Math 342W / 650 Fall 2021
Midterm Examination One

Professor Adam Kapelner

Thursday, March 25, 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 × 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 [8min] (and 8min will have elapsed) George Box and Norman Draper in 1987 wrote “All models are wrong but some are useful”. Below are some conceptual questions about this aphorism and modeling in general.

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

  (a) “models are wrong” since their predictions are not exactly equal to the measurements.

  (b) In the quote, “models are wrong” since they are approximations.

  (c) In the quote, “models are wrong” since they can never be validated.

  (d) In the quote, “models are wrong” since they cannot be learned from data.

  (e) In the quote, “models are wrong” since some are non-mathematical.

  (f) In the quote, “models are wrong” since the prediction target is not well-defined.

  (g) In the quote, “some [models] are useful” since they can be validated.

  (h) In the quote, “some [models] are useful” since they can be learned from data.

  (i) In the quote, “some [models] are useful” since their predictions are “good enough” (where the builder of the model must define exactly what “good enough” means).

  (j) In the quote, “some [models] are useful” because they use precise measurements.

  (k) In the quote, “some [models] are useful” because they are better to use than a naive guess of what the phenomenon will be in a given setting.

  (l) In the quote, “some [models] are useful” because they can perform both regression and binary classification simultaneously.

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

abik

Problem 2 [7min] (and 15min will have elapsed) We examine the famous aphorism "absence makes the heart grow fonder" as g, a model for reality. In case you aren't a poet, "fonder" means "more in love".

- [11 pt / 23 pts] Record the letter(s) of all the following that are **true** in general. At least one will be true.

  (a) There is one target of prediction, y: the degree of the heart's fondness.

  (b) There is one setting, x: the amount of absence.

  (c) The model as stated is mathematical.

  (d) Comparing a binary metric for the degree of the heart's fondness ~~is~~ to a continuous metric for the degree of the heart's fondness, the more accurate metric for the degree of the heart's fondness is continuous.

  (e) The most accurate reading of the aphorism indicates that absence is a binary metric.

  (f) The most accurate reading of the aphorism indicates that absence is a continuous metric.

  (g) This aphorism describes to the reader all of the z's.

  (h) δ will be very large relative to f.

  (i) g is monotonic.

  (j) f is monotonic.

  (k) After establishing metrics and their means of measurement, a mathematical model can be learned from data.

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

abdfhik

Problem 3 [13min] (and 28min will have elapsed) Consider a dataset of n observations.

- [18 pt / 41 pts] Record the letter(s) of all the following that are **true** in general. At least one will be true. Let $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{0, 1\}$ and the dataset has n unique values of x.

  (a) Modeling y is called a "binary classification" problem.

  (b) f must be monotonic.

  (c) g must be monotonic.

  (d) $\mathcal{H} = \{wx : w \in \mathbb{R}\}$ is a reasonable model candidate set.

  (e) $\mathcal{H} = \{w_0 + w_1 x : w_0, w_1 \in \mathbb{R}\}$ is a reasonable model candidate set.

  (f) g_0 is the sample mode of all y observations.

  (g) Any model g will have nonzero δ.

  ~~(h) This model is likely to have estimation error.~~ could be argued both ways

  (i) A reasonable error metric for this model is misclassification error.

  (j) A reasonable error metric for this model is hinge error.

  For the remaining questions in this problem, let $\mathcal{Y} = \mathbb{R}$, $\mathcal{X} = \{0, 1\}$ and the dataset has n unique values of y.

  (k) $\mathcal{H} = \{wx : w \in \mathbb{R}\}$ is reasonable regardless of $\mathcal{A}$.

  (l) $\mathcal{H} = \{w_0 + w_1 x : w_0, w_1 \in \mathbb{R}\}$ is reasonable regardless of $\mathcal{A}$.

  (m) g_0 is the sample mode of all y observations.

  ~~(n) g must be of the form $g(x) = a$ if $x = 0$ and $g(x) = b$ if $x = 1$ where $a, b \in \mathbb{R}$.~~ ambiguously written

  (o) Any model g will have nonzero δ.

  (p) Any model g will have two parameters (i.e. two degrees of freedom).

