

MATH 390.4 / 650.2 Spring 2020 Homework #2

Professor Adam Kapelner

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

(this document last updated 8:20am on Friday 21st 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 L^AT_EX. Links to installing L^AT_EX and program for compiling L^AT_EX is found on the syllabus. You are encouraged to use [overleaf.com](https://www.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 L^AT_EX, print this document *including this first page* and write in your answers. **I do not accept homeworks which are *not* on this printout.**

NAME: _____

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. $t, f, g, h^*, \delta, \epsilon, e, t, z_1, \dots, z_t, \mathbb{D}, \mathcal{H}, \mathcal{A}, \mathcal{X}, \mathcal{Y}, X, y, n, p, x_{:1}, \dots, x_{:p}, x_{1:}, \dots, x_{n:}$, etc.). Connecting this to the modeling framework should really make you think about what Tetlock's observation means for political and historical phenomena.
- (b) [easy] Why did Harry Truman like hedgehogs? Are there a lot of people that think this way?
- (c) [difficult] Why is it that the more education one acquires, the less accurate one's predictions become?
- (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.

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?

- (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.

- (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.

Problem 3

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

- (a) [easy] Describe how the algorithm works. Is k a “hyperparameter”?
- (b) [difficult] Assuming $\mathcal{A} = \text{KNN}$, describe the input \mathcal{H} as best as you can.

- (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).

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} ?
- (b) [easy] Consider the line fit using the ordinary least squares (OLS) algorithm. Prove that the point $\langle \bar{x}, \bar{y} \rangle$ is on this line. Use the formulas we derived in class.
- (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.
- (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 $g(x) = w_0 + w_1x$ 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 $\langle x_i, y_i \rangle$ but now you are also given a set of weights $[w_1 \ w_2 \ \dots \ w_n]$ 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 manner 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.

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

(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.

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