# MATH 390.4 / 650.2 Spring 2020 Homework #2

Tyron Samaroo

Monday 24th February, 2020

## 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.

Using the hedgehog approach when trying to fit a model for a phenomenon $y$ might be saying that their function g is the unknown target function f that predict the phenomenon.So hedgehog approach believe that $g = f$.

Using the fox approach when trying to fit a model for a phenomenon $y$ foxes know that its difficult to figure out the unknown target function $f$ so they know that they can choose from a hypothesis set $\mathcal{H}$ apply a learning algorithm $\mathcal{H}$ to get a function $g$ that can best approximate the $f$ function.

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

Harry Truman like hedgehogs because because hedgehogs make quick answer, he did not like that he had foxes in his administration that couldn't give him a unqualified answer

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

The more education you have the more things you know. The more things the less critical you will be about each bit of information you educate yourself with. You have retain so much information that you will not be able to critically analyze every one of them.

(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.

With probabilistic classifiers you are able to articulate a range of possible outcomes where as vanilla classifiers only return one thing and claim it knows it all.

## 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?

No the $\mathcal{H}$ is same that is used for $\mathcal{A}$ = perceptron learning algorithm

$\mathcal{H} = \{\mathbb{1}_{w_o\bar{x}+b} > 0 : w \in R^p, b \in R\}$

(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.

| | |
|---|---|
| Let $y_i = 1$ | Let $y_i = -1$ |
| w $\cdot \bar{x}_i - (b+1) \geq 0$ | w $\cdot \bar{x}_i - (b+1) \leq 0$ |
| $w \cdot \bar{x}_i - b \geq 1$ | $w \cdot \bar{x}_i - b \leq 1$ |
| Multiply Both Sides by $y_i$ | Multiply Both Sides by $y_i$ |
| $y_i w \cdot \bar{x}_i - b \geq 1(y_i)$ | $y_i w \cdot \bar{x}_i - b \leq 1(y_i)$ |
| Substitute $y_i = 1$ | Substitute $y_i = -1$ |
| $y_i w \cdot \bar{x}_i - b \geq 1$ | $y_i w \cdot \bar{x}_i - b \leq -1$ |
| | Using inequlity rules |
| | $y_i w \cdot \bar{x}_i - b \geq 1$ |

So the cost function(Hinge Error) would be $H_i = max(0, \{(1 - y_i)w \cdot \bar{x}_i - b\})$ Which then will make the sum of hinge error(SHE) would be

$$\sum_{n=1}^{n} max(0, \{(1 - y_i)w \cdot \bar{x}_i - b\})$$

(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 $\lambda$. This is marked easy since there is just one change from the expression given in class.

$\underset{w \in R^p, b \in R}{\text{argmin}} \left\{ \frac{1}{n} \sum_{n=1}^{n} max(0, \{(1 - y_i)w \cdot \bar{x}_i - b\}) + \lambda \, ||w||^2 \right.$

## Problem 3

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

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

In KNN we have the function $g$ that finds the closes $\bar{x}_i$ and returns $y_i$. So we basically classify each points by its "nearest neighbor" by the smallest distance. We have ths Yes $k$ is a "hyperparameter".

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

If our $\mathcal{A}$ is KNN out input $\mathcal{H}$ would be the same $\mathcal{H} = \{\boldsymbol{w}.\bar{x} :\in R^{p+1}\}$ where $\mathcal{H}$ is the set of all linear models.

(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).

When $k = 1$ The estimation is based on the closest neighbor. When test on training set you will get no error since its nearest neighbor will be the point itself. So when new data comes in there will be errors since the model only classify for the data it already knows

## 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 $\mathcal{Y}$?

The $\mathbb{D}$ looks like a line.$\mathcal{X}$ is input values x from $\mathbb{D}$ and our $\mathcal{Y}$ is output value y from $\mathcal{D}$

(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.

$\sum (\mathrm{y}_i - w_o - w_1 x_i)^2$ If we solve this the partial derivative of SSE with respect to partial $w_o$ is in the set of 0 and we get $w_0 = \bar{y} - w_1 \bar{x}$ We solve SSE also with respect to partial $w_1$ in the set of 0 and we get $w_1 = \frac{\sum x_i y_1 + n\bar{x}\bar{y}}{\sum x_i^2 + n\bar{x}^2}$ This shows that the we will have the points $< \bar{x}, \bar{y} >$ 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 $\bar{y}$.

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

(e) [harder] Why is the RMSE usually a better indicator of predictive performance than $R^2$? Discuss in English.

RMSE usually a better indicator of predictive performance because you know how much you are off in your prediction based on readable units. $R^2$ gives just a percentage you can have a model with $R^2$ with accuracy of 99% but model can be off by a huge value

(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 $R^2 < 0$.

