MATH 390.4 / 650.2 Spring 2020 Homework #2

Professor Adam Kapelner

Due 11:59PM Monday KY604, February 24, 2020

(this document last updated 5:20am on Wednesday 26th February, 2020)

Instructions and Philosophy

The path to success in this class is to do many problems. Unlike other courses, exclusively doing reading(s) will not help. Coming to lecture is akin to watching workout videos; thinking about and solving problems on your own is the actual "working out." Feel free to "work out" with others; **I want you to work on this in groups.**

Reading is still *required*. For this homework set, read the first chapter of "Learning from Data" and Chapter 2 of Silver's book. Of course, you should be googling and reading about all the concepts introduced in class online. This is your responsibility to supplement in-class with your own readings.

The problems below are color coded: green problems are considered *easy* and marked "[easy]"; yellow problems are considered *intermediate* and marked "[harder]", red problems are considered *difficult* and marked "[difficult]" and purple problems are extra credit. The *easy* problems are intended to be "giveaways" if you went to class. Do as much as you can of the others; I expect you to at least attempt the *difficult* problems.

This homework is worth 100 points but the point distribution will not be determined until after the due date. See syllabus for the policy on late homework.

Up to 7 points are given as a bonus if the homework is typed using LaTeX. Links to instaling LaTeX and program for compiling LaTeX is found on the syllabus. You are encouraged to use `overleaf.com`. If you are handing in homework this way, read the comments in the code; there are two lines to comment out and you should replace my name with yours and write your section. The easiest way to use overleaf is to copy the raw text from hwxx.tex and preamble.tex into two new overleaf tex files with the same name. If you are asked to make drawings, you can take a picture of your handwritten drawing and insert them as figures or leave space using the "\vspace" command and draw them in after printing or attach them stapled.

The document is available with spaces for you to write your answers. If not using LaTeX, print this document *including this first page* and write in your answers. I do not accept homeworks which are *not* on this printout.

NAME: *Pizon Shetu*

Problem 1

These are questions about Silver's book, chapter 2.

(a) [harder] If one's goal is to fit a model for a phenomenon y, what is the difference between the approaches of the hedgehog and the fox? Answer using notation from class (i.e. $t, f, g, h^*, \delta, \epsilon, e, t, z_1, \ldots, z_t, \mathbb{D}, \mathcal{H}, \mathcal{A}, \mathcal{X}, \mathcal{Y}, X, y, n, p, x_{\cdot 1}, \ldots, x_{\cdot p}, x_{1\cdot}, \ldots, x_{n\cdot}$, etc.). Connecting this to the modeling framework should really make you think about what Tetlock's observation means for political and historical phenomena.

The fox had more features such as $(x_1, x_2....x_n)$ which allowed them to make more accurate predictions, this was due to their $\mathcal{H}$ having a wider range of outcomes whereas the hedgehog had fewer features which led to more errors and inaccurate predictions.

(b) [easy] Why did Harry Truman like hedgehogs? Are there a lot of people that think this way?

He liked hedgehogs because in politics hedgehogs create stories to engage more people allowing it to be more dramatic, plus he was frustrated with the foxes in his administration.

(c) [difficult] Why is it that the more education one acquires, the less accurate one's predictions become?

An overload of information may haze the person, as its harder for a person to recognize which information to use and which not to use. Personal bias also plays a role into predictions.

(d) [easy] Why are probabilistic classifiers (i.e. algorithms that output functions that return probabilities) better than vanilla classifiers (i.e. algorithms that only return the class label)? We will move in this direction in class soon.

A vanilla classifier is based of bias which is poor compared to a probabilistic classifier which uses different types of data and limits to make accurate predictions.

Problem 2

These are questions about the SVM.

(a) [easy] State the hypothesis set $\mathcal{H}$ inputted into the support vector machine algorithm. Is it different than the $\mathcal{H}$ used for $\mathcal{A}$ = perceptron learning algorithm?
Yes they are different we can see the $\mathcal{H}$ for SVM is $\mathcal{H} = 1\vec{w} * \vec{x} + b > 0 : \vec{W} \in \mathbb{R}^3 b \in \mathbb{R}$ whereas the $\mathcal{H}$ for perceptron is $\mathcal{H} = 1\vec{w} * \vec{x} > 0 : \vec{W} \in \mathrm{R}^3 we can see that the perceptron does not have b.$

(b) [difficult] [MA] Prove the SVM converges. State all assumptions. Write it on a separate page.

