Math 390.4 / 650.3 Spring 2018 Midterm Examination One

Solutions

Professor Adam Kapelner

Monday, March 5, 2018

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 **one** page (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.

(a) [4 pt / 4 pts] Explain as best as you can why "all models are wrong but some are useful", a statement made by the famous statisticians George Box and Norman Draper.

Models are "wrong" because they are not reality. Models are "useful" in that they predict reality and those predictions can be accurate. They are also "useful" because their inner workings give clues to how reality operates.

(b) [1 pt / 5 pts] Consider the famous model proposed by Newton in his *Principia Mathematica* in 1687: "Lex II: Mutationem motus proportionalem esse vi motrici impressae, et fieri secundum lineam rectam qua vis illa imprimitur" better known as the "second law of motion" and when translated, is commonly rendered as $a = F/m$. This means that force (denoted F) on an object can accelerate (denoted a) the object but that the object's mass (denoted m) retards this acceleration. Note that all three quantities (F, a, m) can be measured. Is this a *mathematical model*? Yes/no.

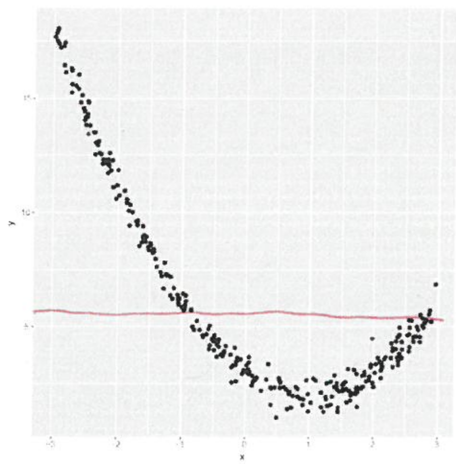
(c) [5 pt / 10 pts] Discuss how Newton's second law of motion can be *validated*.

Since all 3 quantities can be measured according to (b), we can set up an experiment to validate Newton's theory. In the experiment, each observation / case will consist of a different $\langle F_i, m_i \rangle$ pair. For each pair, we measure the response, a_i. We then compare these values to the values computed via Newton's theory i.e. $F_i/m_i = \hat{a}_i$. If the $\hat{a}_i$'s are "close to" the actual a_i's, the model is said to be valid. If not, it is "invalid." We would agree on a definition of "close to" before the experiment.

(d) [2 pt / 12 pts] There is an ancient religion who explains the phenomenon of sunsets as follows: a large but invisible dragon eats the sun. According to Karl Popper, is this model "scientific"? Why or why not?

No. The statement cannot be "falsified" since there is no way to measure the workings of an "invisible dragon".

Problem 2 This question is mostly about the framework of modeling. Consider the phenomenon y with one predictor x. In this case $x \in \mathcal{X} = [-3, 3]$ and $y \in \mathcal{Y} = [4, 16]$. Below is a plot of the data $\mathbb{D}$:



(a) [2 pt / 14 pts] If we are now in the statistical learning framework, what subtype of problem are we most likely solving?

i) regression to predict y

ii) binary classification to predict y

iii) finding t directly

iv) finding optimal n and p for $\mathbb{D}$

v) building $\mathcal{A}$ to find f directly

vi) estimating $\mathcal{X}$ and $\mathcal{Y}$ using $\mathcal{H}$

(b) [3 pt / 17 pts] Illustrate the null model g_0 as a function of x on the plot above.