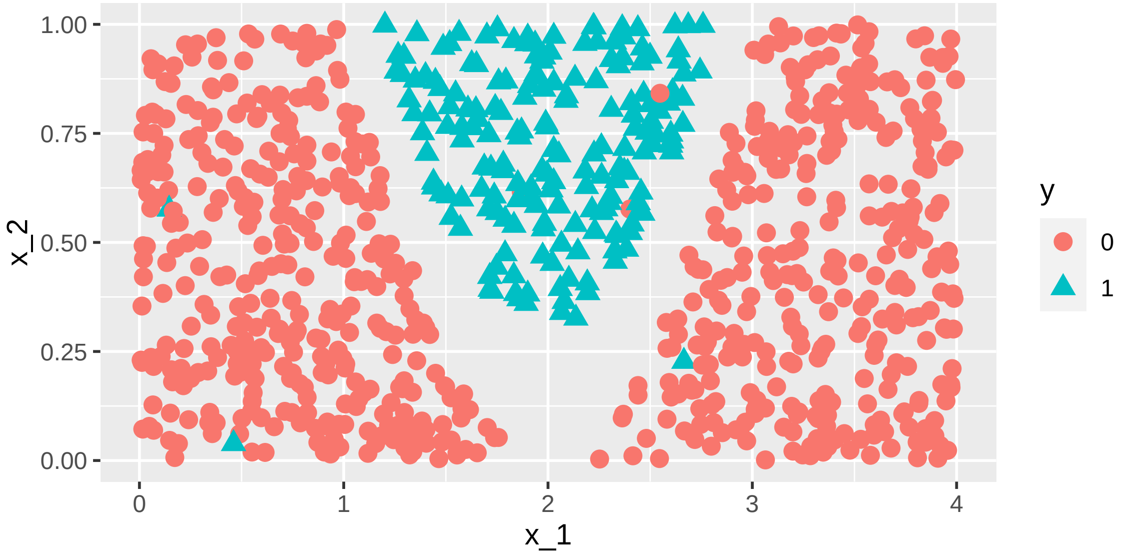
  (q) To decrease misspecification error, we can collect more data (n increases).

  (r) A reasonable error metric for this model is hinge error.

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

afijlo

Problem 4 [5min] (and 33min will have elapsed) $\mathbb{D}$ is illustrated below:



- [7 pt / 48 pts] Record the letter(s) of all the following that are **true** in general. At least one will be true.

(a) Modeling y is called a "binary classification" problem.

(b) $p_{raw} = 2$

(c) $p_{raw} = 3$

(d) The dataset is "linearly separable".

(e) The perceptron algorithm uses a default $\mathcal{H} = \{\mathbb{1}_{w_0 + w_1 x + w_2 x_2 \geq 0} : w_0, w_1, w_2 \in \mathbb{R}\}$

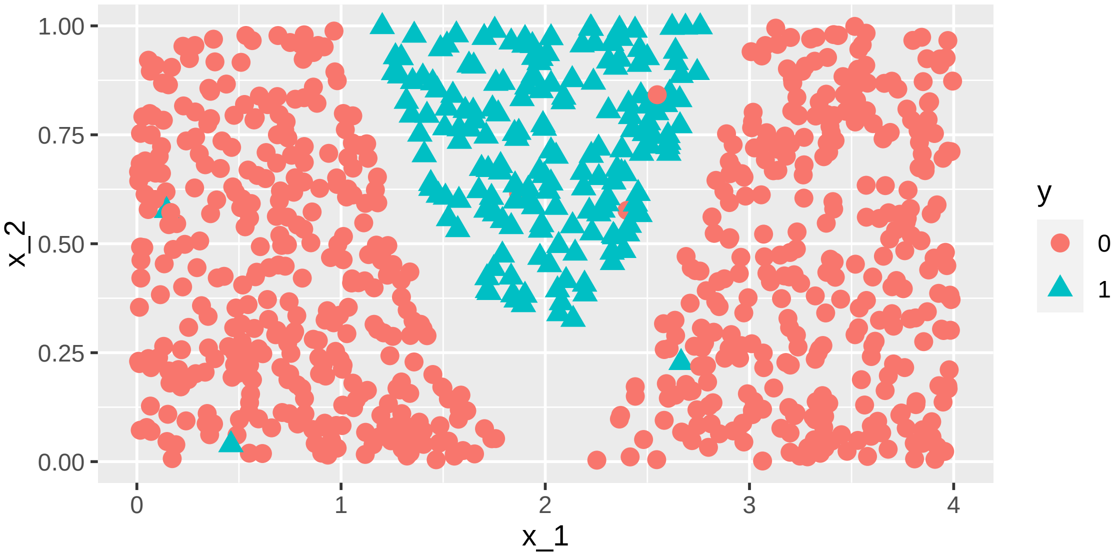
(f) The perceptron algorithm with the default $\mathcal{H}$ run with a max number of iterations of $3,000$ will converge.