(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).

(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?

(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 $b_2 = \bar{y}_g$.

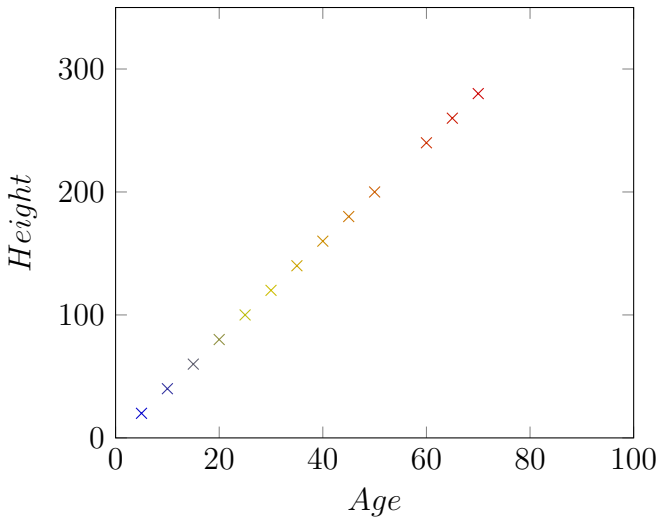
(k) [difficult] In class we talked about $x_{raw} \in \{\text{red}, \text{green}\}$ and the OLS model was the sample average of the inputted $x$. Imagine if you have the additional constraint that $x_{raw}$ is ordinal e.g. $x_{raw} \in \{\text{low}, \text{high}\}$ and you were forced to have a model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

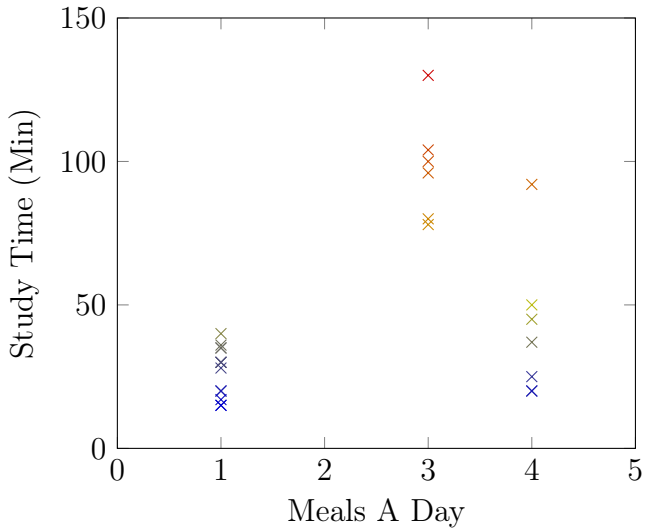
(a) [easy] Give an example of two variables that are both correlated and associated by drawing a plot.

You can see a positive coloration here that as height increase so does age and visa versa. You can easily associate someone height based on their age visa versa.

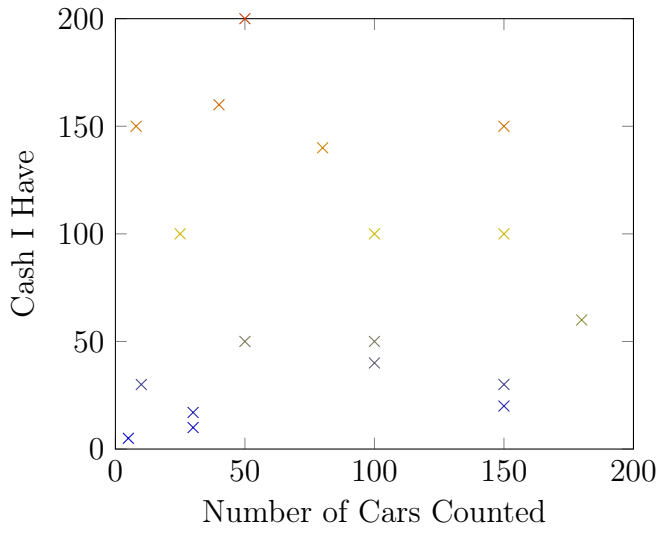



(b) [easy] Give an example of two variables that are not correlated but are associated by drawing a plot.

You cannot tell that the meals you eat will determine the amount of time a person will study but you can associate those who eat one meal to around 20-40 minutes of study time




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




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

Yes two variables can be correlated but not associated. For example you can say there's correlation between people who died and drank water.But you can associated dying with drinking water since you are ignoring that there's tons of other reason why people die.

(e) [difficult] [MA] Prove association $\not\Longrightarrow$ correlation. This requires some probability theory.