Math 342W / 650.4 Spring 2022
Midterm Examination One

Solutions

Professor Adam Kapelner

Wednesday, March 23

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 is 100 points total plus extra credit. Partial credit will be granted for incomplete answers on most of the questions. $\boxed{\text{Box}}$ in your final answers. Good luck!

Problem 1 This question is about science and modeling in general.

- [4 pt / 4 pts] When an object free falls to the ground from height h, an elementary physics provides textbook provides the formula for the predicted time t the object takes to reach the ground as $t = \sqrt{2h/g}$ where g is a constant. Explain why this formula is "wrong but useful".

  This formula is a model, an approximation to reality which is not absolute truth (and hence "wrong"). It ignores features such as air drag and wind. Even in a vacuum, it would not be 100% accurate as there are dependencies on the mass of the object and relativistic effects. However, it is "useful" because it is likely accurate enough to solve the practical problem you are facing.

- [9 pt / 13 pts] Circle the letters of all the following that are **true**.

  (a) A "phenomenon" is anything one finds interesting in the world

  (b) The enterprise of the scientific endeavor is essentially modeling

  (c) The two goals of modeling is to provide predictions of the phenomenon in future settings and explanation of how the settings affect the phenomenon

  (d) Two different people can come to two different predictions for the same observation using a non-mathematical model

  (e) Given one mathematical model g, there can be two different y values for equal $\boldsymbol{x}$ input vectors

  (f) Given one mathematical model g, there can be two different $\hat{y}$ values for equal $\boldsymbol{x}$ input vectors

  (g) The naive model g_0 requires historical data $\boxed{\text{(g) can be argued false}}$

  (h) The naive model g_0 can be used for prediction

  (i) The naive model g_0 cannot be validated since it does not make use of the $\boldsymbol{x}_{i\cdot}$'s

- [6 pt / 19 pts] Circle the letters of all the following that are **true**. In the quote by George Box and Norman Draper in 1987, "All models are wrong but some are useful" means that models ...

  (a) ... must have univariate response

  (b) ... must be constructed using supervised learning

  (c) ... sometimes provide accuracy that meets your prediction goals

  (d) ... never can achieve perfect predictive accuracy

  (e) ... need perfectly accurate input measurements

  (f) ... never describe the pheonomenon absolutely

Problem 2 Consider the diamonds dataset which is part of the `ggplot2` package in `R`. This is a dataset we will be looking at extensively later in the course.

```
> D = ggplot2::diamonds
> dim(D)
[1] 53940    10
> summary(D)
     carat               cut         color      clarity
 Min.   :0.2000   Fair     : 1610   D: 6775   SI1    :13065
 1st Qu.:0.4000   Good     : 4906   E: 9797   VS2    :12258
 Median :0.7000   Very Good:12082   F: 9542   SI2    : 9194
 Mean   :0.7979   Premium  :13791   G:11292   VS1    : 8171
 3rd Qu.:1.0400   Ideal    :21551   H: 8304   VVS2   : 5066
 Max.   :5.0100                     I: 5422   VVS1   : 3655
                                    J: 2808   (Other): 2531
     depth           table           price             x
 Min.   :43.00   Min.   :43.00   Min.   :  326   Min.   : 0.000
 1st Qu.:61.00   1st Qu.:56.00   1st Qu.:  950   1st Qu.: 4.710
 Median :61.80   Median :57.00   Median : 2401   Median : 5.700
 Mean   :61.75   Mean   :57.46   Mean   : 3933   Mean   : 5.731
 3rd Qu.:62.50   3rd Qu.:59.00   3rd Qu.: 5324   3rd Qu.: 6.540
 Max.   :79.00   Max.   :95.00   Max.   :18823   Max.   :10.740

       y                z
 Min.   : 0.000   Min.   : 0.000
 1st Qu.: 4.720   1st Qu.: 2.910
 Median : 5.710   Median : 3.530
 Mean   : 5.735   Mean   : 3.539
 3rd Qu.: 6.540   3rd Qu.: 4.040
 Max.   :58.900   Max.   :31.800
```

1/4 6 6

6 + 4+6+6

6 + 1+6+6

- [1 pt / 20 pts] Using the terminology used in class, what data type is `carat`?

  continuous

- [1 pt / 21 pts] Using the terminology used in class, what data type is `cut`?

  ordinal (categorical)

- [1 pt / 22 pts] Using the terminology used in class, what data type is `color`?

  nominal (categorical)

If we were to model the response `price` using the OLS algorithm ...