(g) The perceptron algorithm with the default $\mathcal{H}$ run with a max number of iterations of $3,000$ will provide a g that can be used for prediction.

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

abeg

Problem 5 [8min] (and 41min will have elapsed) $\mathbb{D}$ is illustrated below and now consider $\mathcal{H}_{new} = \left\{ \mathbb{1}_{[1\ x_1\ x_1^2\ x_2\ x_2^2]^\top \boldsymbol{w} \geq 0} : \boldsymbol{w} \in \mathbb{R}^5 \right\}$. $\mathbb{D}$ has $n = 1000$ with 672 observations where $y = 0$.



- [7 pt / 55 pts] Record the letter(s) of all the following that are **true** in general. At least one will be true.

  (a) In contrast to the default $\mathcal{H}$, $\mathcal{H}_{new}$ now includes polynomial transformations.

  (b) $f \in \mathcal{H}_{new}$.

  (c) Under $\mathcal{H}_{new}$, we have $g_0(x_1, x_2) = 0$.

  (d) A maximum-margin hyperplane model algorithm using $\mathcal{H}_{new}$ will converge to a solution for the five parameters $\boldsymbol{w}$.

  ~~(e) Minimizing the objective function $AHE + \lambda||\boldsymbol{w}||^2$ where AHE is average hinge error for the five parameters $\boldsymbol{w}$ will converge to a solution.~~ we never covered this in class

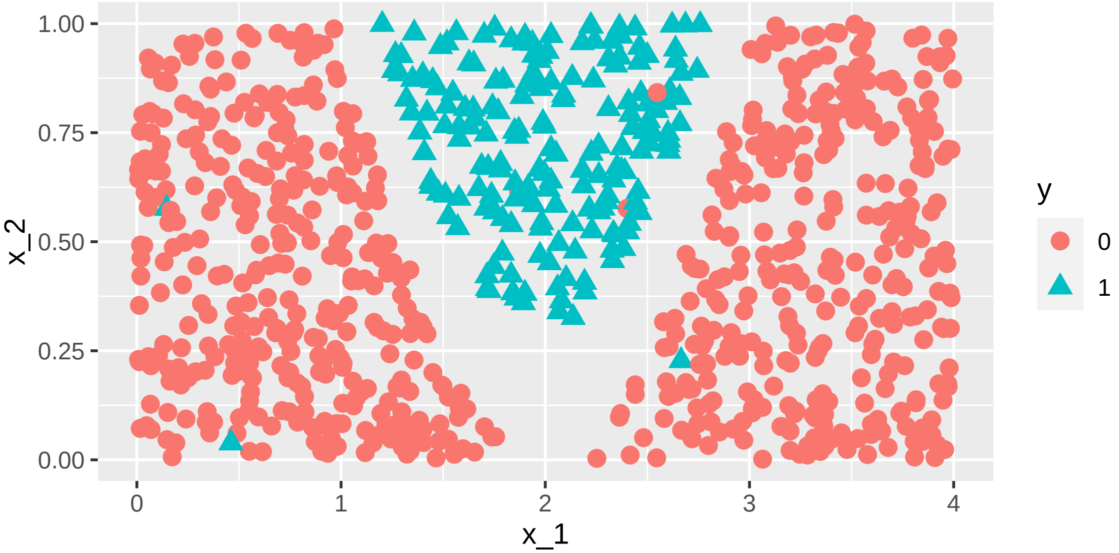
  (f) Minimizing the objective function $AHE + \lambda||[w_1\ w_2\ w_3\ w_4]||^2$ where AHE denotes the average hinge error for the five parameters $\boldsymbol{w}$ will converge to a solution.

