MATH 342W / 650.4 Spring 2024 Homework #2

Benjamin Minkin

Tuesday 20th February, 2024

Problem 1

These are questions about Silver's book, chapter 2, 3. Answer the questions 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).

(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? Connecting this to the modeling framework should really make you think about what Tetlock's observation means for political and historical phenomena.

The hedgehog would try to fit the model towards his broad ideas about how the world functions. This would create a biased model that seeks to confirm his biases. The fox however, would try to understand the phenomenon as a unique process and would create a less biased model but no point of reference.

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

Most people like hedgehogs. They claim to understand everything and can explain historical events from a simplistic perspective. Not many people want to understand history in a nuanced approach that describes events as a result of a multitude of influences.

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

This is not necessarily true but is often the case because of how education is run. Students are taught the beliefs and biases of the teacher. This creates underlying presumptions that can hinder the creation of unbiased models.

(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.

This approach allows for better back-testing of models. One can look back and see how accurate the predictions were. A good probabilistic model will be right at the predicted rate. This is better than vanilla classifiers where it is harder to do so.

(e) [easy] What algorithm that we studied in class is PECOTA most similar to?

This is similar to the nearest k neighbors algorithm.

(f) [easy] Is baseball performance as a function of age a linear model? Discuss.

No. It is presumed that players get better with experience and worse with age. This leads to a peak at some point in the middle of their career. Psychologically, this makes sense. A good player often never quits at his physical peak and usually will retire after he can no longer play at his best.

(g) [harder] How can baseball scouts do better than a prediction system like PECOTA?

They can watch the players play. This firsthand view along with years of experience can cause scouts to have a better prediction.

(h) [harder] Why hasn't anyone (at the time of the writing of Silver's book) taken advantage of Pitch f/x data to predict future success?

There are lots of stats recorded in baseball. Pitch f/x is just one that is collected. It is hard to understand which statistics are meaningful and which are noise.

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?

$$\mathcal{A} = \{\mathbb{1}_{w \cdot x} - b \leq 0 : w \in \mathbb{R}, b \in \mathbb{R}\}$$

This is limited to a line/hyperplane that splits the wedge near the center. This is different than perceptron method as that will output any valid line in the wedge.

(b) [E.C.] Prove the max-margin linearly separable 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.

$$L_i(w, b) = max(0, 1 - y_i(f(x_i)))$$

or:

$$\frac{1}{2}||w||^2$$

(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.

$$L_i(w,b) = max(0, 1 - y_i(f(x_i)) + \epsilon)$$

or:

$$\frac{1}{2}||w||^2 + \epsilon$$

Problem 3

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

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

This algorithm predicts an output based on the closest k matches in the dataset. k is a hyperparameter as it is a adjusted to yield the best results.

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

(c) [easy] 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 predicting on the dataset with k=1, you will have 0 error. This is because the nearest match in the dataset will be that exact value. However, this will not hold in future data or when k is greater than 1.

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 $\mathcal{Y}$?

$\mathbb{D}$ is n different points with one input and output.
$\mathcal{X}$ are the inputs while
$\mathcal{Y}$ are the outputs

(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.

$$b_0 = \bar{y} - b_1\bar{x}$$

$$b_1 = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_1^n (x_i - \bar{x})^2}$$

Setting $b_1 = 0$ yields the line:

$$b_0 = \bar{y} - 0\bar{x} = \bar{y} \rightarrow y = \bar{y} + (0)x$$

This passes through $< \bar{y}, \bar{x} >$

(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 $\bar{y}$.

$$e = y_i - \bar{y}$$
$$e^{bar} = 0 \rightarrow mean(y_i - \bar{y}) = 0 \rightarrow mean(y_i) = \bar{y}$$

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

The average can be defined as:

$$e^{bar} = \frac{1}{n} * \sum_{i}^{n} e$$

$$\sum_{i}^{n} e = 0 \rightarrow e^{bar} = 0$$

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

RMSE clearly outlines the bounds defining the variance. This is more clear than R^2 which can be high but with lots of variance remaining.