(c) [2 pt / 19 pts] If we are now in the statistical learning framework, what will the final output be of the *learning procedure*?

i) $\hat{y}$

ii) $\mathcal{A}$

iii) g

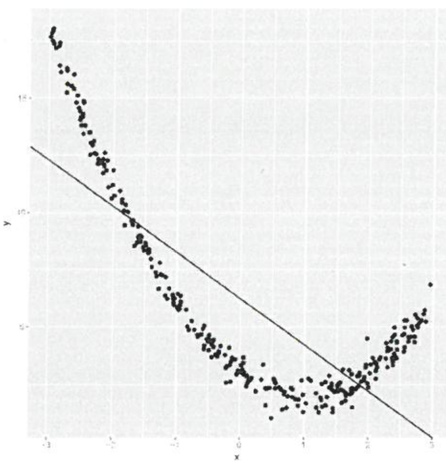
iv) h^*

v) h

vi) f

vii) $z_1, \ldots, z_t$

We will be fitting many models to this data. Consider fit #1 below (the line):



(d) [2 pt / 21 pts] What is the approximate R^2 of fit #1?

i) -50%
ii) 0%
iii) 5%
iv) 50%
v) 95%
vi) 100%

(e) [2 pt / 23 pts] If, instead of using x to model y, we used $\tilde{x} := \mathbb{1}_{x \geq 0}$ to model y and use the same $\mathcal{H}$ (and $\mathcal{A}$) as done in fit #1, the R^2 relative to your answer in (d) would

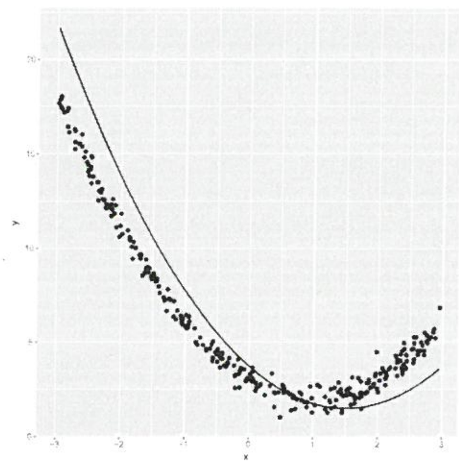
i) decrease
ii) remain the same
iii) increase
iv) not enough information to tell

(f) [2 pt / 25 pts] In fit #1, what is most likely the problem?

i) misspecification of $\mathcal{H}$
ii) g is too far from t
iii) we are too ignorant of $z_1, \ldots, z_t$ since we only know x
iv) the $\mathcal{A}$ is not optimizing its cost function correctly
v) f could never be approximated with this $\mathbb{D}$.
vi) $h^* \notin \mathcal{H}$

(g) [1 pt / 26 pts] In fit #1, is $g \approx h^*$? Yes/no.

Consider fit #2 below:



(h) [2 pt / 28 pts] In fit #2, what is most likely the problem?

i) misspecification of $\mathcal{H}$

ii) g is too far from t

iii) we are too ignorant of $z_1, \ldots, z_t$ since we only know x

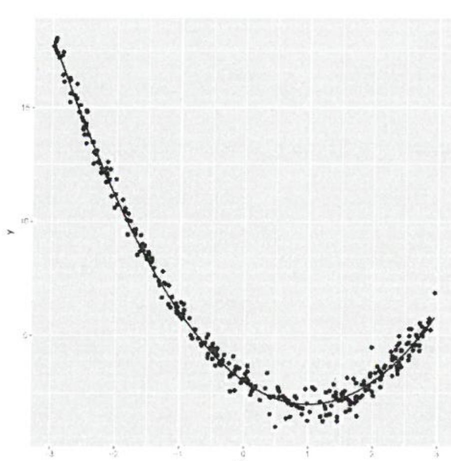
iv) the $\mathcal{A}$ is not optimizing its cost function correctly

v) f could never be approximated with this $\mathbb{D}$.

vi) $h^* \notin \mathcal{H}$

vii) non-linear models are illegal and thus this question cannot be answered

Consider fit #3 below:



(i) [2 pt / 30 pts] What is the approximate R^2 of fit #3?

i) -50%
ii) 0%
iii) 5%
iv) 50%
v) 95%
vi) 100%
vii) R^2 is meaningless if $\mathcal{H} \neq \{\boldsymbol{w} \cdot \boldsymbol{x} : \boldsymbol{w} \in \mathbb{R}^{p+1}\}$.