- [2 pt / 24 pts] ... then $g_0 =$ 3933

- [1 pt / 25 pts] ... with all other columns as regressors, what is the value of n?

53,940

- [2 pt / 27 pts] ... with only `color` as the sole regressor where its levels are dummified, what is the value of p?

$\boxed{6}$

- [2 pt / 29 pts] ... with only `color` as the sole regressor where its levels are dummified, which of the three types of modeling errors is most likely largest?

ignorance

- [2 pt / 31 pts] ... with only `color` as the sole regressor where its levels are dummified, which of the three types of modeling errors is most likely smallest?

estimation

- [2 pt / 33 pts] ... with only `color` as the sole regressor where its levels are dummified, explain in English how you can calculate $\hat{y}$ if $x = G$.

Locate the price values for the diamonds whose color is G and take the average i.e.

$$\hat{y} = \frac{1}{\sum_{i=1}^{n} \mathbb{1}_{x_i = G}} \sum_{i=1}^{n} y_i \mathbb{1}_{x_i = G}$$

- [4 pt / 37 pts] ... with all other columns as regressors, what is the value of p? Hint: there may be multiple acceptable answers.

$p = 19$ if cut is coded numerically
$p = 22$ if cut is dummified

If we were to model the response `clarity` ...

- [1 pt / 38 pts] ... then the model would be a classification model.
- [2 pt / 40 pts] ... then $g_0 =$ SI1
- [1 pt / 41 pts] ... then would the OLS algorithm be suitable?
Circle one: Yes / no

- [1 pt / 42 pts] ... then would the perceptron algorithm be suitable?
Circle one: Yes / no

- [3 pt / 45 pts] ... using the KNN algorithm on price x, provide a legal distance function below for a new input x_*.

$$d(x, x_*) = (x - x_*)^2 \quad or \quad d(x, x_*) = |x - x_*|$$

If we were to model a response cut_is_ideal defined as $y_i := \mathbb{1}_{\texttt{cut}_i\ =\ \texttt{Ideal}}$ ...

- [2 pt / 47 pts] ... then $g_0 =$ 0

The remaining questions require Figure 1, a scatterplot of $y =$ cut_is_ideal on $x_1 =$ table and $x_2 =$ depth.

- [7 pt / 54 pts] Circle the letters of all the following that are **true**.

(a) This dataset is linearly separable

(b) There is an association between $y =$ cut_is_ideal and $x_1 =$ table

(c) There is an association between $y =$ cut_is_ideal and $x_2 =$ depth

(d) There is a large r_{x_1,x_2} i.e. near -1 or +1

(e) Using $\mathcal{A} =$ perceptron to model y with maximum iterations 1,000,000 will return a valid g in exactly 1,000,000 iterations

(f) KNN with the default $K = \sqrt{n}$ will most likely outperform both the perceptron and SVM regardless of the λ hyperparmeter setting in the Vapnik function

(g) There is likely a model g learned from this dataset that can attain zero errors oos

- [1 pt / 55 pts] Using the KNN algorithm to model y based on these two inputs and we use the default $K = \sqrt{n}$, then it seems most likely that the prediction for table = 65 and depth = 65 is ... 0

- [1 pt / 56 pts] Using the KNN algorithm to model y based on these two inputs and we use the default $K = \sqrt{n}$, then it seems most likely that the prediction for table = 55 and depth = 63 is ... 1

- [2 pt / 58 pts] If we were to use the SVM with $\mathcal{H} = \{\mathbb{1}_{\boldsymbol{w}\cdot\boldsymbol{x}+b\geq 0} : \boldsymbol{w} \in \mathbb{R}^2, b \in \mathbb{R}\}$ with a reasonable value of λ, then of the three types of modeling errors, the type most pronounced will likely be ...

