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.

Hedgehogs pull their beliefs from a few big ideas while foxes pull their beliefs from a bunch of smaller ideas, and approaching problems through a multitude of ways.

(b) [easy] Why did Harry Truman like hedgehogs? Are there a lot of people that think this way? bc they can always give a clear, straight forward answer whether or not it's right.
All of media would prefer hedgehogs, it's more entertaining to watc

(c) [difficult] Why is it that the more education one acquires, the less accurate one's predictions become? Bc of information overload. When too much info is gathered it's hard to distinguish between valuable and irrelevant info, leading to confusion and misjudgement

(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.

Bc probalistic classifiers also identify the uncertainty of the prediction, providing insights into the confidence level of the decision. The additional info we gain, allows for more specific desicion making.

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

The nearest neighbor Algorithm

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

It is not, there is an apex of good performance at 27, you are building up to that apex from when you enter the league until 27, and once you peak, your performance starts to decline as you get older

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

The Scouts have easier access to data then PECOTA does. They don't need to derive anything from a model which may or may not be true, they can just find that information w/ a stop watch or radar gun

(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?

The information that Pitch fx uses was traditionally seen a qualit
stat used by scouts, but now it's a quantitative stat use
by models and it was too tough to merge it into one the
stat

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?

$\mathcal{H} = \{h : \vec{w} \cdot \vec{x} - b \geq 0 : \vec{w} \in \mathbb{R}^p, b \in \mathbb{R}$

(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.

$y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1$, the margin is $\frac{2}{\|\vec{w}\|}$ so to minimize it we want to make $\|\vec{w}\|$ as big as possible so we could square it so it can be $\frac{2}{\|\vec{w}\|^2}$

(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.

Problem 3

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

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

If there's a data point and we don't know what group it belongs to, we can find the distance on an x,y plane between the unknown object and it's k closest neighbors, whichever group has more neighbors of the unknown point, we assign it to that group. k is a hyperparameter because we're gonna decide how many neighbors we're looking for before the problem

(b) [difficult] [MA] Assuming $\mathcal{A} = \text{KNN}$; 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).

bc if we're looking at just one neighbor, then it will just draw us to the closest one. we won't have to choose between 2 neighbors that aren't the same type. This is not a good estimation bc if our current closest neighbor is red, then our element is red, but if new data comes in and our next closest element is blue, our element will now become blue

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 $< \bar{x}, \bar{y} >$ is on this line. Use the formulas we derived in class.

$$\text{slope} = m = \frac{(x_1 - \bar{x})(y_1 - \bar{y})}{(x_1 - \bar{x})^2} \qquad b = \bar{y} - m\bar{x} \longrightarrow y = mx + b \longrightarrow \bar{y} = m\bar{x} + b$$

$$\longrightarrow \bar{y} = m\bar{x} + (\bar{y} - m\bar{x}) = \boxed{\bar{y} = \bar{y}}$$

(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 $\mathbb{D}$.

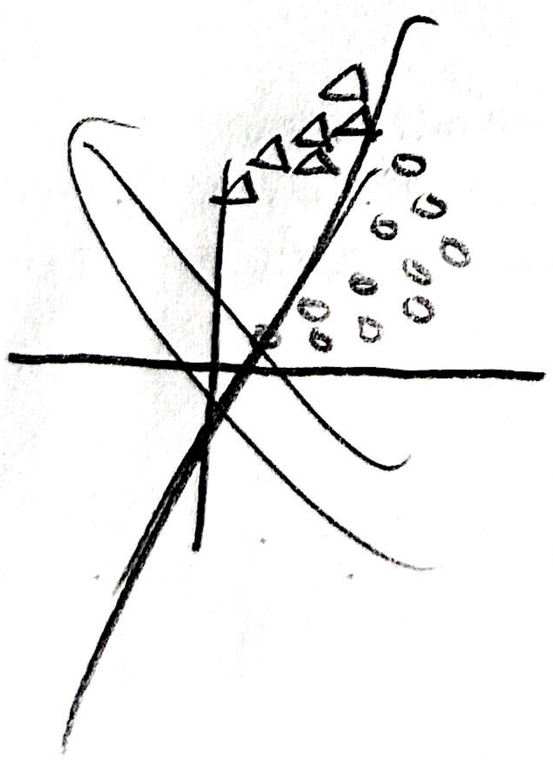
$\hat{y}_i = mx_i + b \quad e_i = y_i - \hat{y}_i \longrightarrow \bar{e} = \bar{y} - m\bar{x} - b \longrightarrow$

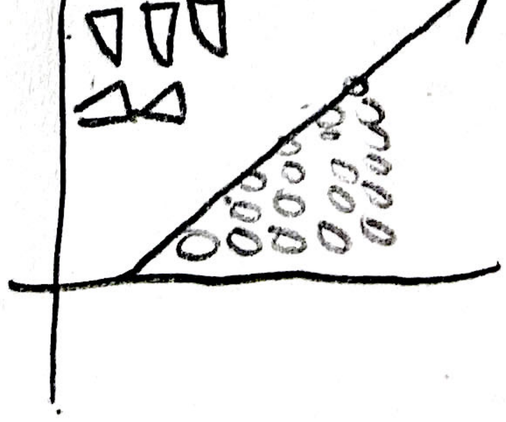
$\bar{e} = b + m\bar{x} - m\bar{x} - b \longrightarrow \bar{e} = 0$

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

bc RMSE directly measures the average prediction error, while R^2 shows the variance ratio which doesn't always translate to predictive accuracy

(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 $R^2 < 0$.