(j) [2 pt / 32 pts] What is the approximate RMSE of fit #3?

i) -4
ii) 0
iii) 0.4
iv) 4
v) 40
vi) RMSE is always $\bar{y}$
vii) RMSE is meaningless if $\mathcal{H} \neq \{\boldsymbol{w} \cdot \boldsymbol{x} : \boldsymbol{w} \in \mathbb{R}^{p+1}\}$.

(k) [3 pt / 35 pts] In fit #3, choose the *likely* largest source of error and indicate a strategy to mitigate it in the future.

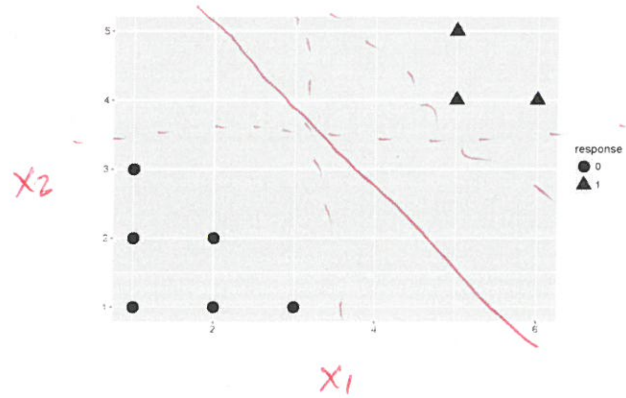
i) error due to ignorance which can be reduced by measuring more relevant predictors
ii) misspecification error which can be reduced by ______________________
iii) estimation error which can be reduced by ______________________

(l) [1 pt / 36 pts] Regardless of what you think about the fit of the model, consider the following situation. The phenomenon y that we were attempting to predit is *price of a metal per gram in dollars*. And, if the predicted price has to be wrong by more than \$0.25/g for our business to lose money, would this be a "good model"? Yes / no.

(m) [1 pt / 37 pts] Regardless of what you think about the fit of the model, consider the following situation. The phenomenon y that we were attempting to predit is *toxicity of a substance*. And, if the predicted toxicity is wrong by even 0.1 then the patient could die, would this be a "good model"? Yes / no.

(n) [1 pt / 38 pts] Could fits #1, 2 and 3 be fit with the support vector machine algorithm we discussed in class? Yes/no.

Problem 3 This question is another modeling example. Below is a plot of $\mathbb{D}$:



(a) [2 pt / 40 pts] If we are now in the supervised statistical learning framework, what subtype of problem are we most likely solving?

i) regression to predict y

ii) binary classification to predict y

iii) finding t directly

iv) finding optimal n and p for $\mathbb{D}$

v) building $\mathcal{A}$ to find f directly

vi) estimating $\mathcal{X}$ and $\mathcal{Y}$ using $\mathcal{H}$

(b) [2 pt / 42 pts] What is p in this supervised learning problem? What is n? Answer both numerically.

$p = 2,\ n = 9$

(c) [2 pt / 44 pts] What is the null model g_0 in this case? $g_0 = 0$

(d) [3 pt / 47 pts] If you were to use the *perceptron learning algorithm* beginning from random $\boldsymbol{w}$ locations, draw 3 possible outputs from the algorithm on the plot above. Use dashed lines to illustrate. Also, label the axes.

(e) [4 pt / 51 pts] For any of the 3 possible outputs from the *perceptron learning algorithm*, provide the function g below explicitly.