(c) [difficult] Let $\mathcal{Y} = \{-1, 1\}$. Rederive the cost function whose minimization yields the SVM line in the linearly separable case.

$\forall y_i = -1 = w \cdot X - b = -1$
$\forall y_i = 1 = w \cdot X - b = 1$
Ideally we would like all 1's above the line for $y_i = 1$ and all the -1's to be below for all $y_i = -1$
if $y_i = 1$, $w \cdot X - b \geq 1$
if $y_i = -1$, $w \cdot X - b \leq -1$
Minimize $||w||$ with the constraint $\forall i$, $y_i(w \cdot X - b) \geq 1$

(d) [easy] Given your answer to (c) rederive the cost function using the "soft margin" i.e. the hinge loss plus the term with the hyperparameter λ. This is marked easy since there is just one change from the expression given in class.

$SHE = \sum_{i=1}^{n} max(0, 1 - y_i(\vec{w} * \vec{x} - b))$
$argmin(\frac{1}{n} SHE + \lambda||w||)$

Problem 3

These are questions are about the k nearest neighbors (KNN) algorithm.

(a) [easy] Describe how the algorithm works. Is k a "hyperparameter"?

Pick a x^*, the algorithm looks at the k data points in the training set, $\mathbb{D}$, that are closest to x^*, (usually Euclidean distance is used). Then, the algorithm returns the mode of the y's of the neighbors. K is indeed an hyper-parameter which is determined by the user not the algorithm

(b) [difficult] Assuming $\mathcal{A}$ = KNN, describe the input $\mathcal{H}$ as best as you can.

The input for $\mathcal{H}$ is an indicator function with an argmin function, where $\mathcal{H}$ will contain all the possible K differences. Meaning it will find a the smallest distance between our x^* and all its neighbors.

(c) [difficult] When predicting on $\mathbb{D}$ with $k = 1$, why should there be zero error? Is this a good estimate of future error when new data comes in? (Error in the future is called *generalization error* and we will be discussing this later in the semester).

There should be no error as if K = 1 then the algorithm will select x^* itself which would have distance of 0 thus it is the smallest distance.

Problem 4

These are questions about the linear model with $p = 1$.

(a) [easy] What does $\mathbb{D}$ look like in the linear model with $p = 1$? What is $\mathcal{X}$? What is $\mathcal{Y}$?

If $p = 1$ $\mathbb{D}$ is a 1×n vector with $\mathcal{X} = \mathbb{R}$ $\mathcal{Y} = \mathbb{R}$

(b) [easy] Consider the line fit using the ordinary least squares (OLS) algorithm. Prove that the point $< \bar{x}, \bar{y} >$ is on this line. Use the formulas we derived in class.
Recall that $y = b_0 + b_1 x$ and that

$$b_1 = \frac{\Sigma x_i y_i - n\bar{y}\bar{x}}{\Sigma x_i^2 - n\bar{x}^2} \qquad \text{plus we have} \qquad b_0 = \bar{y} - \frac{\Sigma x_i y_i - n\bar{y}\bar{x}}{\Sigma x_i^2 - n\bar{x}^2}\bar{x}$$

Now we have to show that $\bar{y} = b_0 + b_1\bar{x}$

$$y = \frac{\Sigma x_i y_i - n\bar{y}\bar{x}}{\Sigma x_i^2 - n\bar{x}^2}\bar{x} + \bar{y} - \frac{\Sigma x_i y_i - n\bar{y}\bar{x}}{\Sigma x_i^2 - n\bar{x}^2}\bar{x}$$

$$y = \bar{y}$$

Thus showing that $< \bar{x}, \bar{y} >$ is on the line

(c) [harder] Consider the line fit using OLS. Prove that the average prediction $\hat{y}_i := g(x_i)$ for $x_i \in \mathbb{D}$ is $\bar{y}$.
Recall that

$$\hat{\boldsymbol{y}} = b_0 + b_1\bar{x}$$

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{y}}_i &= \frac{1}{n}\sum_{i=1}^{n} b_0 + \frac{1}{n}\sum_{i=1}^{n} b_1 x_i \\ &= \frac{1}{n} n b_0 + b_1 \frac{1}{n}\sum_{i=1}^{n} x_i \\ &= b_0 + b_1\bar{x} \\ &= \bar{y}\end{aligned}$$

(d) [harder] Consider the line fit using OLS. Prove that the average residual e_i is 0 over $\mathbb{D}$.

