Math 342W / 642 / RM742 Spring 2025
Midterm Examination One

Professor Adam Kapelner

March 20, 2025

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 (variable time per question) and closed-book. You are allowed **two** pages (front and back) of a "cheat sheet", blank scrap paper (provided by the proctor) and a graphing calculator (which is not your smartphone). Please read the questions carefully. Within each problem, I recommend considering the questions that are easy first and then circling back to evaluate the harder ones. No food is allowed, only drinks.

Problem 1 This question relates to conceptual issues.

(a) [8 pt / 8 pts] Circle the letters of all the following that are **true**. The notation below should be assumed the same from class. c d f i j k m n

(a) A model g for a real-world, non-man-made phenomenon y which is a function of real-world measurements $x_1, \ldots, x_p$ could be absolutely true, (i.e. always have zero error when predicting in for future units).

(b) A model g for phenomenon y can be proved true analytically (using a mathematical deductive proof).

(c) A null model g_0 outputs a value in the set $\mathcal{Y}$.

(d) A null model g_0 can still be "useful" to the model's user.

(e) Any model g which is a function of real-world measurements $x_1, \ldots, x_p$ *always* beats g_0 on predictive performance.

(f) A model g for phenomenon y which is of type nominal categorical with possible values within the set $C_1, C_2, \ldots, C_L$ must return a value within the set $C_1, C_2, \ldots, C_L$.

(g) A model g for phenomenon y which is of type nominal categorical with possible values within the set $C_1, C_2, \ldots, C_L$ must have a proxy x_j which is also type nominal categorical with possible values within the set $C_1, C_2, \ldots, C_L$.

(h) If x_j is of type nominal categorical with possible values within the set $C_1, C_2, \ldots, C_L$, then it is reasonable when modeling that those categorical values should be replaced by $1, 2, \ldots, L$.

(i) If x_j is of type nominal categorical with possible values within the set $C_1, C_2, \ldots, C_L$, then it is reasonable when modeling that those categorical values should be replaced by $L-1$ binary variables indicating the presence of absence of categories $C_2, \ldots, C_L$.

(j) If x_j is of type ordinal with possible values within the set $C_1, C_2, \ldots, C_L$, assuming that set is ordered from smallest to largest, then it is reasonable when modeling that those categorical values *can* be replaced by $1, 2, \ldots, L$.

(k) If x_j is of type ordinal with possible values within the set $C_1, C_2, \ldots, C_L$, then it is reasonable when modeling that those categorical values *can* be replaced by $L-1$ binary variables indicating the presence of absence of categories $C_2, \ldots, C_L$.

(l) In-sample performance metrics gives you an honest idea about how accurate your model is when using it for future prediction *always*.

(m) In-sample performance metrics gives you an honest idea about how accurate your model is when using it for future prediction *only* when n is much larger than p.

(n) Out-of-sample validation gives you an idea about how accurate your model is when using it for future prediction *always*.

(o) Out-of-sample validation gives you an idea about how accurate your model is when using it for future prediction *only* when n is much larger than p.

(p) Out-of-sample validation gives you an idea about how accurate your model is when using it for future prediction *only* when $\mathcal{Y} \subseteq \mathbb{R}$.

Problem 2 Consider the following data frame $\mathbb{D}$. Below is some `R` code (lines beginning with > are commands that were entered, other lines are output from those commands).

```
> skimr::skim(D)
── Data Summary ────────────────────────
                           Values
Name                       D
Number of rows             592
Number of columns          4
Key                        NULL
────────────────────────────────────────
Column type frequency:
  character                2
  numeric                  2
────────────────────────────────────────
Group variables            None

── Variable type: factor ───────────────
  skim_variable n_missing complete_rate ordered n_unique top_counts
1 eye_color             0             1 FALSE          4 Bro: 220, Blu: 215,
   Haz: 93, Gre: 64
2 hair_color            0             1 FALSE          4 Bro: 286, Blo: 127,
   Bla: 108, Red: 71

── Variable type: numeric ──────────────
  skim_variable n_missing complete_rate  mean    sd p0 p25 p50 p75 p100
1 is_male               0             1 0.471 0.500  0   0   0   1    1
2 is_not_male           0             1 0.529 0.500  0   0   1   1    1

> D$is_not_male = 1 - D$is_male
> D[sample(1 : nrow(D), 10), ]
    is_male eye_color hair_color is_not_male
71        1     Brown      Brown           0
446       0     Hazel      Brown           1
474       0     Green      Brown           1
75        1     Brown      Brown           0
489       0     Brown        Red           1
107       1     Brown      Brown           0
473       0     Green      Brown           1
69        1     Brown      Brown           0
447       0     Hazel      Brown           1
137       1      Blue      Brown           0
```

(a) [1 pt / 9 pts] Circle one letter: the output on lines 28–37 is ...