(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 can be true if the null model has less variance than the linear model. This is possible when there are a few outliers that greatly shift the line.

(g) [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).

$$\mathbb{D} = \sum_{i}^{n} w_i(y_i - b_0 - b_1 x_i)^2$$

Then take the partial derivative of b_0 and b_1 respectively:

$$\frac{d}{d(b_0)} = -2 * \sum_{i}^{n} w_i(y_i - b_0 - b_1 x_i)$$

$$\frac{d}{d(b_1)} = -2 * \sum_{i}^{n} w_i(y_i - b_0 - b_1 x_i) * (x_i)$$

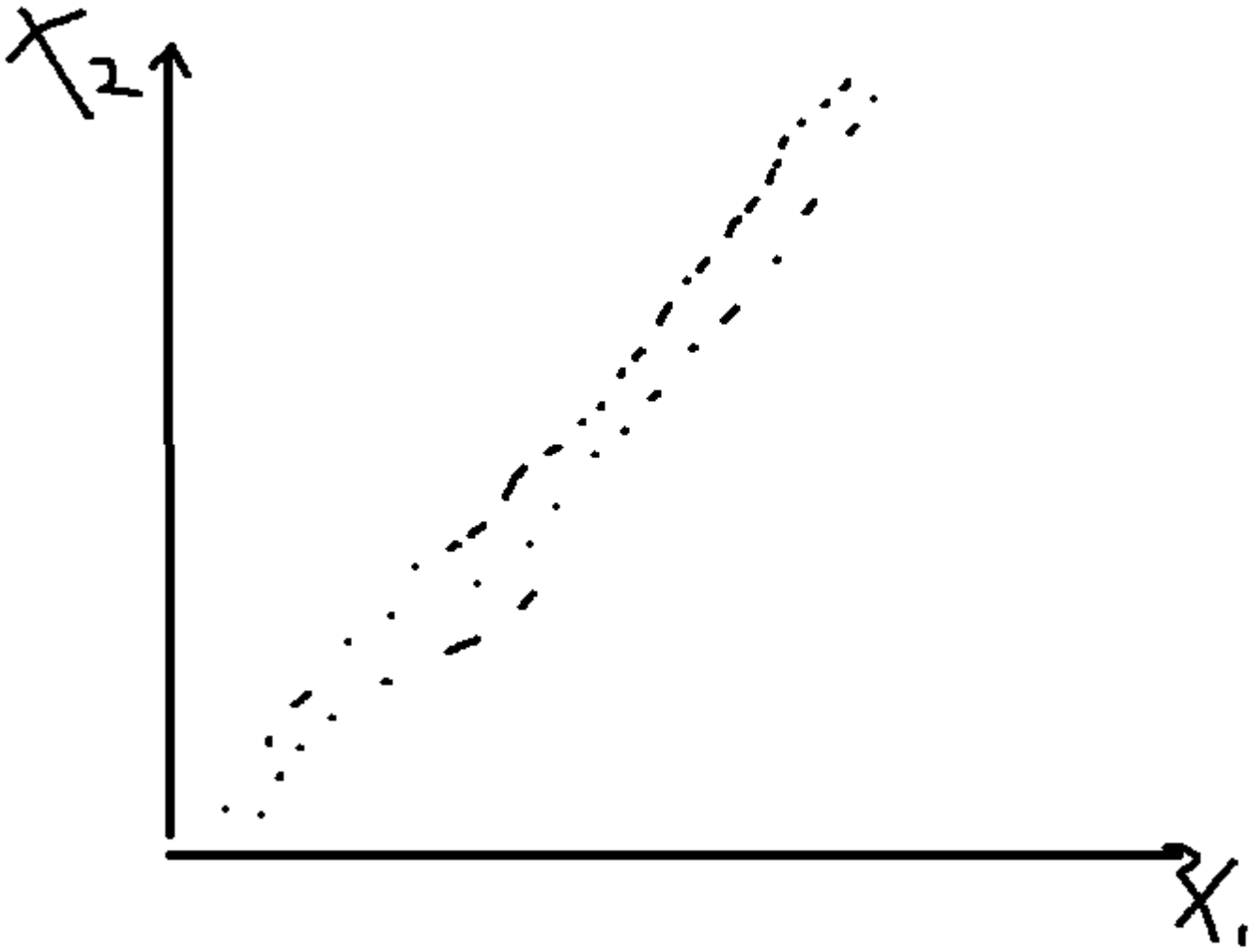
(h) [harder] Interpret the ugly sums in the b_0 and b_1 you derived above and compare them to the b_0 and b_1 estimates in OLS. Does it make sense each term should be altered in this matter given your goal in the weighted least squares?