  (g) The value of λ in (e) and (f) is specified before the model is fit.

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

acdfg

Problem 6 [5min] (and 46min will have elapsed) $\mathbb{D}$ is illustrated below and now consider $\mathcal{H}_{new} = \left\{ \mathbb{1}_{[1 \; x_1 \; x_1^2 \; x_2 \; x_2^2]^\top \boldsymbol{w} \geq 0} : \boldsymbol{w} \in \mathbb{R}^5 \right\}$. $\mathbb{D}$ has $n = 1000$ with 672 observations where $y = 0$.



- [5 pt / 60 pts] Record the letter(s) of all the following that are **true** in general. At least one will be true.

  Consider the situation where you remove the points from $\mathbb{D}$ that would allow for separability using a model from $\mathcal{H}_{new}$, fit a maximum margin hyperplane g, then add those points back into $\mathbb{D}$.

  Note: only one of (a), (b), (c) is true.

  (a) The order of magnitude of the average hinge error in $\mathbb{D}$ for g is 10^{-1} in units of y.

  (b) The order of magnitude of the average hinge error in $\mathbb{D}$ for g is 10^{-2} in units of y.

  (c) The order of magnitude of the average hinge error in $\mathbb{D}$ for g is 10^{-3} in units of y.

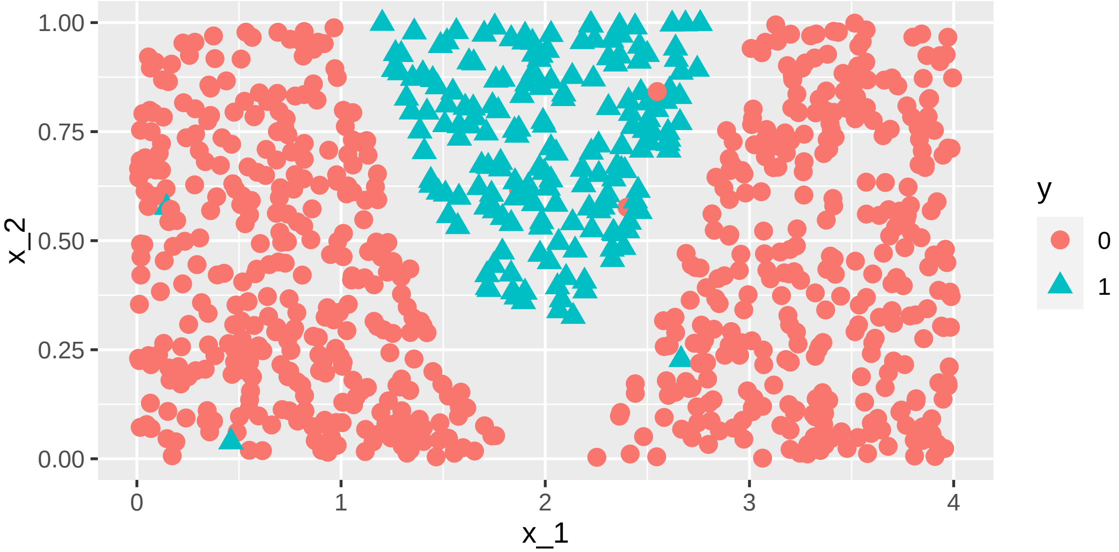
  (d) If K-fold cross validation were employed where $K = 10$, the out of sample misclassification error will be $\approx 0.5\%$.

  (e) If K-fold cross validation were employed where $K = 5$, the out of sample misclassification error will be $\approx 0.5\%$.

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

bde

Problem 7 [6min] (and 52min will have elapsed) $\mathbb{D}$ is illustrated below. $\mathbb{D}$ has $n = 1000$ with 672 observations where $y = 0$. We now fit g using the KNN algorithm with the default distance metric.



- [6 pt / 66 pts] Record the letter(s) of all the following that are **true** in general. At least one will be true.