A) the first 10 rows of data frame $\mathbb{D}$

B) a random 10 rows of data frame $\mathbb{D}$

C) the last 10 rows of data frame $\mathbb{D}$

(b) [1 pt / 10 pts] What type of variable is `eye_color`? Be as specific as you can.

nominal (or unordered) categorical (or factor)

Let y be defined as the variable `is_male` and all other variables be considered features.

(c) [2 pt / 12 pts] What is the g_0 function equal to explicitly?

$g_0 = 0$

(d) [3 pt / 15 pts] Circle all of the following modeling procedures that would be appropriate for this modeling setting considering the response space.

A) OLS

B) Perceptron

C) SVM

D) SVM with Vapnik objective function (and a numeric choice for hyperparameter λ)

E) KNN with the Euclidean distance function

F) KNN with a non-Euclidean, but still valid distance function

(e) [3 pt / 18 pts] Regardless of your answer in (c), how many columns will the full-rank design matrix X have? Hint: make sure you include the $\mathbf{1}_n$ as the first column in X.

8 = 1 for intercept + 3 for eye_color + 3 for hair_color + 1 for is_not_male

(f) [2 pt / 20 pts] Regardless of your answer in (c), consider using the SVM to fit g. What would the in-sample misclassification error for g be?

0

(g) [2 pt / 22 pts] Regardless of your answer in (c), consider using the OLS to fit g. What would the in-sample misclassification error for R^2 be?

1

Now we will change the modeling target. Let y be defined as the variable `eye_color` and all other variables be considered features.

(h) [2 pt / 24 pts] What is the g_0 function equal to explicitly?

g_0 = Brown

(i) [3 pt / 27 pts] Circle all of the following modeling procedures that would be appropriate for this modeling setting considering the response space.

A) OLS

B) Perceptron

C) SVM

D) SVM with Vapnik objective function (and a numeric choice for hyperparameter λ)

E) KNN with the Euclidean distance function

F) KNN with a non-Euclidean, but still valid distance function

(j) [3 pt / 30 pts] Regardless of your answer in (g), how many columns will the full-rank design matrix X have? Hint: make sure you include the $\mathbf{1}_n$ as the first column in X.

5 = 1 for intercept + 3 for hair_color + 1 for is_male (this is_not_male is dropped as it's a linear combination of the intercept and is_male)

Now we will change the modeling target again. Let y be defined as the variable `hair_color` and we only use the feature `eye_color`.

(k) [5 pt / 35 pts] Let the reference level of `eye_color` be Green. For the subset of the data shown on rows 28–37 of the code, provide the same 10-row subset of the full-rank design matrix X. Name the columns of X appropriately.

(intercept)	eye_color_is_brown	eye_color_is_blue	eye_color_is_hazel
1	1	0	0
1	0	0	1
1	0	0	0
1	1	0	0
1	1	0	0
1	1	0	0
1	0	0	0
1	1	0	0
1	0	0	1
1	0	1	0

(l) [2 pt / 37 pts] Assume you use KNN with the Euclidean distance function to fit g. Of the three sources of error, which is likely the largest source of error in g's predictions?

ignorance

Problem 3 Here are some theoretical questions. For all notation used, assume it has the same as it did in class.

(a) [4 pt / 41 pts] Assume $\mathcal{A}$ = OLS (linear regression). Prove that $\boldsymbol{b} = R^{-1}Q^\top \boldsymbol{y}$.

$$\begin{aligned}
\boldsymbol{b} &= (X^\top X)^{-1} X^\top \boldsymbol{y} \\
&= ((QR)^\top QR)^{-1} (QR)^\top \boldsymbol{y} \\
&= (R^\top Q^\top QR)^{-1} R^\top Q^\top \boldsymbol{y} \\
&= (R^\top R)^{-1} R^\top Q^\top \boldsymbol{y} \\
&= R^{-1} (R^\top)^{-1} R^\top Q^\top \boldsymbol{y} \\
&= R^{-1} Q^\top \boldsymbol{y}
\end{aligned}$$

(b) [10 pt / 51 pts] Assume $\mathcal{A}$ = OLS (linear regression). For each of the following, if the operation is illegal, write "illegal". If the operation is legal and the expression allows for simplification, simplify as best you can after an "=" sign. If you cannot simplify, then leave the line *blank*. If your answer is the zero or one vector, please indicate the dimension as a subscript to get full credit.