Recall that

$$\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \bar{y}$$

And,

$$\frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$$

So

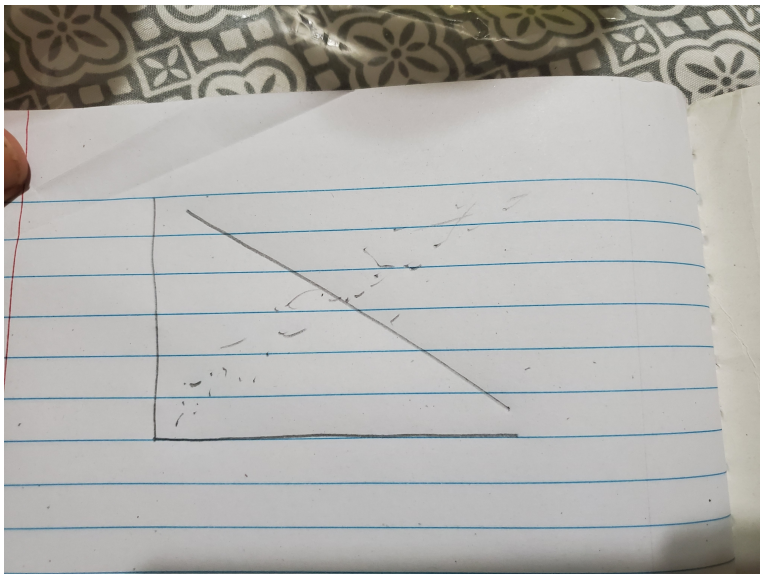
$$\frac{1}{n}\sum_{i=1}^{n} e_i = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i - \frac{1}{n}\sum_{i=1}^{n} y_i$$

Thus

$$\frac{1}{n}\sum_{i=1}^{n} e_i = \bar{y} - \bar{y} = 0$$

(e) [harder] Why is the RMSE usually a better indicator of predictive performance than R^2? Discuss in English.
While the R^2 tells us how good of a "fit" our model is while this can be useful this is not as indicative like the RMSE where it tells us how far are the prediction is from the model.

(f) [harder] R^2 is commonly interpreted as "proportion of the variance explained by the model" and proportions are constrained to the interval $[0, 1]$. While it is true that $R^2 \leq 1$ for all models, it is not true that $R^2 \geq 0$ for all models. Construct an explicit example $\mathbb{D}$ and create a linear model $g(x) = w_0 + w_1 x$ whose $R^2 < 0$.
This is when our model does a worse job than the null model for instances like this graph



(g) [harder] [MA] Prove that the OLS line always has $R^2 \in [0, 1]$ on a separate page.

(h) [difficult] You are given $\mathbb{D}$ with n training points $< x_i, y_i >$ but now you are also given a set of weights $[w_1\ w_2\ \ldots\ w_n]$ which indicate how costly the error is for each of the i points. Rederive the least squares estimates b_0 and b_1 under this situation. Note that these estimates are called the *weighted least squares regression* estimates. This variant $\mathcal{A}$ on OLS has a number of practical uses, especially in Economics. No need to simplify your answers like I did in class (i.e. you can leave in ugly sums).

SSWE $= \sum_{i=1}^{n} w_i(y_i - (b_0 + b_1 x_i)^2$
SSWE $= \sum_{i=1}^{n} w_i(y_i^2 + b_0^2 + b_1^2 x_i^2 - 2y_i b_0 - 2y_i b_1 x_i + 2b_0 b_1 x_i^2)$

$$= \sum_{i=1}^{n} y_i^2 w_i + b_0^2 \sum_{i=1}^{n} w_i + b_1^2 \sum_{i=1}^{n} w_i y_i^2 - 2b_0 \sum_{i=1}^{n} w_i y_i - 2b_1 \sum_{i=1}^{n} w_i x_i y_i + 2b_0 b_1 \sum_{i=1}^{n} w_i x_i$$

$$\frac{\partial}{\partial b_0}(SSWE) = 2b_0 \sum w_i - 2 \sum w_i y_i + 2b_1 \sum w_i x_i = 0$$