misspecification

Problem 3 Let $\boldsymbol{X} = [\boldsymbol{1}_n \mid \boldsymbol{x}_1 \mid \ldots \mid \boldsymbol{x}_p] \in \mathbb{R}^{n \times (p+1)}$ a non-orthogonal matrix whose entries after the first column are iid standard random normals, $\text{rank}[\boldsymbol{X}] = p+1 < n$, $\boldsymbol{y} \in \mathbb{R}^n$ whose average is $\bar{y}$ and sample variance is s_y^2. The modeling task is to model the response using the n observations. Let $\boldsymbol{b}$ be the coefficients for the $p+1$ features, generated via the following $\mathcal{A}$,

$$\boldsymbol{b} = \underset{\boldsymbol{w} \in \mathbb{R}^{p+1}}{\arg\min} \left\{ (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \right\},$$

let $\boldsymbol{\beta}$ be the slope coefficients in the model that optimally fits $f(\boldsymbol{x})$, $\boldsymbol{H}$ be the orthogonal projection matrix onto the colsp $[\boldsymbol{X}]$, $\boldsymbol{Q}$ be the result of running Gram-Schmidt algorithm on $\boldsymbol{X}$, $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$, $\hat{\boldsymbol{y}}$ is the vector of predictions for the n observations, $\boldsymbol{e}$ are the residuals where at least one $e_i \neq 0$, $\boldsymbol{X}_\perp$ denotes matrix whose columns form the span for $\mathbb{R}^n$ that are not included in the columns of $\boldsymbol{X}$ and $\boldsymbol{H}_\perp$ be the orthogonal projection matrix onto the colsp $[\boldsymbol{X}_\perp]$.

- [26 pt / 84 pts] Circle the letters of all the following that are **true**.

  (a) This algorithm is OLS

  (b) $\boldsymbol{b} = \underset{\boldsymbol{w} \in \mathbb{R}^{p+1}}{\arg\min} \left\{ \sum_{i=1}^{n} (y_i - \boldsymbol{x}_{i\cdot} \boldsymbol{w})^2 \right\}$.

  (c) SSR < SST

  (d) As p increases, the dimension of $\boldsymbol{H}$ increases

  (e) As p increases, the rank of $\boldsymbol{H}$ increases

  (f) $\text{rank}[\boldsymbol{H}] = n$ if $\boldsymbol{X}\boldsymbol{b} = \boldsymbol{y}$

  (g) $\boldsymbol{H}\boldsymbol{y} = \boldsymbol{y}$

  (h) $\boldsymbol{H}\hat{\boldsymbol{y}} = \hat{\boldsymbol{y}}$

  (i) $\boldsymbol{H}_\perp \boldsymbol{y} = \boldsymbol{y}$

  (j) $\boldsymbol{H}_\perp \boldsymbol{y} = \boldsymbol{e}$

  (k) $\boldsymbol{H}_\perp \boldsymbol{e} = \boldsymbol{e}$

  (l) $[\boldsymbol{X} \vdots \boldsymbol{X}_\perp] = \boldsymbol{I}_n$

  (m) $\boldsymbol{H} + \boldsymbol{H}_\perp = \boldsymbol{I}_{p+1}$

  (n) $\boldsymbol{H} + \boldsymbol{H}_\perp = \boldsymbol{I}_n$

  (o) $\boldsymbol{y} \cdot \boldsymbol{e} = 0$

  (p) $\boldsymbol{b} \cdot \boldsymbol{e} = 0$

  (q) $\boldsymbol{h}^* = \boldsymbol{X}\boldsymbol{\beta}$ where $\boldsymbol{h}^*$ is the n-dimensional column vector of all the $h^*(\boldsymbol{x}_{i\cdot})$'s

  (r) $\boldsymbol{h}^* \cdot \boldsymbol{e} = 0$

  (s) $\boldsymbol{X} \left(\boldsymbol{X}^T \boldsymbol{X} \right)^{-1} \boldsymbol{X}^T \boldsymbol{q}_{\cdot 3} = \boldsymbol{0}_n$

(t) $\boldsymbol{X}\boldsymbol{b} = \hat{\boldsymbol{y}}$

(u) $\boldsymbol{Q}\boldsymbol{b} = \hat{\boldsymbol{y}}$

(v) $\boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{X} = \boldsymbol{X}$

(w) $\boldsymbol{I}_n - \boldsymbol{Q}\boldsymbol{Q}^\top = \boldsymbol{H}_\perp$

(x) An analysis of the entries in $\boldsymbol{H}_\perp$ can inform us if g is overfit

(y) Gram-Schmidt will produce the same $\boldsymbol{Q}$ if it is run on $\boldsymbol{X}'$ whose columns are the same as $\boldsymbol{X}$ except in a different order

