MATH 390.4 Spring 2020 Homework #2

Tziporah Horowitz

Tuesday 25th February, 2020

1 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.

Foxes are better at prediction than hedgehogs because they improve as more information becomes available. Hedgehogs only believe in one hypothesis set, $\mathcal{H}$, while foxes will adjust their hypotheses based on the x's observed.

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

Truman believed that foxes were too uncertain with their answers. Like many others, Truman liked bold predictions.

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

The more information a hedgehog acquires, the faster he can attribute it toward a prediction.

(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 are better than vanilla classifiers because they give an almost definite prediction of the percentage of correct outcomes, rather than an objective prediction that has a higher likelihood of being wrong.

2 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?

The SVM's hypothesis set differs from the perceptron's hypothesis set by a factor of b:

$$\mathcal{H} = \left\{ \mathbb{1}_{\vec{w}\cdot\vec{x}+b\geq 0} : \vec{w} \in \mathbb{R}^p, b \in \mathbb{R} \right\}$$

(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.

$$\begin{aligned}
l_a: \quad \forall y_i = 1, \qquad & \vec{w}\cdot\vec{x_i} - (b+1) \geq 0 \\
& \vec{w}\cdot\vec{x_i} - b \geq 1 \\
& y_i(\vec{w}\cdot\vec{x_i} - b) \geq y_i = 1 \\
l_b: \quad \forall y_i = -1, \quad & \vec{w}\cdot\vec{x_i} - (b-1) \leq 0 \\
& \vec{w}\cdot\vec{x_i} - b \leq -1 \\
& -y_i(\vec{w}\cdot\vec{x_i} - b) \leq y_i = -1 \\
& y_i(\vec{w}\cdot\vec{x_i} - b) \geq 1
\end{aligned}$$

$$l: \quad \text{maximize } \frac{2}{||\vec{w}||} \text{ such that } y_i(\vec{w}\cdot\vec{x_i} - b) \geq 1$$

(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.

$$Cost = \underset{\vec{w}\in\mathbb{R}^p, b\in\mathbb{R}}{\arg\min} \left\{ \frac{1}{n} \max\left\{0, 1 - y_i(\vec{w}\cdot\vec{x_i} - b)\right\} + \lambda||\vec{w}||^2 \right\}$$

3 The k Nearest Neighbors (KNN) Algorithm

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

For $\vec{x^*}$, find $x_{i_{(1)}}, \ldots, x_{i_{(k)}}$ where the distances $d(\vec{x^*}, \vec{x_{i_{(k)}}})$ are the k smallest and let $\hat{y} = mode[y_{i_{(1)}}, \ldots, y_{i_{(k)}}]$. k is a hyperparameter.

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

$\mathcal{H}$ is defined by a distance metric and the hyperparameter, k.

(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 error will be 0 because $\vec{x^*}$ will be mapped to itself. However, it is not a good estimate for future error because a datapoint that was not previously observed will not be able to map to itself.

4 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 $\mathcal{Y}$?

$$\mathbb{D} = \left\{ \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \right\}$$

$$\mathcal{X} = \mathbb{R}^{n \times (p+1)}$$
$$\mathcal{Y} = \mathbb{R}^n$$

(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.

$$\begin{aligned} g(x) &= b_0 + b_1 x \\ b_1 &= r\frac{s_y}{s_x} \\ b_0 &= \bar{y} - b_1 \bar{x} \\ &= \bar{y} - \bar{x} r \frac{s_y}{s_x} \\ g(\bar{x}) &= \bar{y} - \bar{x} r \frac{s_y}{s_x} + \bar{x} r \frac{s_y}{s_x} = \bar{y} \end{aligned}$$

(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 $\bar{y}$.