(a) If K in the KNN algorithm was set to be 1, then there would be zero in-sample misclassification error.

(b) If K in the KNN algorithm was set to be 1, then there would be zero out-of-sample misclassification error.

(c) If K in the KNN algorithm was set to be 7, then there would be zero in-sample misclassification error.

(d) If K in the KNN algorithm was set to be 1, then $g(0.5, 0.5) = 0$.

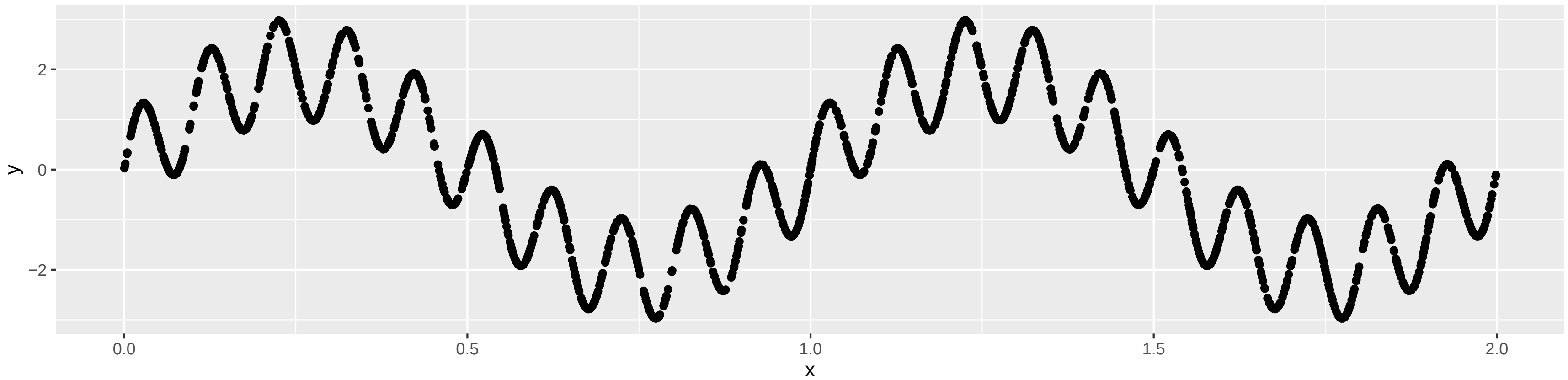
(e) If K in the KNN algorithm was set to be 1, then $g(2, 0.13) = 0$.

(f) If K in the KNN algorithm was set to be 100, then $g(2, 0.13) = 0$.

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

ad

Problem 8 [8min] (and 60min will have elapsed) Let the random variable X and Y be the models that realized the rows in $\mathbb{D}$. An $n = 2,000$ example $\mathbb{D}$ is housed in an R `data.frame` object called `Xy`. Below is a plot of this data frame.



- [8 pt / 74 pts] Record the letter(s) of all the following that are **true** in general. At least one will be true.

(a) X are Y are correlated.

(b) X are Y are associated.

(c) Running `cov(Xy$x, Xy$y)` in R would return zero.

too much duplication

~~(d) Running `coef(lm(y ~ x))` in R would return a vector of dimension two where the values are both near zero.~~

(e) The model produced by `lm(y ~ x)` in R suffers mostly from misspecification error.

(f) A linear polynomial model of degree 5 would produce a model with lower out of sample error than the model produced by `lm(y ~ x)` in R.

(g) A linear polynomial model of degree 5 has risky predictive performance when extrapolating.

(h) It is reasonable to believe that $z \approx x$ and that $t \approx f$ in this case (at least within $\mathcal{X} = [0, 2]$).