(z) colsp $[\boldsymbol{X}\boldsymbol{R}]$ = colsp $[\boldsymbol{Q}]$

- [7 pt / 91 pts] Prove $\sum_{i=1}^n \hat{y}_i = n\bar{y}$ for all p.

$$\sum \hat{y}_i = \vec{1}_n^\top \vec{\hat{y}} = \vec{1}_n^\top (H\vec{y}) = \vec{1}_n^\top H^\top \vec{y} = (H\vec{1}_n)^\top \vec{y} = \vec{1}_n^\top \vec{y} = \sum y_i = n\bar{y}$$

- [7 pt / 98 pts] On an axis below, plot the in-sample RMSE for this algorithm as a function of p using a line or points. Label the axes and label all critical points using the notation provided in the problem header.

Since $\vec{x}_{\cdot j}$'s are random values, there is an R^2 contribution of $\frac{1}{n-1}$ per feature. Hence $R^2 = \frac{1}{n-1} \cdot p$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{SSE}{n-p-1}}$$

$$R^2 = 1 - \frac{SSE}{SST} \Rightarrow SST R^2 = SST - SSE$$

$$\Rightarrow SSE = SST(1 - R^2)$$

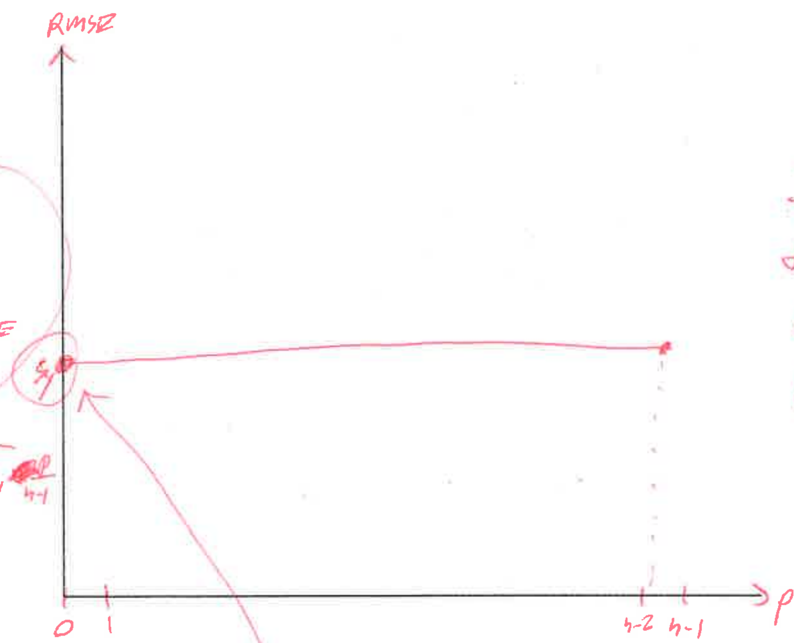
$$RMSE = \sqrt{\frac{SST(1-R^2)}{n-p-1}} = \sqrt{\frac{SST}{n-p-1} - \frac{SST}{n-p-1}\frac{p}{n-1}}$$

$$= \sqrt{SST}\sqrt{\frac{1}{n-p-1} - \frac{p}{(n-p-1)(n-1)}}$$

$$= \sqrt{SST}\sqrt{\frac{n-1}{(n-p-1)(n-1)} - \frac{p}{(n-p-1)(n-1)}}$$

$$= \sqrt{SST}\sqrt{\frac{\cancel{n-p-1}}{\cancel{(n-p-1)}(n-1)}} \quad \text{if } n-p-1 \neq 0$$

$$= \sqrt{\frac{SST}{n-1}} = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}} = s_y$$



In practice, this function will wobble up and down but remain flat

Note: RMSE is undefined for $p = n-1$

Problem 4 Assume $\boldsymbol{X}$ and $\boldsymbol{y}$ have the same values as in the previous problem but now the coefficients are generated via a new algorithm $\mathcal{A}_{new}$,

$$\boldsymbol{b}_{new} = \underset{\boldsymbol{w} \in \mathbb{R}^{p+1}}{\arg\min} \left\{ \sum_{i=1}^{n} (y_i - \boldsymbol{x}_{i \cdot} \boldsymbol{w})^4 \right\},$$

which produces new predictions $\hat{\boldsymbol{y}}_{new}$ and new residuals $\boldsymbol{e}_{new}$.

