MATH 342W / 650.4 Spring 2024 Homework #2

Mohammed Hasan

Monday 26th 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's approach resembles more of a single, all-encompassing model h^* that relies heavily on a general principle, aiming to minimize error ϵ through a complex model with many parameters θ. On the other hand, The fox's approach involves employing a variety of models $\{g_1, g_2, ..., g_m\}$ that capture different aspects of the data D and phenomenon y.

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

Harry Truman liked hedgehogs more because he wanted a "one-handed economist" meaning he wanted one straightforward answers rather than multiple conditional answers which foxes gave. Yes, many people think this way are considered part of a type A culture which contain industries like television, business, and politics.

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

"the more facts hedgehogs have at their command, the more opportunities they have to permute and manipulate them in ways that confirm their biases." This implies that when individuals with a hedgehog-like focus acquire more education and knowledge, their predictions can become less accurate due to an overconfidence in their own perspectives. This overconfidence leads them to manipulate information in ways that confirm their existing biases.

(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.

Probabilistic classifiers show a range of possible outcomes rather than a single prediction which vanilla classifiers would show.

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

PECOTA is most similar to KNN model, the K-nearest neighbors.

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

No. baseball performance as a function of age does not follow a linear model because age deterioration causes more of a bell curve with the higher points being the prime of a humans lifetime.

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

Baseball scouts can do better than a prediction system like PECOTA by looking into aspects of a player's performance and potential that are difficult to quantify through statistical analysis. PECOTA and other statistical models can analyze past performance data and compare players to other recorded players, scouts can assess intangible factors such as a player's work ethic, mental toughness, adaptability, and potential for growth beyond what their current statistics might suggest. Silver also experiences this first hand when he confronted Pedroia. "Pedroia might have let the scouting reports go to his head and never have made the big leagues." which shows that the player had other factors such as mental toughness which allowed him to excel.

(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?

"But someone will come along and take advantage of Pitch f/x data in a smart way, or will figure out how to fuse quantitative and qualitative evaluations of player performance." I believe this means that no one at the time had the abilities to evaluate players based on their quantitative and qualitative data. Which is why Pitch f/x wasnt being taken advantage of at the time.

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?

For the SVM algorithm, H consists of linear functions of the form $h(x) = w \cdot x + b$, where w is the weight vector, b is the bias term, and x is the input feature vector. The entire point of the SVM is to find the optimal separating hyperplane that maximizes

the margin between two classes in the feature space. On the other hand, for the Perceptron Learning Algorithm, the hypothesis set is also composed of linear functions $h(x) = w \cdot x + b$. The difference being w and b. The Perceptron Learning Algorithm updates its weights based on misclassified examples, without considering the margin between the classes. Meanwhile, the SVM explicitly maximizes this margin, which often results in a better generalization to unseen data.

(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.

For a linearly separable case, the SVM's goal is to find a hyperplane that separates the classes with the maximum margin while correctly also satisying the dataset. The cost function's minimization gives us the SVM line that is based on maximizing the margin between the two classes, which is inversely proportional to the norm of the weight vector w. In which case the problem is:

$$\min \{w, b\} \frac{1}{2}\|w\|^2$$

Where $y_i \in \{-1, 1\}$ are the class labels, and x_i are the feature vectors of the training examples.

(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"?

a predefined, k, is used to consider the nearest neighbors to consider for making a prediction. This k is a "hyperparameter," it is a parameter that is set before the learning process and directly influences the performance of the model.

(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).

The algorithm will select the closest point in the dataset to the query point and predict its value, this means there should be zero error because each point is its own neighbor as well. No, K=1 KNN model is is to constrained considering the dataset.

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 linear model p = 1 is a regular linear regression model that showcases a linear relationship between X and Y. This is when X is an independent prediction and Y is a dependant response.

(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.

The equation for OLS is:

$$y = \beta_0 + \beta_1 x$$

where β_0 is the y-intercept, and β_1 is the slope of the line.

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

and to prove that $< \bar{x}, \bar{y} >$ is on this line, we can just sub in $\bar{x}$ for 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 $\bar{y}$.

The average prediction $\hat{y}_i = g(x_i)$ for $x_i \in D$ is $\bar{y}$, we would use the simple linear regression model discussed earlier: $g(x) = \hat{\beta}_0 + \hat{\beta}_1 x$, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the OLS estimates for the intercept and slope. The average prediction over the dataset D is shown as:

$$\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

we can just isolate $\hat{\beta}_0$ so that:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Substituting $\hat{\beta}_0$ into our average prediction equation.

