) In procedure [A], each of the five models above predict on the test set
  (f) In procedure [A], the model that predicts on the test set was fit on $n = 6000$ observations
  (g) In procedure [A], the final model was fit on $n = 6000$ observations
  (h) Procedure [B] is more computationally costly than procedure [A]
  (i) Procedure [B] has lower variance in its estimate of future performance
  (j) Procedure [B] is likely to select a model with better future performance than procedure [A]
  (k) There are a total of 16 models fit during procedure [B] including the final model
  (l) There are a total of 26 models fit during procedure [B] including the final model
  (m) There are a total of 65 models fit during procedure [B] including the final model
  (n) There are a total of 76 models fit during procedure [B] including the final model
  (o) During procedure [B], for each of the K_{test} outer resamplings, there could be a different model that predicts on the test set
  (p) In procedure [B], the final model will be the modal model out of the K_{test} models selected

Your answer will consist of a lowercase string (e.g. `aebgd`) where the order of the letters does not matter.

ADFHIJO

Problem 8 [8min] (and 67min will have elapsed) Consider the `diamonds` data frame, which has $n = 53940$ diamonds with 9 measurements each and the response variable `price` measured in USD. Beginning with the design matrix $\tilde{\boldsymbol{X}}$ created by the OLS model of $y \sim .*.*.$ we consider the *greedy forward stepwise* linear model algorithm. The dataset is sampled into a distinct training set of size $n = 3,000$, a distinct select set of $n = 3,000$ and a distinct test set of $n = 3,000$ and these three subsets are used in accordance with the modeling selection procedure we learned about in class.

- [7 pt / 76 pts] Record the letter(s) of all the following that are **true** in general. At least one will be true.

  (a) This algorithm at most has $p + 1$ iterations where $p + 1$ is the number of columns in $\tilde{\boldsymbol{X}}$ and in each iteration, it fits a different linear model

  (b) This algorithm likely has less than $p + 1$ iterations

  (c) The prediction error in the training set is monotonically decreasing

  (d) The prediction error in the select set is monotonically decreasing

  (e) The prediction error in the test set is monotonically decreasing

  (f) When predicting on the test set, the model will definitely have the same number of degrees of freedom as the final model

  (g) The future performance of the model selected by the *greedy forward stepwise* linear model algorithm will be better than the future performance of the OLS model of $y \sim .$ based on your knowledge of this dataset from class

Your answer will consist of a lowercase string (e.g. `aebgd`) where the order of the letters does not matter.

ABCG

Problem 9 [7min] (and 74min will have elapsed) Consider the `diamonds` data frame, which has $n = 53,940$ diamonds with 9 measurements each and the response variable `price` measured in USD. Beginning with the design matrix $\tilde{\boldsymbol{X}}$ created by the OLS model of $y \sim .*.*$. we consider the *greedy forward stepwise* linear model algorithm. The dataset is sampled into a distinct training set of size $n = 3,000$, a distinct select set of $n = 3,000$ and a distinct test set of $n = 3,000$ and these three subsets are used in accordance with the modeling selection procedure we learned about in class. The design matrix $\tilde{\boldsymbol{X}}$ has $p + 1 = 1477$ number of columns. Consider the case where we run through all of the $t = 1, 2, \ldots, 1477$ iterations of this stepwise algorithm.

- [7 pt / 83 pts] Record the letter(s) of all the following that are **true** in general. At least one will be true.

  Let g_t denote the model fit at every iteration t.

  (a) As t increases the bias of g_t increases monotonically

  (b) As t increases the variance of g_t increases monotonically

  (c) As t increases the MSE of g_t increases monotonically

  (d) $g_1, g_2, \ldots, g_{1477}$ are independent models

  Consider the model g_{avg} that averages models $g_1, g_2, \ldots, g_{1477}$.

  (e) g_{avg} is called a "bagged" model

  (f) g_{avg} has lower bias than g_{1477}

  (g) g_{avg} always has lower MSE than the model that was selected by the greedy forward stepwise linear model algorithm

Your answer will consist of a lowercase string (e.g. `aebgd`) where the order of the letters does not matter.

