Math 390.4 / 650.3 Spring 2019
Midterm Examination One

Professor Adam Kapelner

Tuesday, March 5, 2019

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] Why aren't the numbers observed in simulations equal to the measured data for the same inputs?

(b) [4 pt / 8 pts] Are the "laws of science" a model or the reality? Explain.

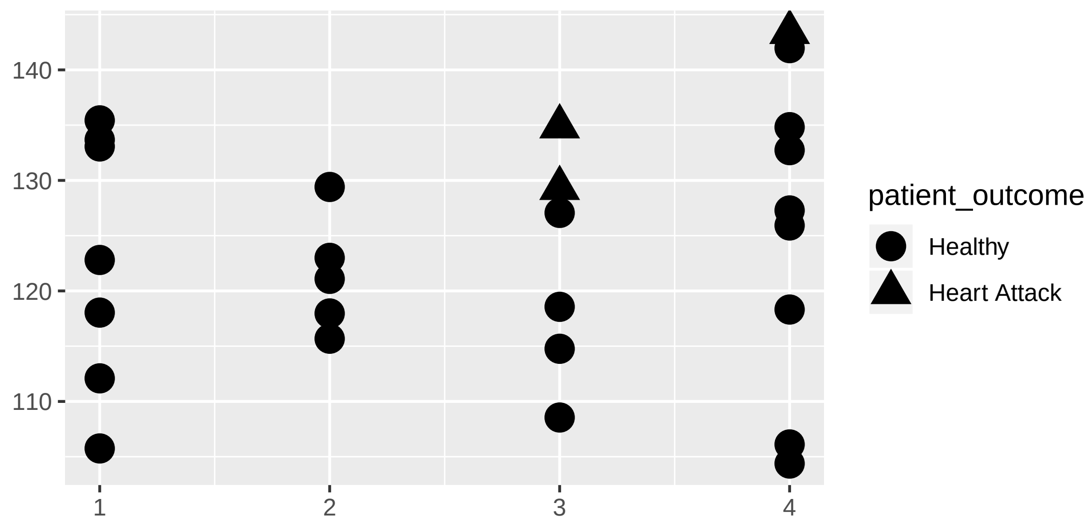
(c) [1 pt / 9 pts] Scientists now say that eating turmeric root can prevent cancer. Is this a *mathematical model*? Yes/no.

(d) [5 pt / 14 pts] [Extra Credit] If you wrote yes to (b), describe the mathematical model. If you wrote no to (b), explain how you would upgrade this statement to be a mathematical model.

Problem 2 We are trying to predict heart attacks based on two risk factors: blood pressure and cholesterol.

Blood pressure levels are based on the classification recommended by the Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation and Treatment of High Blood Pressure and are defined as follows: normal (systolic blood pressure <120 mm Hg and a diastolic blood pressure <80 mm Hg); pre-hypertension (systolic blood pressure 120–139 mm Hg or diastolic blood pressure 80–89 mm Hg); hypertension stage 1 (systolic blood pressure 140–159 mm Hg or diastolic blood pressure 90–99 mm Hg); and hypertension stage 2 (systolic blood pressure >160 mm Hg or diastolic blood pressure >100 mm Hg). Persons are classified into the higher blood pressure group if the systolic and diastolic values fall within more than one category. Cholesterol is measured as the LDL cholesterol.

Below is a plot of some historical pilot data.



(a) [2 pt / 16 pts] Based on the description in the problem header label the axes above.

(b) [2 pt / 18 pts] If we are going to build a model from this data, what type of statistical learning are we doing? Circle all that apply.

i) regression to predict y

ii) binary classification to predict y

iii) multiclass (i.e. specifically non-binary) classification to predict y

iv) finding t directly

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

vi) supervised learning

vii) unsupervised learning

(c) [2 pt / 20 pts] Assume we are doing supervised learning. Relabel the axes and legend above using parentheses around x_j and y where j is the index on the variable number which you need to determine. For example, an axis may be labeled "# of phone calls (x_{17})".

(d) [2 pt / 22 pts] Is this data linearly separable? Yes / no and explain your answer.

(e) [2 pt / 24 pts] Denote $\mathbb{D} :=< X, \boldsymbol{y} >$. What is $\dim[X]$?

(f) [2 pt / 26 pts] Given $\mathbb{D}$, find g_0, the null model.