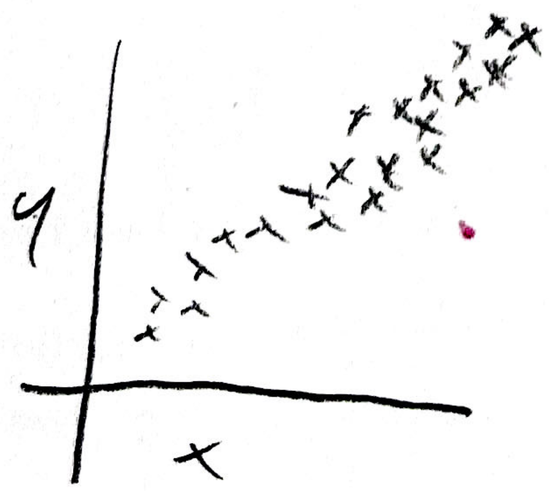
The 1D would be the same as it is for when $R^2 > 0$

(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.

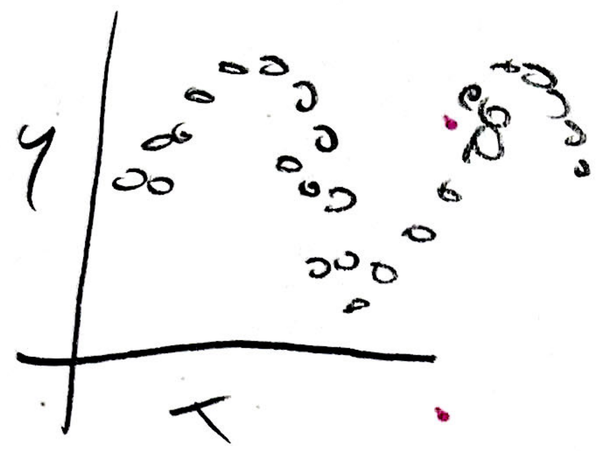
Problem 5

These are questions about association and correlation.

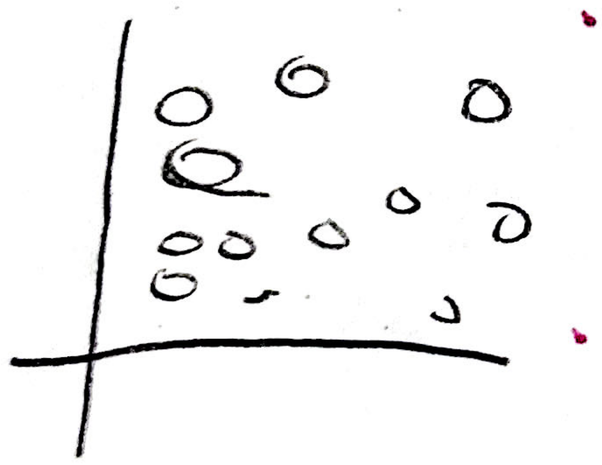
(a) [easy] Give an example of two variables that are both correlated and associated by drawing a plot.



(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, when 2 variables are correlated they have a dependent relationship of each other, meaning there's some sort of relationship between them, which by definition is what association is

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.

$||y - X\vec{b}||^2$ → when we set the derivative = 0 (w/ respect to $\vec{b}$)

we set $X^T X\vec{b} = X^T y$ → Since X is full rank and $X^T X$ is invertible

we get $\boxed{\vec{b} = (X^T X)^{-1} X^T y}$

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

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

You can reduce the feature set to eliminate dependencies or use different algorithms to approximate a solution

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

Idk how to prove but one of the properties of transposed matrices, is when you multiply a matrix by it's transpose, the rank is the same as the original matrix, i.e. $\text{rank}(x) = \text{rank}(x^\top x)$

(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.

$\hat{b} = (X^T X)^{-1} X^T y$ ⟹ $g(x) = x^T \hat{b}$ and when you substitute the $+ bars$ into g, when $\bar{x}_i > 0$ it doesn't contribute anything since their mean is 0. only $\hat{b}_0$ (intercept) is left, which is $\bar{y}$

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

$e = y - X\hat{b}$, the first col in X is a 1's column for the intercept $\hat{b}_0$. By the equation $X^T X \hat{b} = X^T y$ $\hat{b}_0$ is chosen that the sum of e_i's $= 0$

$\sum e_i = \sum y_i - n\hat{b}_0 = 0$ so $n\hat{b}_0 = \sum y_i$ by the def of $\hat{b}_0$

$\therefore \bar{e} = \frac{1}{n}\sum e_i = 0 \rightarrow \bar{e} = 0$

(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.

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