(i) [E.C.] 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 $g(\text{low}) \leq g(\text{high})$. Write about an algorithm $\mathcal{A}$ that can solve this problem.

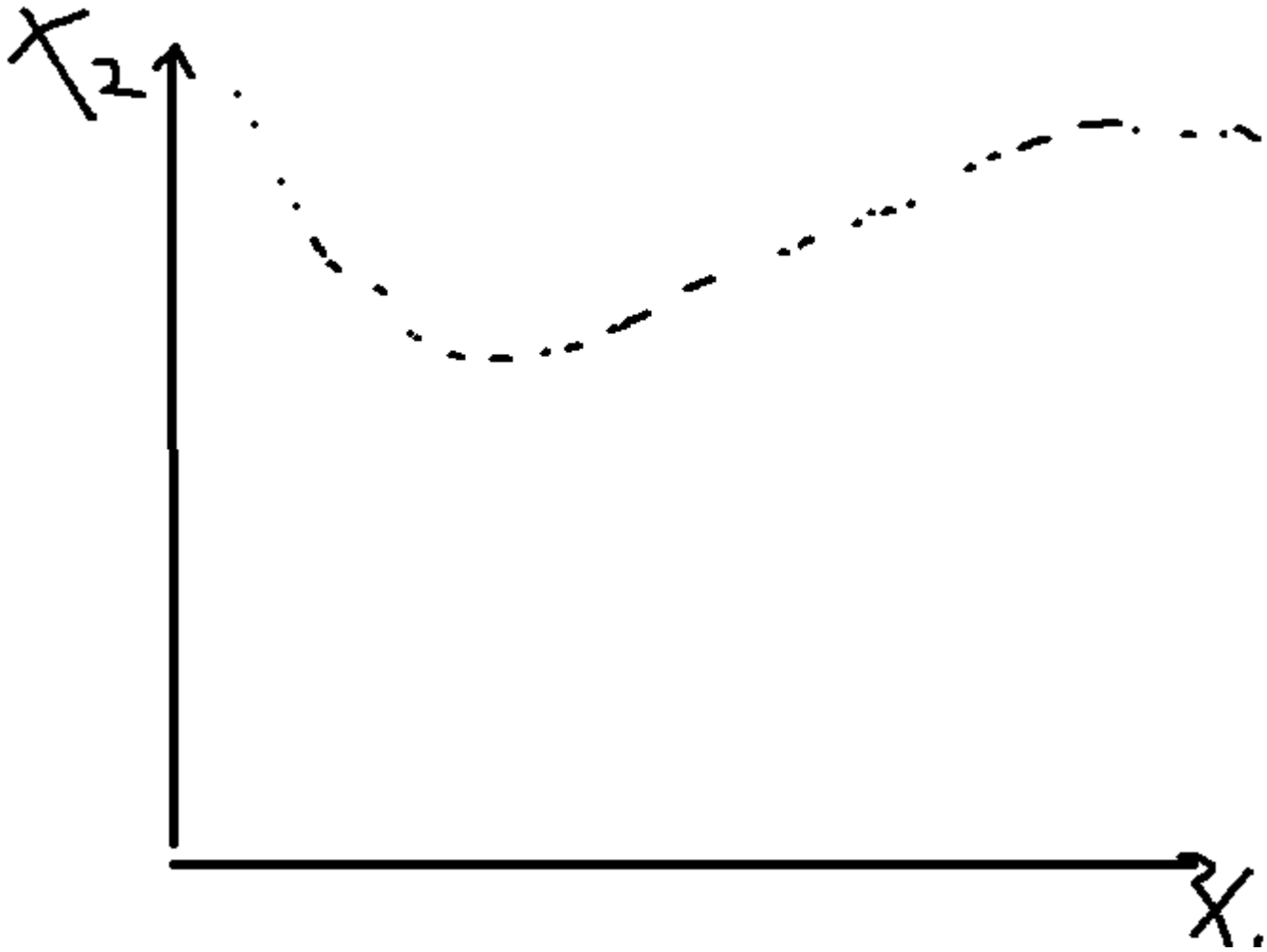