- XQ illegal
- XR
- $X\boldsymbol{b} = \hat{\boldsymbol{y}}$
- $HQ = Q$

- QH illegal
- $Q\boldsymbol{b}$
- $\boldsymbol{y}^\top H = \hat{\boldsymbol{y}}^\top$
- $\boldsymbol{e}^\top H = \mathbf{0}_n^\top$
- $QQ^\top \mathbf{1}_n = \mathbf{1}_n$
- $\text{rank}\,[H]\,/\text{rank}\,[X] = 1$

Problem 4 Consider the Galton dataset we examined in lab. Below is some `R` code (lines beginning with > are commands that were entered, other lines are output from those commands).

```
> skimr::skim(HistData::Galton)
── Data Summary ────────────────────────
                           Values
Name                       HistData::Galton
Number of rows             928
Number of columns          2
────────────────────────────
Column type frequency:
  numeric                  2

── Variable type: numeric ──────────────────────────────────────────────────────
  skim_variable n_missing complete_rate mean   sd   p0  p25  p50  p75 p100
1 parent                0             1 68.3 1.79 64   67.5 68.5 69.5 73
2 child                 0             1 68.1 2.52 61.7 66.2 68.2 70.2 73.7
```

(a) [1 pt / 52 pts] What is the sample size of this dataset?

928

As in the lab, consider the dependent variable `child` which measures the height of a child and consider the independent variable `parent` which measures the average height of the mother and father of the child.

(b) [1 pt / 53 pts] What is p in this modeling context?

1

(c) [3 pt / 56 pts] "The unit is one ___" (fill in the blank below by describing the unit).

father-mother-child trio

(d) [2 pt / 58 pts] Regardless of the $\mathcal{A}$ employed to create a model, which will likely be the greatest of the three sources of error?

ignorance (since there's only one feature and the phenomenon of interest is complicated)

(e) [2 pt / 60 pts] If $\mathcal{A}$ = OLS, what are the contents of the set $\mathcal{H}$?

$$\mathcal{H} = \{w_0 + w_1 x : w_0, w_1 \in \mathbb{R}\} = \left\{[1\ x]\boldsymbol{w} : \boldsymbol{w} \in \mathbb{R}^2\right\}$$

(f) [3 pt / 63 pts] If $\mathcal{A}$ = OLS, provide one specific $< x, y >$ point on the OLS line (i.e. the g model).

since $< \bar{x}, \bar{y} >$ is always on the OLS line, the only point you could know about without more output is $< 68.3, 68.1 >$

(g) [5 pt / 68 pts] Prove numerically that there is "regression to the mean" for the phenomenon of interest in this modeling setting. Assume that the correlation coefficient between the two variables is 0.46.

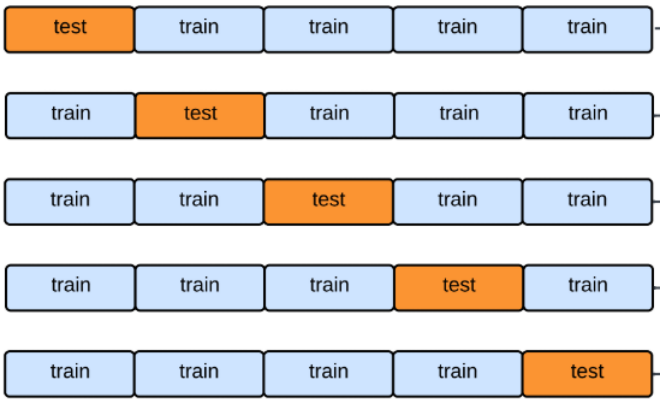
Regression to the mean means that we predict a child to have a smaller height than their parents for parents with above-average height and we predict a child to have a larger height than their parents for parents with below-average height. This is equivalent to showing $b_1 < 1$. We now calculate b_1 and show it is less than 1.