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

abefgh

Problem 9 [14min] (and 74min will have elapsed) Let $\boldsymbol{X} = [\boldsymbol{x}_{\cdot 0} \mid \boldsymbol{x}_{\cdot 1} \mid \ldots \mid \boldsymbol{x}_{\cdot p}] \in \mathbb{R}^{n \times (p+1)}$, rank $[\boldsymbol{X}] = p+1$, let $\boldsymbol{x}_i$ denote the ith row of the matrix $\boldsymbol{X}$ and $\boldsymbol{y} \in \mathbb{R}^n$. Your modeling task is to model the response using the n observations. All notation is standard ~~form~~ class and we consider:

$$\begin{aligned}
\mathcal{H} &= \left\{\boldsymbol{w}^\top \boldsymbol{x} : \boldsymbol{w} \in \mathbb{R}^{p+1}\right\} \\
\boldsymbol{b} &= \underset{\boldsymbol{w} \in \mathbb{R}^{p+1}}{\arg\min} \left\{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})\right\} \\
\hat{y}_i &= g(\boldsymbol{x}_i) = \boldsymbol{x}_i \boldsymbol{b}
\end{aligned}$$

- [15 pt / 89 pts] Record the letter(s) of all the following that are **true** in general. At least one will be true.

(a) The algorithm $\mathcal{A}$ that returns g minimizes $\sum_{i=1}^n e_i^2$.

(b) The algorithm $\mathcal{A}$ that returns g minimizes $\sum_{i=1}^n \mathcal{E}_i^2$.

(c) The algorithm $\mathcal{A}$ that returns g is called "ordinary least squares" (OLS) regression if $\boldsymbol{x}_{\cdot 0} = \boldsymbol{1}_n$.

(d) The vector $\hat{\boldsymbol{y}} := [\hat{y}_1, \ldots, \hat{y}_n]^\top$ is in the span of the columns of $\boldsymbol{X}$.

(e) If p is substantially less than n, there would be no overfitting in this model.

(f) SST = SSR + SSE

(g) If $p = n - 1$, then $R^2 = 1$.

(h) This algorithm can accomodate additional columns in $\boldsymbol{X}$ that are log transformations of original columns in $\boldsymbol{x}$.

(i) rank $[\boldsymbol{H}] = p + 1$

(j) rank $[\boldsymbol{H}\boldsymbol{X}] = n$

(k) $\boldsymbol{H}\boldsymbol{Q} = \boldsymbol{X}$

(l) colsp $[\boldsymbol{H}]$ = colsp $[\boldsymbol{Q}]$

(m) If $\exists j$ such that $\boldsymbol{y} = \boldsymbol{x}_{\cdot j}$, then MSE = 0.

(n) $\exists \boldsymbol{w} \in \mathcal{H}$ where $\boldsymbol{w} \neq \boldsymbol{b}$ and this $\boldsymbol{w}$ provides a higher RMSE.

(o) Consider $\boldsymbol{A}$, an $n \times (2p+2)$ matrix of $\boldsymbol{X}$ and $\boldsymbol{Q}$ column-binded together. $\boldsymbol{A}$ is full rank.

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

acdfghilmn

Problem 10 [7min] (and 81min will have elapsed) Let $\boldsymbol{X} = [\boldsymbol{x}_{\cdot 0} \mid \boldsymbol{x}_{\cdot 1} \mid \ldots \mid \boldsymbol{x}_{\cdot p}] \in \mathbb{R}^{n \times (p+1)}$, rank $[\boldsymbol{X}] = p + 1$, let $\boldsymbol{x}_i$ denote the ith row of the matrix $\boldsymbol{X}$ and $\boldsymbol{y} \in \mathbb{R}^n$. Your modeling task is to model the response using the n observations. All notation is standard form class and we consider:

$$\begin{aligned}\mathcal{H} &= \left\{\boldsymbol{w}^\top \boldsymbol{x} : \boldsymbol{w} \in \mathbb{R}^{p+1}\right\} \\ \boldsymbol{b} &= \underset{\boldsymbol{w} \in \mathbb{R}^{p+1}}{\arg\min} \left\{\left|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\right|^\top \boldsymbol{1}_n\right\} \\ \hat{y}_i &= g(\boldsymbol{x}_i) = \boldsymbol{x}_i \boldsymbol{b}\end{aligned}$$

- [9 pt / 98 pts] Record the letter(s) of all the following that are **true** in general. At least one will be true.

(a) The algorithm $\mathcal{A}$ that returns g minimizes $\sum_{i=1}^n e_i^2$.