$$b_0 = \frac{\sum w_i y_i}{\sum wi} - b_1 \frac{\sum w_i x_i}{\sum w_i}$$

$$\frac{\partial}{\partial b_1}(SSWE) = 2b_1 \sum w_i y_i^2 - 2 \sum w_i x_i y_i + 2b_0 \sum w_i x_i = 0$$

replace b_0 with $\frac{\sum w_i y_i}{\sum wi} - b_1 \frac{\sum w_i x_i}{\sum w_i}$

$$2b_1 \sum w_i y_i^2 - 2 \sum w_i x_i y_i + 2(\frac{\sum w_i y_i}{\sum wi} - b_1 \frac{\sum w_i x_i}{\sum w_i}) \sum w_i x_i = 0$$

$$b_1 \sum w_i y_i^2 - b_1 \frac{(\sum w_i x_i)^2}{\sum w_i} = \sum w_i x_i y_i - \frac{(\sum w_i y_i)(\sum w_i x_i)}{\sum wi}$$

$$b_1 = \frac{\sum w_i x_i y_i - \frac{(\sum w_i y_i)(\sum w_i x_i)}{\sum wi}}{\sum w_i y_i^2 - \frac{(\sum w_i x_i)^2}{\sum w_i}}$$

(i) [harder] [MA] Interpret the ugly sums in the b_0 and b_1 you derived above and compare them to the b_0 and b_1 estimates in OLS. Does it make sense each term should be altered in this matter given your goal in the weighted least squares?

Since we have to account for the set or the weights of all the $x_i's$ and $y_i's$ it makes sense for our terms to be altered. Our OLS will differ for each i for the weights

(j) [difficult] [MA] In class we talked about $x_{raw} \in \{\text{red}, \text{green}\}$ and the OLS model was the sample average of the inputted x where $b_0 = \bar{y}_r$ and $b_1 = \bar{y}_g - \bar{y}_r$. Reparameterize $\mathcal{H} = \{w_1 \mathbb{1}_{x_{raw} = \text{ red}} + w_2 \mathbb{1}_{x_{raw} = \text{ green}} \ : \ w_1, w_2 \in \mathbb{R}\}$ and prove that the OLS estimates are $b_1 = \bar{y}_r$ and $b_2 = \bar{y}_g$.

$x_{raw} \in \{\text{red}, \text{green}\}$ $x_i = \mathbb{1}_{x_{raw}=\text{green}}$ $\hat{\boldsymbol{y}} = b_o + b_1 x_1 - > b_0 = \bar{y}r, b_1 = \bar{y}g * \bar{y}r$
$x_{raw} \in \{\text{low,high}\}$ $x_1 = \mathbb{1}_x = \mathcal{H}$ $\hat{\boldsymbol{y}} = \bar{y}_L$if x = low$or = \bar{y}_H$if x = high

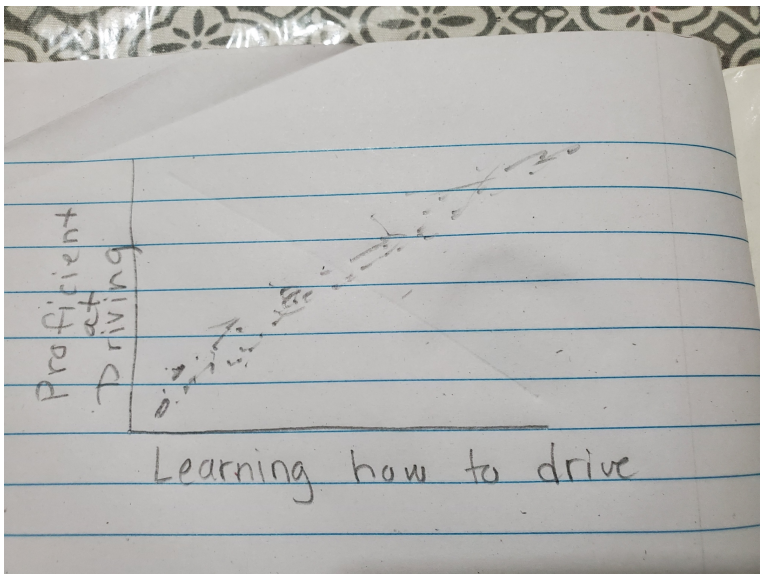
(k) [difficult] In class we talked about $x_{raw} \in \{\text{red}, \text{green}\}$ and the OLS model was the sample average of the inputted x. Imagine if you have the additional constraint that x_{raw} is ordinal e.g. $x_{raw} \in \{\text{low}, \text{high}\}$ and you were forced to have a model where

$g(\text{low}) \leq g(\text{high})$. Invent an algorithm $\mathcal{A}$ that can solve this problem.