Problem 5

These are questions about association and correlation.

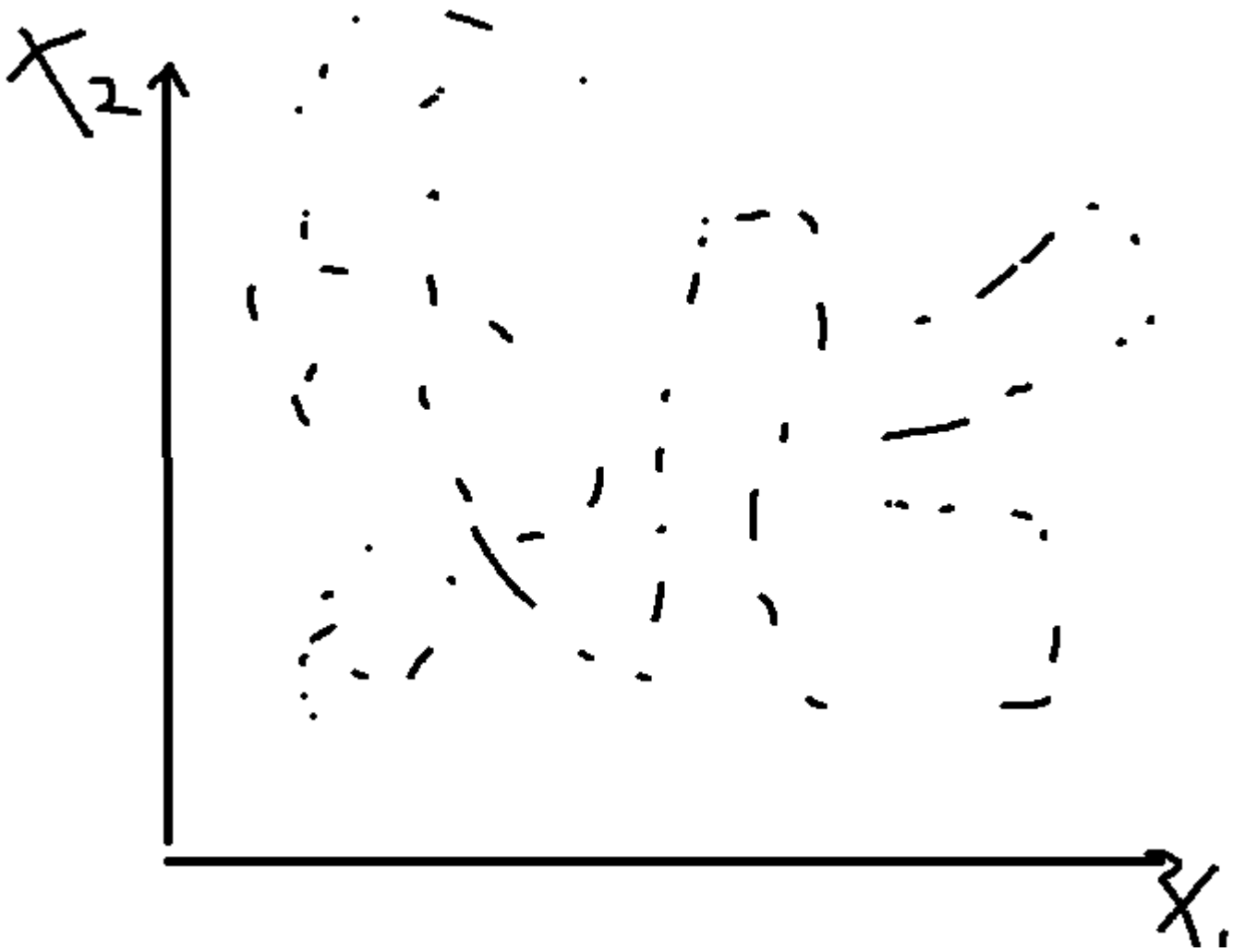
(a) [easy] Give an example of two variables that are both correlated and associated by drawing a plot.