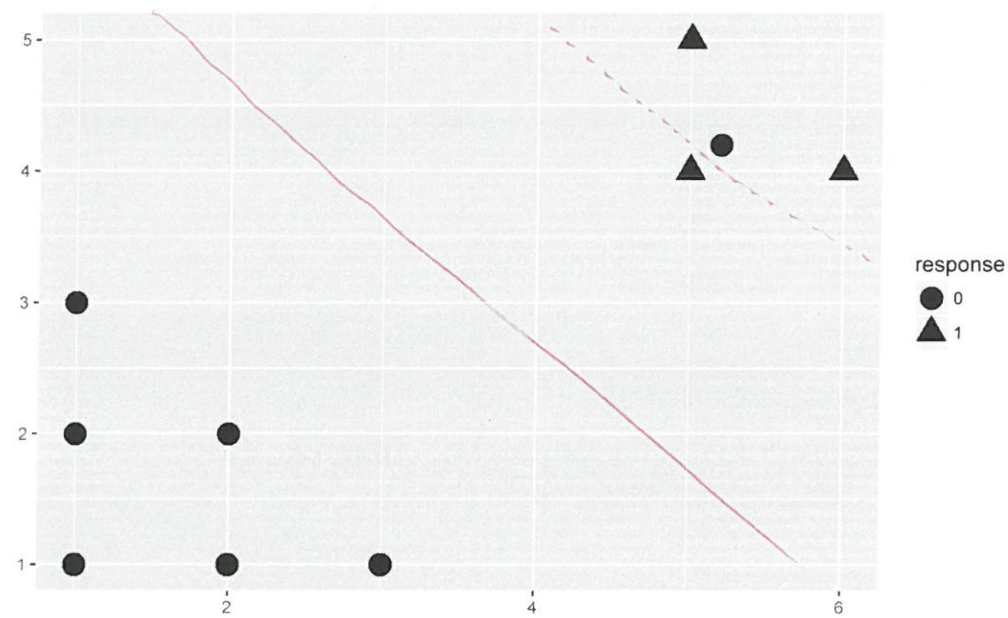
$g(x_1, x_2) = \mathbb{1}_{x_2 \geq 3.5}$

(f) [2 pt / 53 pts] If you were to use the *linear support vector machine algorithm* for linearly separable $\mathbb{D}$, draw one possible output from the algorithm on the plot above. Use a solid line to illustrate.

(g) [1 pt / 54 pts] Of the four models you imagined in parts (c) and (e), which one will give the best performance prediction by and large under general conditions?

The svm in (f).

Now consider the same data except with one new data point.



(h) [3 pt / 57 pts] Consider employing the *linear support vector machine algorithm* for nonlinearly separable $\mathbb{D}$. Let $\lambda = 0$. Draw the most likely output line on the above plot. Use a dashed line to illustrate. Also, label the axes the same as previously.

(i) [3 pt / 60 pts] Approximate the average hinge loss of the dashed line you drew in the previous question.

i) < 0

ii) 0

iii) 0.02

iv) 0.2

v) 2

vi) >2

(j) [3 pt / 63 pts] Consider employing the *linear support vector machine algorithm* for nonlinearly separable $\mathbb{D}$. And let $\lambda > 0$ so that the term that λ effects becomes important in the optimization but does not drown out the hinge loss. Draw the most likely output line on the above plot. Use a solid line to illustrate.

(k) [1 pt / 64 pts] How would the solution you proposed in (i) be fit on a computer?

I) There is an analytic solution that is pre-programmed

II) The computer uses numerical optimization which is heuristic

(l) [2 pt / 66 pts] Consider $\boldsymbol{x}^* = [5.4 \ 4.2]$. What is $\hat{y}^*$ if you employ the KNN algorithm where $k = 1$?

0

(m) [3 pt / 69 pts] Consider $\boldsymbol{x}^* = [5.4 \ 4.2]$. What is $\hat{y}^*$ if you employ the KNN algorithm where $k = 3$?

1

(n) [3 pt / 72 pts] Consider $\boldsymbol{x}^* = [5.4 \ 4.2]$. What is $\hat{y}^*$ if you employ the KNN algorithm where $k = 10$?