(b) The algorithm $\mathcal{A}$ that returns g minimizes $\sum_{i=1}^n \mathcal{E}_i^2$.

(c) The algorithm $\mathcal{A}$ that returns g is called "ordinary least squares" (OLS) regression if $\boldsymbol{x}_{\cdot 0} = \boldsymbol{1}_n$.

(d) The vector $\hat{\boldsymbol{y}} := [\hat{y}_1, \ldots, \hat{y}_n]^\top$ is in the span of the columns of $\boldsymbol{X}$.

(e) If p is substantially less than n, there would be no overfitting in this model.

(f) SST = SSR + SSE

(g) If $p = n - 1$, then $R^2 = 1$.

(h) This algorithm can accomodate additional columns in $\boldsymbol{X}$ that are log transformations of original columns in $\boldsymbol{x}$.

(i) This algorithm can accomodate additional columns in $\boldsymbol{X}$ that are first-order interaction transformations of original columns in $\boldsymbol{x}$.

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

dghi

Problem 11 [8min] (and 89min will have elapsed) Consider a continuous response fit by OLS. During lecture we illustrated in-sample s_e and out-of-sample s_e as a function of "model complexity" for a given dataset of size n with $p_{raw} < n$ features. Below are questions related to this illustration. The quantity K is the value that controls the size of the train-test split setting as we discussed in class.

- [7 pt / 105 pts] Record the letter(s) of all the following that are **true** in general. At least one will be true.

(a) In-sample s_e converges to zero as the number of features increases to n.

(b) Logging y and rerunning the model will not change the in-sample s_e curve.

(c) Out-of-sample s_e is an honest metric of future performance for any valid K.

(d) Out-of-sample s_e is always smaller (for any degree of model complexity) if we employ K-fold CV.

(e) The out-of-sample s_e curve is smoother if we employ K-fold CV if compared to not performing K-fold CV.

(f) If you were to begin with p_{raw} features and then add columns consisting of random noise until there are a total of n columns, the minimum of the out-of-sample s_e curve would be expected at the model built with only the p_{raw} features.

(g) If the minimum of the out-of-sample s_e curve is at a p much larger than p_{raw}, this means the additional $p - p_{raw}$ features allowed the OLS algorithm to fit non-linearities and/or interactions among the features.

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

aceg

Problem 12 [11min] (and 100min will have elapsed) Consider the following four models in R:

```
> mod1 = lm(medv ~ log(tax) + poly(rm, 2) + poly(zn, 2) + poly(nox, 2), MASS::Boston)
> mod2 = lm(medv ~ ., MASS::Boston)
> mod3 = lm(medv ~ . * rm, MASS::Boston)
> mod4 = lm(medv ~ . * rm + poly(rm, 2) + poly(zn, 2) + poly(nox, 2), MASS::Boston)
```

Recall that `medv` is the response, a continuous metric that measures average price in \$1,000's; `rm` is a continuous metric that measures the average number of rooms in the houses and it ranges between 3.56 and 8.78; and `chas` is a dummy variable indicating whether the property is on the Charles River or not.

- [11 pt / 116 pts] Record the letter(s) of all the following that are **true** in general. At least one will be true.

(a) For model 1, you can say `medv` increases by b_1 times a proportion change in `tax`.

(b) Model 1 can fit a non-linear monotonic relationship in rm for the input space of rm.

(c) Model 1 can fit a non-linear monotonic relationship in rm for all values in $\mathbb{R}$.

(d) The model matrix for model 1 will include a column for the raw values of rm and a transformed column that takes these raw values and squares them element-wise.

(e) Model 2 has a higher R^2 than model 1.

(f) Model 3 has a higher R^2 than model 2.

(g) Model 4 has a higher R^2 than model 3.

(h) Model 2 has a higher oos R^2 than model 1.

(i) Model 3 has a higher oos R^2 than model 2.

(j) Model 4 has a higher oos R^2 than model 3.

(k) Within $\boldsymbol{b}$ in model 3, the 18th element is named `chas:rm` and has a value of -1.643. You can interpret this slope coefficient as follows: as the number of rooms increments (i.e. `rm` increases by 1), the response `medv` is predicted to decrease by \$1,643.

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

abfg