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i &= \frac{1}{n}\sum_{i=1}^{n}(b_0 + b_1 x_i) \\ &= b_0 + \frac{b_1}{n}\sum_{i=1}^{n} x_i \\ &= b_0 + b_1\bar{x} \\ &= \bar{y}\end{aligned}$$

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

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}e_i &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \\ &= \bar{y} - \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i \\ &= \bar{y} - \frac{1}{n}\sum_{i=1}^{n}(b_0 + b_1 x_i) \\ &= \bar{y} - \bar{y} \\ &= 0\end{aligned}$$

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

RMSE tells you how far your model is from the true values, while R^2 is a unitless percentage.

(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 $R^2 < 0$.

When $g(x)$ does not fit the data and therefore predicts worse than $g_0(x)$:

$$\mathbb{D} = \left\{ \begin{bmatrix} 1 & 2 \\ 1 & 5 \\ 1 & 8 \end{bmatrix} \quad \begin{bmatrix} 4 \\ 10 \\ 15 \end{bmatrix} \right\} \qquad\qquad g(x) = -\frac{x}{4} - 3$$

(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).

$$
\begin{aligned}
\text{Weighted SSE} &= \sum_{i=1}^{n} w_i \big(y_i - (b_0 + b_1 x_i)\big)^2 \\
&= \sum_{i=1}^{n} w_i \big(y_i^2 - 2y_i(b_0 + b_1 x_i) + (b_0 + b_1 x_i)^2\big) \\
&= \sum_{i=1}^{n} w_i (y_i^2 - 2y_i b_0 - 2y_i b_1 x_i + b_0^2 + 2b_0 b_1 x_i + b_1^2 x_i^2) \\
&= \Sigma w_i y_i^2 - 2b_0 \Sigma w_i y_i - 2b_1 \Sigma w_i y_i x_i + b_0^2 \Sigma w_i + 2b_0 b_1 \Sigma w_i x_i + b_1^2 \Sigma w_i x_i^2 \\
\frac{\partial}{\partial b_0} &= -2 \sum_{i=1}^{n} w_i y_i + 2b_0 \sum_{i=1}^{n} w_i + 2b_1 \sum_{i=1}^{n} w_i x_i \overset{\text{set}}{=} 0 \\
b_0 \sum_{i=1}^{n} w_i &= \sum_{i=1}^{n} w_i y_i - b_1 \sum_{i=1}^{n} w_i x_i \\
b_0 &= \frac{\sum_{i=1}^{n} w_i y_i - b_1 \sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \\
\frac{\partial}{\partial b_1} &= -2 \sum_{i=1}^{n} w_i y_i x_i + 2b_0 \sum_{i=1}^{n} w_i x_i + 2b_1 \sum_{i=1}^{n} w_i x_i^2 \overset{\text{set}}{=} 0 \\
b_1 &= \frac{\sum_{i=1}^{n} w_i y_i x_i - b_0 \sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i x_i^2}
\end{aligned}
$$

$$
\begin{aligned}
b_1 &= \frac{\sum_{i=1}^{n} w_i y_i x_i}{\sum_{i=1}^{n} w_i x_i^2} - \frac{\sum_{i=1}^{n} w_i y_i - b_1 \sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \cdot \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i x_i^2} \\
&= \frac{\sum_{i=1}^{n} w_i y_i x_i}{\sum_{i=1}^{n} w_i x_i^2} - \frac{\sum_{i=1}^{n} w_i y_i \sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i x_i^2} + \frac{b_1 \big(\sum_{i=1}^{n} w_i x_i\big)^2}{\sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i x_i^2} \\
b_1 - \frac{b_1 \big(\sum_{i=1}^{n} w_i x_i\big)^2}{\sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i x_i^2} &= \frac{\sum_{i=1}^{n} w_i y_i x_i}{\sum_{i=1}^{n} w_i x_i^2} - \frac{\sum_{i=1}^{n} w_i y_i \sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i x_i^2} \\
b_1 &= \left(\frac{\Sigma w_i y_i x_i}{\Sigma w_i x_i^2} - \frac{\Sigma w_i y_i \Sigma w_i x_i}{\Sigma w_i \Sigma w_i x_i^2} \right) \cdot \left(\frac{1}{1 - \frac{(\Sigma w_i x_i)^2}{\Sigma w_i}} \right)
\end{aligned}
$$

(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?