$$b_1 = r\frac{s_y}{s_x} = 0.46 \cdot \frac{2.52}{1.79} = 0.65 < 1$$

(h) [5 pt / 73 pts] If $\mathcal{A} = \text{OLS}$, calculate the RMSE to two decimals.

$$\begin{aligned} SST &= (n-1)s_y^2 = 927 \cdot 2.52^2 = 5886.821 \\ R^2 &= 1 - SSE/SST \quad \Rightarrow \quad SSE = SST(1-R^2) = 5886.821(1-.46^2) = 4650.588 \\ RMSE &= \sqrt{\frac{SSE}{n-(p+1)}} = \sqrt{\frac{4650.588}{926}} = 2.24 \end{aligned}$$

Problem 5 Consider the following illustration of cross-validation for $\mathbb{D}$ with $n = 200$.



(a) [1 pt / 74 pts] What is the value of K?

5

(b) [1 pt / 75 pts] How many units will be in $\mathbb{D}_{train}$ in the first split?

$n - \frac{1}{K}n = 160$

(c) [1 pt / 76 pts] If you calculate oos residuals in the second split, how many oos residuals will there be?

It's the same as the size of $\mathbb{D}_{test}$ which is 40.

(d) [2 pt / 78 pts] If g was overfit on $\mathbb{D}_{train}$, then how would oosRMSE be expected to compare with in-sample RMSE?

oosRMSE > in-sample RMSE

(e) [3 pt / 81 pts] If you calculate the oosRMSE from the residuals in (c), this oosRMSE is expected to be greater than, less than or equal to the oosRMSE you expect in the future when predicting on g_{final} which is fit with all $\mathbb{D}$?

greater than

(f) [3 pt / 84 pts] If you aggregate all the residuals from all K folds together and calculate the oosRMSE, this oosRMSE is expected to be greater than, less than or equal to the oosRMSE you expect in the future when predicting on g_{final} which is fit with all $\mathbb{D}$?

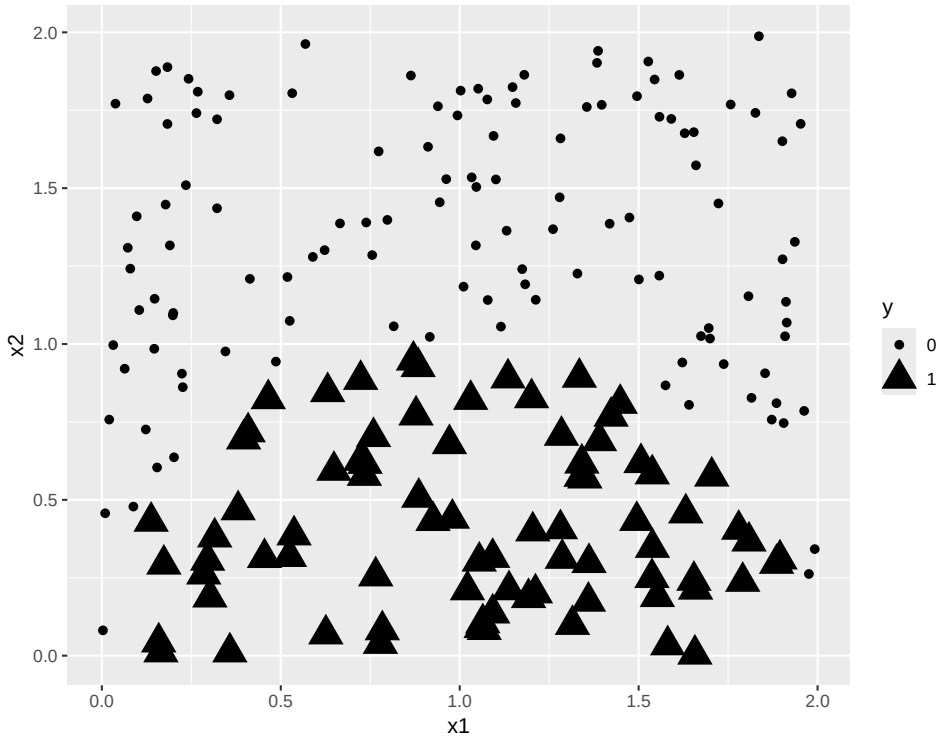
greater than

Problem 6 Consider the following $\mathbb{D}$ illustrated within the questions below.