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

The residual for each observation i is defined as $e_i = y_i - \hat{y}_i$, where $\hat{y}_i$ is the predicted value from our OLS model. The average residual over the dataset D is:

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

By definition of the OLS, the sum and the average of residuals is zero, as the OLS minimizes the sum of squared residuals, $SSR = \sum_{i=1}^{n} e_i^2$.

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

The RMSE and the coefficient of determination (R^2) both measure model performance, but RMSE is a better indicator of predictive performance because it directly quantifies the average error made by the model in the units of the outcome variable. RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

While R^2 measures the proportion of the variance in the dependent variable that is predictable from the independent variables, it does not give information that the RMSE does.

(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) [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).

Multi-variable linear regression used when you want $\vec{\hat{y}}$ to come close to $\vec{y}$. To do this we must use an intelligent weight $\vec{w}$. In the WLS model, each squared term in the loss function is weighted by a factor, which indicates how much weight or importance is given to that observation. $\vec{\hat{y}}$ are the predictions and $\vec{y}$ are the responses

$$\hat{y} = \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \text{X } \vec{w}$$

$\vec{b} = \text{argminSSE}$

$\frac{\partial}{\partial \vec{w}}$ [SSE] = $\tilde{\mathbf{0}}_{p+1}$

$=> \frac{\partial}{\partial \vec{w}}[\hat{y}^T\hat{y} - 2w^T X^T \vec{y} + \vec{w}^T X^T X \vec{w}]$

$=> \mathbf{0}_p \vec{+} 1$ - $2\frac{\partial}{\partial \vec{w}}[\vec{w}^T X^T \hat{y}$ + $\frac{\partial}{\partial \vec{w}}[w^T \vec{x^T} X \vec{w}]$

where $\mathbf{0}_p \vec{+} 1$ satisfies Rule 0, - $2\frac{\partial}{\partial \vec{w}}[\vec{w}^T X^T \hat{y}$ satisfies Rule 1, $\frac{\partial}{\partial \vec{w}}[w^T \vec{x^T} X \vec{w}]$ satisfies rule 3 and the unity satisfies rule 2.

(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?

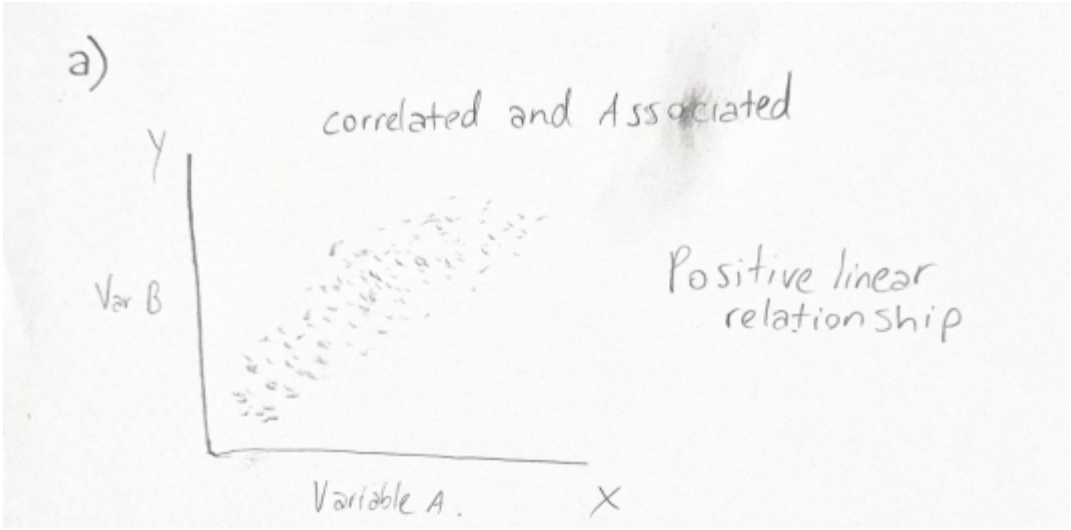
In the standard OLS estimates, b_0 and b_1 are calculated without weighting the errors, which means all observations are ecaluated at an equal weight. This means that every observation contributes equally to the direction and position of the best-fit line, regardless of the reliability or importance of that observation. On the other hand, the WLS estimates adjust each term by the corresponding weight, w_i, reflecting the proportionate importance of each observation.