0

Problem 4 Consider a subset of the Boston Housing Data that has been preprocessed. Below is some R code that gives background on the `boston` data frame which will be referenced throughout this problem. Note that this problem contains some coding exercises.

```
> dim(boston)
[1] 506    4
> head(boston)
           chas rad  room medv
1 NOT_ON_RIVER   1 6.575 24.0
2 NOT_ON_RIVER   2 6.421 21.6
3 NOT_ON_RIVER   2 7.185 34.7
4 NOT_ON_RIVER   3 6.998 33.4
5 NOT_ON_RIVER   3 7.147 36.2
6 NOT_ON_RIVER   3 6.430 28.7
> summary(boston)
           chas          rad             room            medv
 NOT_ON_RIVER:471   24     :132   Min.   :3.561   Min.   : 5.00
 ON_RIVER    : 35   5      :115   1st Qu.:5.886   1st Qu.:17.02
                    4      :110   Median :6.208   Median :21.20
                    3      : 38   Mean   :6.285   Mean   :22.53
                    6      : 26   3rd Qu.:6.623   3rd Qu.:25.00
                    2      : 24   Max.   :8.780   Max.   :50.00
                    (Other): 61
```

(a) [2 pt / 74 pts] Using the terminology used in class, what type of predictor is `rad`?

nominal categorical

(b) [2 pt / 76 pts] Using the terminology used in class, what type of predictor is `room`?

Continuous

(c) [2 pt / 78 pts] Write one line of R code below that pulls out the 7th observation as a vector.

boston[7,]

(d) [2 pt / 80 pts] Write one line of R code below that pulls out the first 30 observations.

boston[1:30,]

(e) [3 pt / 83 pts] Write one line of R code below that pulls out all observations where: chas is "ON_RIVER" or medv is less than 20.

boston[boston$chas == "ON_RIVER" | boston$medv < 20,]

(f) [3 pt / 86 pts] Write one line of R code below that creates a new data frame boston_random containing the same observations as boston but where the order of the observations is random.

boston_random = boston[sample(1:nrow(boston)),]

(g) [2 pt / 88 pts] Write one line of R code below that finds the 25%ile of the variable medv.

quantile(boston$medv, 0.25)

(h) [2 pt / 90 pts] What would the following code produce?

```
as.matrix(boston[1 : 2, 3 : 4])
```

	1	2
1	6.575	24.0
2	6.421	21.6

(i) [1 pt / 91 pts] What would the following code produce?

```
class(as.matrix(boston[, 3: 4])[1, ])
```

numeric

(j) [1 pt / 92 pts] What would the following code produce?

```
class(as.matrix(boston[, 3: 4])[1, , drop = FALSE])
```

matrix

Problem 5 This last problem contains pure coding exercises.

(a) [5 pt / 97 pts] Complete the function below to spec. You don't have to use all the free lines given (in fact, it can be done in one line). You are free to use the `mean`, `sd`, `cov`, `cor` and other base `R` functions (but you cannot use `lm`).

```
#' This function implements the linear least squares regression algorithm
#' popularized by Sir Francis Galton in 1886.
#'
#' @param x      the continuous predictor
#' @param y      the continuous response
#' @return       a list containing a key ``b_0'' whose value is the inter-
#'               cept and a key ``b_1'' whose value is the slope
linear_least_squares_algorithm = function(x, y){

  b_1 = cor(x,y) * sd(y) / sd(x)
  b_0 = mean(y) - b_1 * mean(x)
  list(b_0 = b_0, b_1 = b_1)







}
```

(b) [2 pt / 99 pts] What does the following code produce?

```
xs = rep(NA, 5)
xs[3] = -8
xs[5] = 7
tot = 0
for (x in xs){
  if (is.na(x)){
    next
  }
  tot = tot + x
}
tot
```

-1

(c) [1 pt / 100 pts] What does the following code produce?

```
my_function = function(x, y = 2, z = 3, p = 4, q = 6, r = 0){
  (x + y + z) / (p + q + r)
}
my_function(1)
```

0.6

Problem 6 Some final theory for extra credit.

(a) [5 pt / 105 pts] [Extra credit] Consider the non-linearly separable SVM algorithm we studied in class. Now, describe an alternative $\mathcal{A}$ which instead of returning a function whose range is only $\{0, 1\}$, returns a function that can estimate $\mathbb{P}(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x})$. Describe it in English and use diagrams if necessary.