(a) [2 pt / 86 pts] Examine the plot of $\mathbb{D}$ in the following question. If y is the response variable and the other two variables are considered features, what is the formal name of the type of modeling scenario most likely depicted?

binary classification

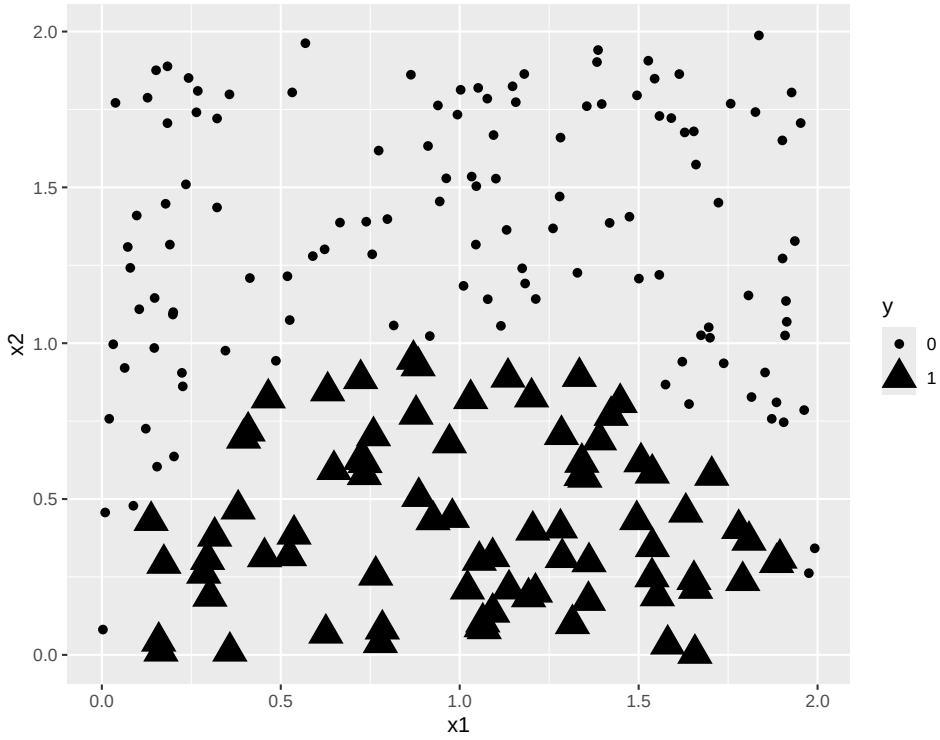
(b) [2 pt / 88 pts] If $\mathcal{A}$ = perceptron with the $\mathcal{H}$ we considered thus far in class, draw a possible g that outputs from this algorithm after 10,000 iterations on the illustration below:



any straight line is acceptable

(c) [1 pt / 89 pts] If $\mathcal{A}$ = perceptron with the $\mathcal{H}$ we considered thus far in class, will the algorithm eventually converge? Circle one: yes / no

(d) [4 pt / 93 pts] If $\mathcal{A}$ = SVM with the $\mathcal{H}$ we considered thus far in class and the Vapnik function with hinge errors and $\lambda = 0.001$, draw the g that most likely outputs from this algorithm on the illustration below:



With the hyperparameter set to such a low value, we're really just minimizing sum of hinge errors which would likely be the horizontal line at $x_2 \approx 0.75$.

(e) [3 pt / 96 pts] If $\mathcal{A}$ = SVM with the $\mathcal{H}$ we considered thus far in class and the Vapnik function with hinge errors and $\lambda = 0.001$, what is the most likely of the three sources of error in g?

Misspecification since we can clearly see a semicircle threshold with diameter running across $x_2 = 0$ can obtain near zero misclassification error.

(f) [2 pt / 98 pts] If $\mathcal{A}$ = KNN with the Euclidean distance function and $K = 1$, what would be $g(0, 0)$?

The closest point to $< 0, 0 >$ has $y = 0$ so thus $\hat{\boldsymbol{y}} = 0$.

(g) [2 pt / 100 pts] If $\mathcal{A}$ = KNN with the Euclidean distance function and $K = 10$, what would be $g(0, 0)$?

Draw a circle arc from the y-axis to the x-axis with $< 0, 0 >$ at its center such that 10 points are included. It is clear that there are more triangles than dots so that the modal value $\hat{\boldsymbol{y}} = 1$.