(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.

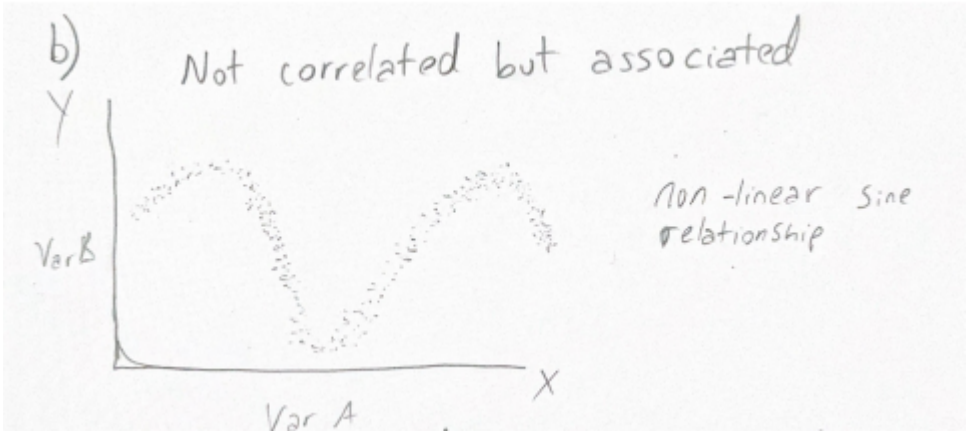
Problem 5

These are questions about association and correlation.

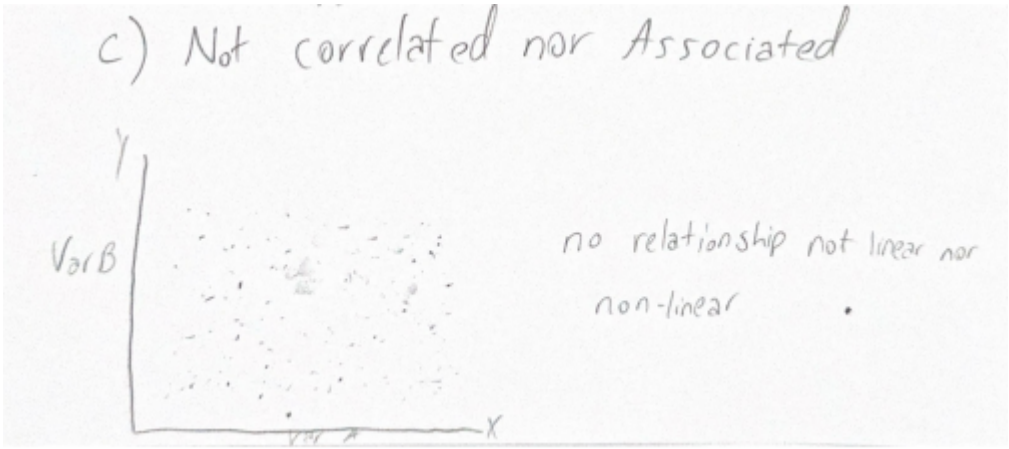
(a) [easy] Give an example of two variables that are both correlated and associated by drawing a plot.



(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, I believe correlation is a type of association and can't even visibly shown through a graph the possibility of "correlated but not associated"

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.

The least squares solution for the coefficients b in a linear model is derived from minimizing the sum of squared residuals. we can use the equation:

$$b = (X^\top X)^{-1} X^\top y$$

(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 $b_1 = r\frac{s_y}{s_x}$.

When $p = 1$, the matrix X consists of two columns: a column of ones (for the intercept) and a column for the single predictor variable. The solution derived in (b) simplifies to the formulas for b_0 and b_1 in simple linear regression:

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

$$b_1 = \frac{r_{xy} s_y}{s_x}$$

where r_{xy} is the correlation between x and y, s_y and s_x are the standard deviations of y and x respectively, and $\bar{y}$ and $\bar{x}$ are the means of y and x.

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

When X is rank deficient, the matrix $X^\top X$ is not invertible, which complicates finding a direct solution using the normal equation. we can add a term to the loss function, making $X^\top X$ invertible by ensuring it is not rank deficient.

(e) [difficult] Prove rank $[X]$ = 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.

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

(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.

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