(g) [2 pt / 28 pts] Let $\mathcal{A}$ = perceptron learning algorithm. What will its output be? Circle all that apply.

i) $\hat{y}$

ii) $\mathcal{A}$

iii) g

iv) h^*

v) $x_{\cdot 1}, \ldots, x_{\cdot p}$

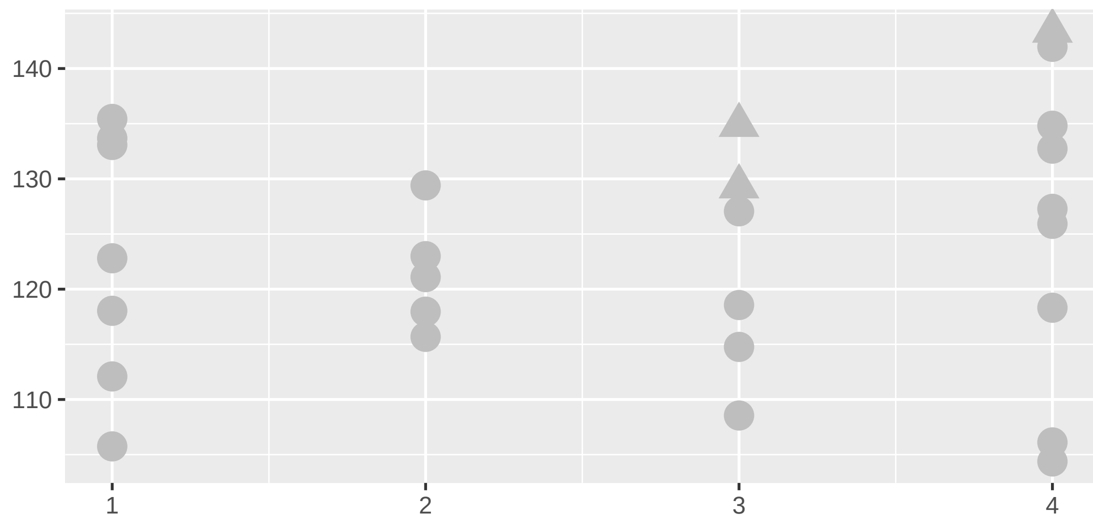
vi) f

vii) $z_1, \ldots, z_t$

(h) [4 pt / 32 pts] Let $\mathcal{A}$ = perceptron learning algorithm. Starting the algorithm at the zero vector of the appropriate dimension, draw the algorithm's output line on the plot as a dotted line or explain below why you are unable to do so.

(i) [2 pt / 34 pts] Let $\mathcal{A}$ = an algorithm that minimizes average hinge loss. Write the objective function of the algorithm below.

(j) [3 pt / 37 pts] Let $\mathcal{A}$ = SVM with the Vapnik objective function with a λ specified to be small but non-zero. Draw the algorithm's output line on the plot below or explain below why you are unable to do so.



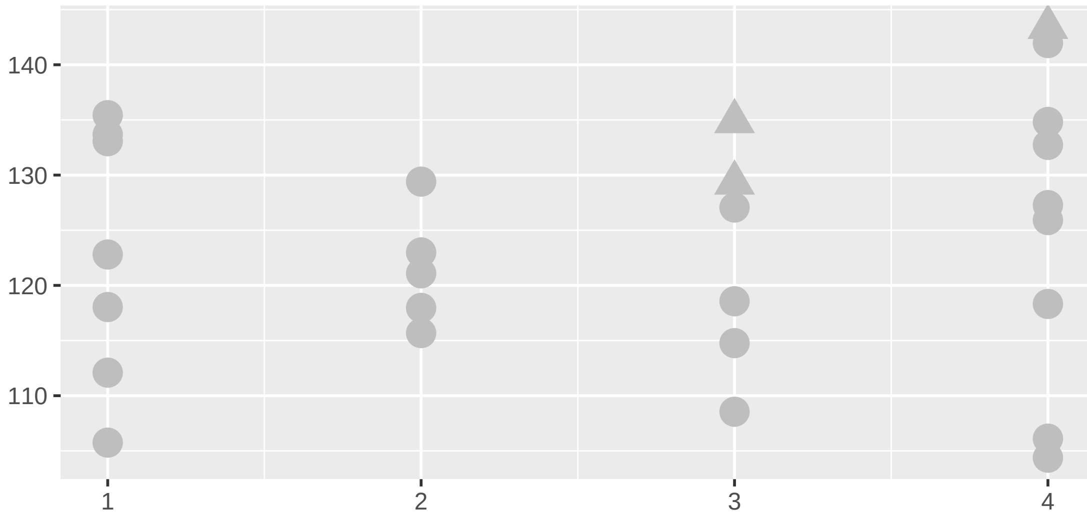
(k) [2 pt / 39 pts] What is the average classification error of the model in the previous question?

(l) [2 pt / 41 pts] What is the approximate R^2 of this fit?

i) <0%
ii) 0%
iii) 2.5%
iv) 25%
v) 95%
vi) 100%
vii) $\mathcal{A}$ cannot produce an output thus R^2 cannot be estimated.

(m) [2 pt / 43 pts] Why is R^2 an inappropriate metric to be using here to measure model performance?