(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.

No. Everything that is correlated is associated. Correlation is just one type of (linear) association.

Problem 6

These are questions about multivariate linear model fitting using the least squares algorithm.

(a) [difficult] Derive $\frac{\partial}{\partial \boldsymbol{c}} \left[\boldsymbol{c}^\top A \boldsymbol{c}\right]$ where $\boldsymbol{c} \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$ but *not* symmetric. Get as far as you can.

(b) [easy] Given matrix $X \in \mathbb{R}^{n \times (p+1)}$, full rank and first column consisting of the $\mathbf{1}_n$ vector, rederive the least squares solution $\boldsymbol{b}$ (the vector of coefficients in the linear model shipped in the prediction function g). No need to rederive the facts about vector derivatives.

(c) [harder] Consider the case where $p = 1$. Show that the solution for $\boldsymbol{b}$ you just derived in (b) is the same solution that we proved for simple regression. That is, the first element of $\boldsymbol{b}$ is the same as $b_0 = \bar{y} - r\frac{s_y}{s_x}\bar{x}$ and the second element of $\boldsymbol{b}$ is $b_1 = r\frac{s_y}{s_x}$.

Where p=1, the x matrix will be a 1xn column vector like before.

(d) [easy] If X is rank deficient, how can you solve for $\boldsymbol{b}$? Explain in English.

The first step would be to identify and remove the linearly dependant columns. Then proceed with full rank.

(e) [difficult] Prove $\text{rank}\,[X] = \text{rank}\left[X^\top X\right]$.

columns = nullity + rank
Assume x has nullity of 0. nullity $X^\top X = 0$

$$\text{rank}\,[X] = \text{rank}\left[X^\top X\right]$$

(f) [harder] [MA] If $p = 1$, prove $r^2 = R^2$ i.e. the linear correlation is the same as proportion of sample variance explained in a least squares linear model.

(g) [harder] Prove that $g([1\ \bar{x}_1\ \bar{x}_2\ \ldots\ \bar{x}_p]) = \bar{y}$ in OLS.

$$e = y_i - \bar{y}$$
$$e^{bar} = 0 \rightarrow mean(y_i - \bar{y}) = 0 \rightarrow mean(y_i) = \bar{y}$$

(h) [harder] Prove that $\bar{e} = 0$ in OLS.

The average can be defined as:

$$e^{bar} = \frac{1}{n} * \sum_i^n e$$

$$\sum_i^n e = 0 \rightarrow e^{bar} = 0$$

(i) [difficult] If you model $\boldsymbol{y}$ with one categorical nominal variable that has levels A, B, C, prove that the OLS estimates look like $\bar{y}_A$ if $x = A$, $\bar{y}_B$ if $x = B$ and $\bar{y}_C$ if $x = C$. You can choose to use an intercept or not. Likely without is easier.

$$(\bar{y}|x = A) = A$$
$$(\bar{y}|x = B) = B$$
$$(\bar{y}|x = C) = C$$

(j) [harder] [MA] Prove that the OLS model always has $R^2 \in [0, 1]$.