The sums in the previous answer resemble the sums in the OLS estimates, but each sum in the weighted least squares estimates is altered by a factor of w_i. This makes sense because we are scaling the least squares estimates by w_i.

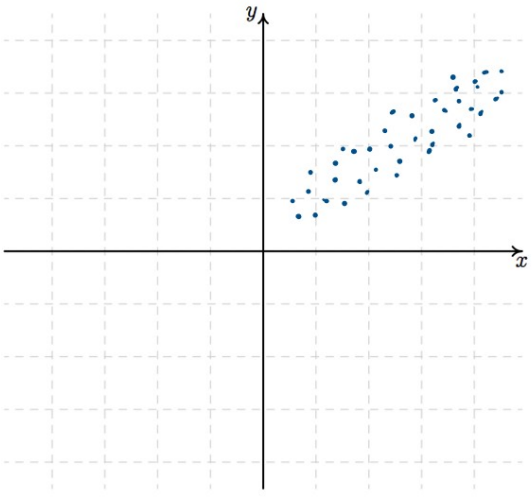
(j) [difficult] [MA] In class we talked about $x_{raw} \in \{\text{red}, \text{green}\}$ and the OLS model was the sample average of the inputted x where $b_0 = \bar{y}_r$ and $b_1 = \bar{y}_g - \bar{y}_r$. Reparameterize $\mathcal{H} = \{w_1 \mathbb{1}_{x_{raw} = \text{ red}} + w_2 \mathbb{1}_{x_{raw} = \text{ green}} \ : \ w_1, w_2 \in \mathbb{R}\}$ 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.

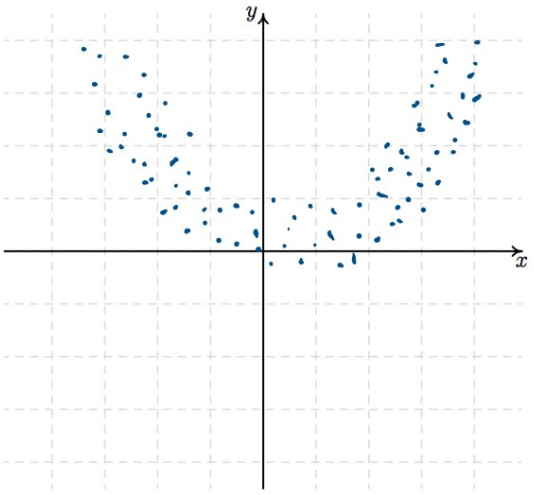
$$g(x) = \begin{cases} 1 & \text{if } b_0 + b_1 x_i \leq \delta \\ 2 & \text{if } b_0 + b_1 x_i > \delta \end{cases}$$

5 Association and Correlation

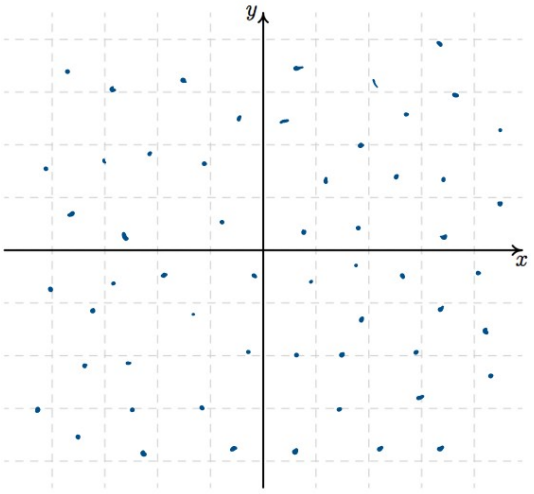
(a) [easy] Give an example of two variables that are both correlated and associated by drawing a plot.



(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. Since correlation is a measure of linear association, two variables must be linearly associated if they are correlated.

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