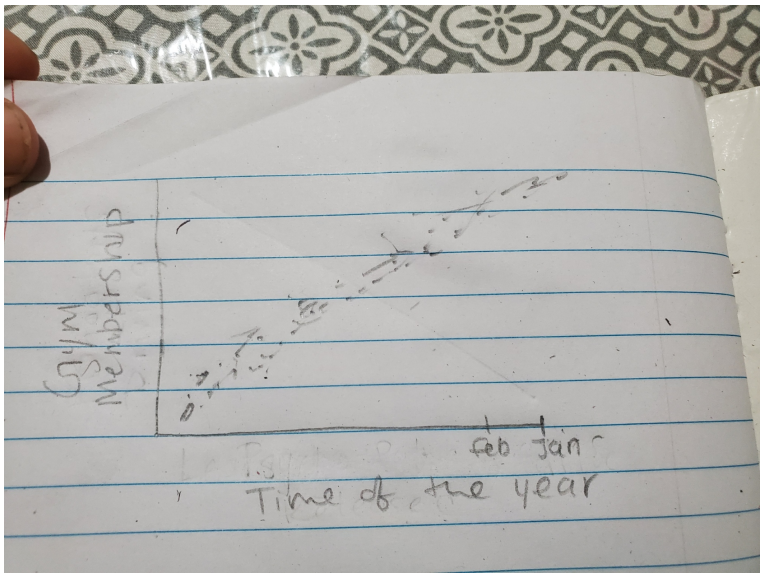
Problem 5

These are questions about association and correlation.

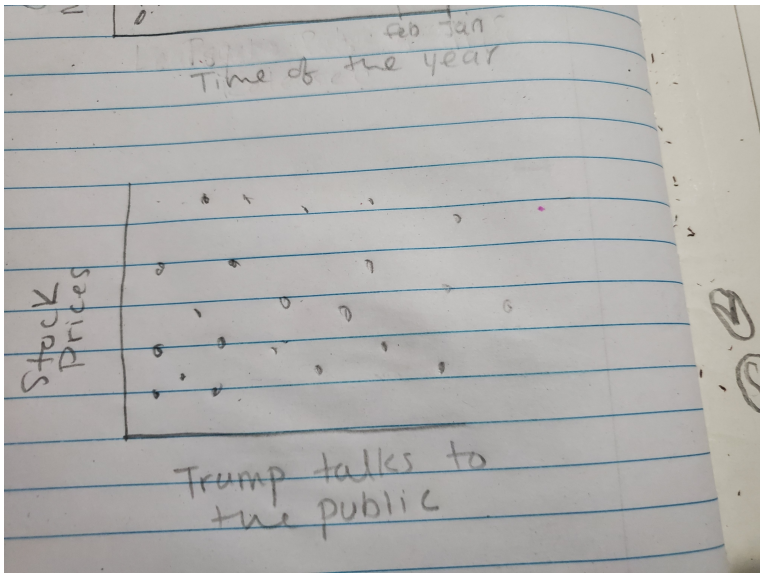
(a) [easy] Give an example of two variables that are both correlated and associated by drawing a plot. Learning to drive and being proficient at driving



(b) [easy] Give an example of two variables that are not correlated but are associated by drawing a plot.



[easy] Give an example of two variables that are not correlated nor associated by drawing a plot.



[easy] Can two variables be correlated but not associated? Explain.

If there is any correlation that implies there some level of relation, meaning some level of association. But just because two things are associated does not imply correlation.

[difficult] [MA] Prove association $\not\Rightarrow$ correlation. This requires some probability theory.

E[X] = 0, E[Y] = 0 where $Y = X^2$
$COV[X,Y] = E[(X - M_X)(Y - M_y)]$
$COV[X,Y] = E[(X - 0)(Y - 1)]$
$COV[X,Y] = E[(XY - X)]$
$COV[X,Y] = E[XY] - E[X]$
$COV[X,Y] = E[X^3] - E[X]$
$COV[X,Y] = 0 - 0$
$thus COV[X,Y] = 0$