B

Problem 10 [12min] (and 86min will have elapsed) Consider the `adult` data frame, which has data on $n = 32,560$ people with 14 measurements each and the response variable `income` which is binary (1 = the person has an income >50K and 0 = the person has an income ≤50K). Consider the following model fit to `adult_train`, a training set of $n = 10,000$ leaving the remainder of the data as a holdout set. The $\boldsymbol{b}$ vector for the fitted model is displayed below on the last line.

```
> lmod = glm(income ~ age + hours_per_week + capital_gain + education_num, adult_train, family = "binomial")
Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
> coef(lmod)
   (Intercept)             age  hours_per_week    capital_gain   education_num
       -8.1582          0.0454          0.0380          0.0003          0.3200
```

- [12 pt / 95 pts] Record the letter(s) of all the following that are **true** in general. At least one will be true.

  (a) The vector $\boldsymbol{b}$ was computed solely using linear algebra calculations

  (b) For a future person $\boldsymbol{x}_*$, the quantity $\boldsymbol{x}_*\boldsymbol{b}$ can be used to produce probability predictions in the set (0, 1)

  (c) For a future person $\boldsymbol{x}_*$, if $\boldsymbol{x}_*\boldsymbol{b} < 0$, this model is predicting that it is impossible for this person's income to be >50K

  (d) The Brier score for this model's predictions in-sample will likely be closer to zero than the Brier score for predictions out-of-sample.

  (e) The probability predictions of this model will be the same predictions as a probit model fit to the same data

  (f) "For a future person $\boldsymbol{x}_*$, the probability that this person's income >50K will increase by 0.044 if the person becomes one year older".

  (g) "For a future person $\boldsymbol{x}_*$, the probability that this person's income >50K will increase by 0.044 if the person becomes one year older as long as the three other measurements for this person do not change".

  (h) If `age` increases and the three other variables remain the same, the predicted probability will increase for any $\boldsymbol{x}$

  (i) The warning message means there was numerical underflow and/or overflow during the computation of $\boldsymbol{b}$

  (j) For a 20yr-old with no job, capital gains or education, the probability he has an income of >50K is 0.9993 to the nearest four significant digits

  (k) For a 20yr-old with no job, capital gains or education, the probability he has an income of >50K is 0.0007095 to the nearest four significant digits

  (l) You have all the information necessary in this problem statement to trace out a DET curve

Your answer will consist of a lowercase string (e.g. `aebgd`) where the order of the letters does not matter.

BDHIKL

Problem 11 [14min] (and 100min will have elapsed) Consider the `adult` data frame, which has data on $n = 32,560$ people with 14 measurements each and the response variable `income` which is binary (1 = the person has an income >50K and 0 = the person has an income ≤50K). Considering the model from the previous quesiton fit to `adult_train`, a training set of $n = 10,000$ leaving the remainder of the data as a holdout set, we now predict on `adult_test` and perform binary classification:

```
> yhat = as.numeric(predict(lmod, adult_test, type = "response") > 0.9)
> table(y_test, yhat)
              yhat
y_test        0     1
      0   15107    30
      1    4439   585
```

- [9 pt / 104 pts] Record the letter(s) of all the following that are **true** in general. At least one will be true.

  (a) This is likely an asymmetric cost classification model

  (b) The FDR, FOR, FPR calculated using the above table are honest estimates of future performance

  (c) This model makes mistakes 22.2% of the time to the nearest three significant digits

  (d) This classification model is absolutely the best classification model you could build from `lmod` to minimize costs when the cost of a FP is \$9 and the cost of a FN is \$9

  (e) This model implies the point (0.00198, 0.116) on an ROC curve to the nearest three significant digits

  (f) When predicting in the future using this classification model, if $\hat{y} = 1$, then the probability of making an error is 4.88% to the nearest three signification digits

  (g) If the cost of a false positive is \$100 and the cost of false negatives was negligible and there is no reward for classifying correctly, then you expect to lose \$0.15 per prediction to the nearest two signification digits

  (h) If the probability threshold was increased within the binary classification model, then the number of FP's would decrease

  (i) You have all the information necessary in this problem statement to compute the AUC for `lmod`

Your answer will consist of a lowercase string (e.g. `aebgd`) where the order of the letters does not matter.

ABCEFGH