- [8 pt / 106 pts] Circle the letters of all the following that are **true**.

  (a) This algorithm is OLS

  (b) $\boldsymbol{X}\boldsymbol{b}_{new} = \hat{\boldsymbol{y}}_{new}$

  (c) $\boldsymbol{b}_{new} = \boldsymbol{b}$

  (d) $\hat{\boldsymbol{y}}_{new} = \hat{\boldsymbol{y}}$

  (e) $||\boldsymbol{y}||^2 = ||\hat{\boldsymbol{y}}_{new}||^2 + ||\boldsymbol{e}_{new}||^2$

  (f) $\hat{\boldsymbol{y}}_{new} \in \text{colsp}\left[\boldsymbol{X}\right]$

  (g) $\hat{\boldsymbol{y}}_{new} \in \text{colsp}\left[\boldsymbol{Q}\right]$

  (h) $\boldsymbol{e}_{new} \in \text{colsp}\left[\boldsymbol{X}_{\perp}\right]$

Problem 5 Assume a dataset $\mathbb{D} := \langle \boldsymbol{X}, \boldsymbol{y} \rangle$ where X is an $n \times p$ matrix and $\boldsymbol{y}$ is an $n \times 1$ column vector. The dataset is split into a train and test set of n_{train} observations and n_{test} observations. Let $\mathbb{D}_{\text{train}} := \langle \boldsymbol{X}_{\text{train}}, \boldsymbol{y}_{\text{train}} \rangle$ and $\mathbb{D}_{\text{test}} := \langle \boldsymbol{X}_{\text{test}}, \boldsymbol{y}_{\text{test}} \rangle$ just like we did in class and lab by taking a random partition of the indices $1, 2, \ldots, n$. Let $g_{\text{train}} = \mathcal{A}(\mathbb{D}_{\text{train}}, \mathcal{H})$, $g_{\text{test}} = \mathcal{A}(\mathbb{D}_{\text{test}}, \mathcal{H})$ and $g_{\text{final}} = \mathcal{A}(\mathbb{D}, \mathcal{H})$. We will assume stationarity of the phenomenon of interest as it related to the covariates in $\boldsymbol{X}$.

- [15 pt / 121 pts] Record the letters of all the following that are **true**. Your answer will consist of a string (e.g. `aebgd`) where the order of the letters does not matter.

  (a) If stationarity is not assumed, then supervised learning models cannot be validated without collecting data in addition to what was provided in $\mathbb{D}$

  (b) Validation in-sample is always dishonest

  (c) If $\mathbb{D}_{\text{train}}$ and $\mathbb{D}_{\text{test}}$ were generated from a different random partition of the indicies $1, 2, \ldots, n$, then the oos validation metrics are expected to be the same as the first random partition

  (d) $n_{\text{train}} + n_{\text{test}} = n$

  (e) If $K = 2$, then $\dim\left[\boldsymbol{y}_{\text{train}}\right] = \dim\left[\boldsymbol{y}_{\text{test}}\right]$

  (f) If $K = n$, then $\dim\left[\boldsymbol{y}_{\text{train}}\right] = \dim\left[\boldsymbol{y}_{\text{test}}\right]$

  (g) RMSE is calculated by using predictions from g_{train} and comparing them to $\boldsymbol{y}_{\text{train}}$

  (h) oosRMSE can be calculated by using predictions from g_{test} and comparing them to $\boldsymbol{y}_{\text{test}}$

(i)-(n) should have the phrase "in expectation" after the words "same" and "higher". So (i) can be argued false but there is less of an argument for (n).

(i) If $K > 2$, then oosRMSE will likely be the same as the RMSE of g_{train} when used to predict on future observations

(j) If $K > 2$, then oosRMSE will likely be higher than the RMSE of g_{train} when used to predict on future observations

(k) If $K > 2$, then oosRMSE will likely be the same as the RMSE of g_{test} when used to predict on future observations

(l) If $K > 2$, then oosRMSE will likely be higher than the RMSE of g_{test} when used to predict on future observations

(m) If $K > 2$, then oosRMSE will likely be the same as the RMSE of g_{final} when used to predict on future observations

(n) If $K > 2$, then oosRMSE will likely be higher than the RMSE of g_{final} when used to predict on future observations

(o) The larger K becomes, the less trustworthy oos performance statistics become