(n) [2 pt / 45 pts] Draw a model below that has zero average hinge error.



(o) [4 pt / 49 pts] Let $\mathcal{A}$ = KNN with d = Euclidean distance. Evaluate the following:

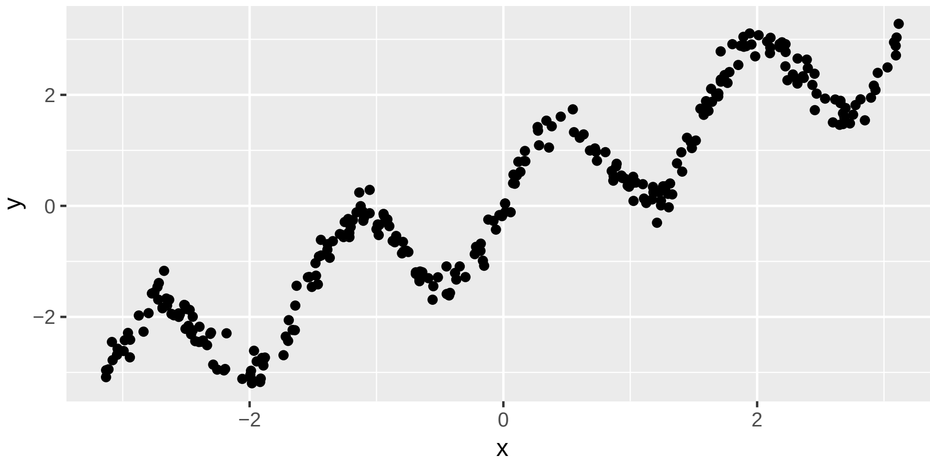
i) $K = 1,\ g(4, 145) =$

ii) $K = 1,\ g(3, 125) =$

iii) $K = 4,\ g(3, 125) =$

(p) [5 pt / 54 pts] Is there a problem with using $\mathcal{A}$ = perceptron, $\mathcal{A}$ = SVM and $\mathcal{A}$ = KNN with $\mathbb{D}$? Yes / no. Discuss.

Problem 3 The following dataset is a mock view of a financial asset. The x axis represents time and the y axis represents value. There are $n = 300$ data points and $s_x^2 = 3.322$ and $s_y^2 = 3.196$.



(a) [2 pt / 56 pts] Estimate $\mathcal{X}$ and $\mathcal{Y}$.

(b) [2 pt / 58 pts] Estimate the equation for g_0.

(c) [2 pt / 60 pts] Estimate the R^2 for g_0.

(d) [5 pt / 65 pts] If $\mathcal{A} = \text{OLS}$, estimate the percentage of the RMSE of g belonging to each of the three errors.

i) Name of error:

Percentage of RMSE:

ii) Name of error:

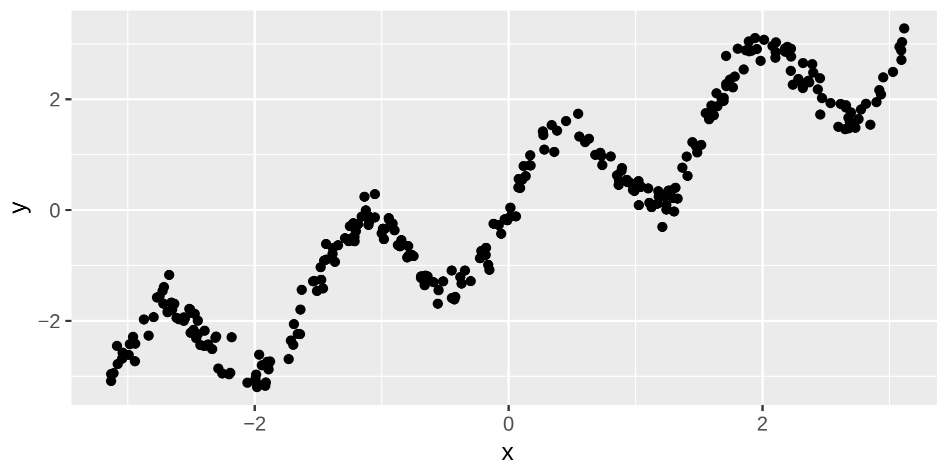
Percentage of RMSE:

iii) Name of error:

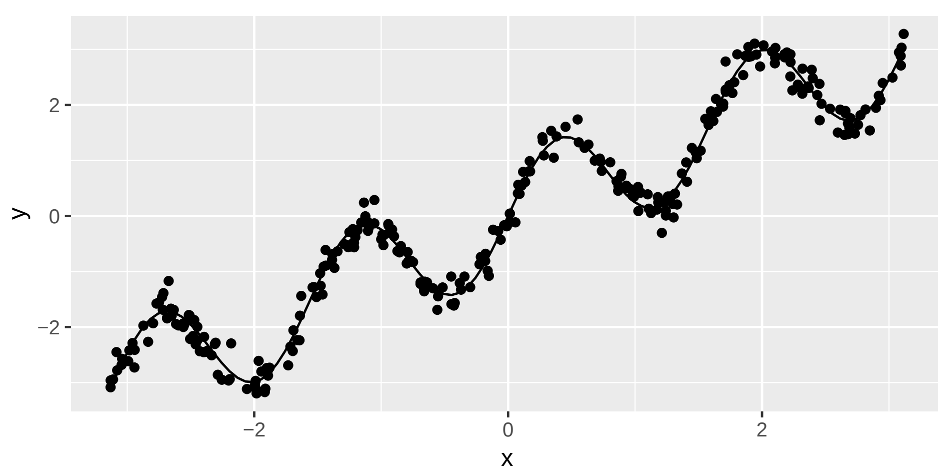
Percentage of RMSE:

(e) [4 pt / 69 pts] Let $\mathcal{A}$ = minimize the least squares error but provide a better $\mathcal{H}$ than the set $\{w_0 + w_1 x \ : \ w_0, w_1 \in \mathbb{R}\}$ where "better" means that the elements h can much better approximate f.

(f) [2 pt / 71 pts] Draw h^*, an element of your set in (d) on the plot below:



(g) [5 pt / 76 pts] Compute the R^2 of the model depicted as the solid line below *as best as you can.*



Problem 4 We will now be looking at the `diamonds` dataset. Below is some `R` code that gives background on this data frame which will be referenced throughout this problem. The response variable that is usually modeled is `price`. This problem contains some coding exercises.

```
> diam = ggplot2::diamonds
> dim(diamonds)
[1] 53940    10
> summary(diamonds)
     carat               cut          color      clarity          depth
 Min.   :0.2000   Fair     : 1610   D: 6775   SI1    :13065   Min.   :43.00
 1st Qu.:0.4000   Good     : 4906   E: 9797   VS2    :12258   1st Qu.:61.00
 Median :0.7000   Very Good:12082   F: 9542   SI2    : 9194   Median :61.80
 Mean   :0.7979   Premium  :13791   G:11292   VS1    : 8171   Mean   :61.75
 3rd Qu.:1.0400   Ideal    :21551   H: 8304   VVS2   : 5066   3rd Qu.:62.50
 Max.   :5.0100                     I: 5422   VVS1   : 3655   Max.   :79.00
                                    J: 2808   (Other): 2531
     table           price             x                y                z
 Min.   :43.00   Min.   :  326   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
 1st Qu.:56.00   1st Qu.:  950   1st Qu.: 4.710   1st Qu.: 4.720   1st Qu.: 2.910
 Median :57.00   Median : 2401   Median : 5.700   Median : 5.710   Median : 3.530
 Mean   :57.46   Mean   : 3933   Mean   : 5.731   Mean   : 5.735   Mean   : 3.539
 3rd Qu.:59.00   3rd Qu.: 5324   3rd Qu.: 6.540   3rd Qu.: 6.540   3rd Qu.: 4.040
 Max.   :95.00   Max.   :18823   Max.   :10.740   Max.   :58.900   Max.   :31.800
```

(a) [2 pt / 78 pts] In this $\mathbb{D}$, what is n and p?

(b) [2 pt / 80 pts] What is the type of the variable `carat`?

(c) [2 pt / 82 pts] What is the type of the variable `cut`?

(d) [2 pt / 84 pts] If you were fitting an OLS model of `price` using `color`, what would p be in that model?

(e) [2 pt / 86 pts] Write code that extracts every 50th diamond observation.

(f) [3 pt / 89 pts] Write code that adds a new variable to the data frame named `customer_favorite` that is 1 if the `cut` is `ideal` and the `color` is either `G` or `J` and the `depth` is 90%ile or above.

(g) [3 pt / 92 pts] Assume that `x`, `y` and `z` are the spatial dimensions of the stone. Write code below that creates a data frame called `tinies` that contains diamonds that have volume less than 50.

(h) [3 pt / 95 pts] Describe the output of this script as completely as possible.

```r
> dict = list()
> for (color in unique(diam$color)){
  dict[[color]] = diam[diam$color == color, "price"]
}
> dict[["D"]]
```

Problem 5

This last problem contains a pure coding exercise.

(a) [5 pt / 100 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`).

```r
#' This function implements the linear least squares regression algorithm
#' for one covariate 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, a key ``b_1'' whose value is the slope, a key ``Rsq''
#'            that is the R-squared of the fit.
linear_least_squares_algorithm = function(x, y){
  b_1 = cor(x, y) * sd(y) / sd(x)
  b_0 = mean(y) - b_1 * mean(x)











}
```