New York University Tandon School of Engineering
Computer Science and Engineering

CS-UY 4563: Written Homework 1 Solutions

Problem 1: Practice Framing a Supervised Learning Problem (10pts)

A university admissions office wants to predict the success of students based on their application material. They have access to past student records to learn a good algorithm.

(a) (3pts) To formulate this as a supervised learning problem, identify a possible target variable. This should be some variable that measures success in a meaningful way and can be easily collected (in an automated manner) by the university. There is no one correct answer to this problem.

(b) (1pts) Is the target variable continuous or discrete-valued?

Some example answers:

- Students graduating college GPA (continuous valued)
- If the student graduates in 4 years (yes/no discrete valued)
- Salary of students first post-grad job (continuous valued, some schools collect this data)

(c) (3pts) State one possible variable that can act as the predictor for the target variable you chose in part (a).

Some example answers:

- Students graduating high school GPA .
- Students SAT scores.
- Number of honors level classes take in high school.

(d) (3pts) Without looking at data, would a linear model for the data be reasonable? If so, what sign do you expect the slope to be? If note, what might be a better model?

Really anything goes here, as long as they get the slope correct, and say *something* about why linear makes sense or doesn't.

Problem 2: Practice Minimizing a Loss Function (10pts)

Consider a linear model of the form

$$f_\beta(x) = \beta x,$$

which is the same as the linear model we saw in class, but with the intercept forced to zero. Such models are used when we want to force the predicted value $f_\beta(x) = 0$ when $x = 0$. For example, if we are modeling $y =$ output power of a motor vs. $x =$ the input power, we would expect $x = 0 \Rightarrow y = 0$.

(a) (4pts) Given data $(x_1, y_1), \ldots, (x_n, y_n)$, write the equation for a loss function which measures prediction accuracy using the sum-of-squared distances between the predicted values and target values.

Make sure they remember to properly subscript y_i, x_i.

$$L(\beta) = \sum_{i=1}^{n} (y_i - \beta x_i)^2$$

(b) (6pts) Derive an expression for the β that minimizes this loss function. Do you get the same expression that we got for β_1 in the full linear model?

- Set derivative to zero: $\frac{d}{d\beta} L(\beta) = 0$
- Calculate derivative $\frac{d}{d\beta} L(\beta) = \sum_{i=1}^{n} -2x_i \cdot (y_i - \beta x_i)$
- Solve for β: $\beta = (\sum_{i=1}^{n} x_i y_i)/(\sum_{i=1}^{n} x_i^2)$

This expression is **different** than what we got for the full linear model – its close, but you don't subtract means off of x's and y's first.

Problem 3: Building Intuition about Different Loss Functions. (10pts)

Suppose we have data $y_1, \dots, y_n \in \mathbb{R}$ and we want to choose some value $m \in \mathbb{R}$ which is "most representative" of our dataset. This is sometimes called the "central tendency" problem in statistics. A machine learning approach to this problem would measure how representative m is of the data using a loss function.

(a) (4pts) Consider the loss function $L(m) = \sum_{i=1}^n (y_i - m)^2$. Show that $L(m)$ is minimized by setting $m = \bar{y}$, where $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ is the *mean* of our data.

- Set derivative to zero: $\frac{d}{dm}L(m) = \sum_{i=1}^n -2 \cdot (y_i - m) = 0$
- Solve for m: $m = \frac{1}{n}\sum_{i=1}^n y_i$

(b) (4pts) Consider the loss function $L(m) = \max_i |y_i - m|$. What value of m minimizes this loss? Hint: Using derivatives will not help here – try just thinking about the minimization problem directly.

Let $y_{\min} = \min_i y_i$ and let $y_{\max} = \max_i y_i$. We should set $m = (y_{\min} + y_{\max})/2$. I.e. place m exactly between the largest and smallest value in our data set. No other point can achieve a better value on the loss because it will be further than $(y_{\min} + y_{\max})/2$ from either $y_{\min}$ or $y_{\max}$.

(c) **Bonus (5pts):** Consider the loss function $L(m) = \sum_{i=1}^n |y_i - m|$. Show that $L(m)$ is minimized by setting m to the *median* of the data.

There are a number of ways to prove this. One is as follows:

- Sort our numbers from smallest to largest so that $y_1 \leq y_2 \leq \dots, \leq y_m$.
- Split $\sum_{i=1}^n |y_i - m|$ into $\sum_{i=1}^n |y_i - m| = (|y_1 - m| + |y_n - m|) + (|y_2 - m| + |y_{n-1} - m|) + \dots + (|y_{n/2} - m| + |y_{n/2+1} - m|)$.
- Then observe that $(|y_1 - m| + |y_n - m|)$ is minimized at cost $|y_n - y_1|$ by *any m that lies between y_1 and y_n*. Similarly $(|y_2 - m| + |y_{n-1} - m|)$ is minimized by *any m that lies between y_2 and y_{n-1}*, so on and so forth.
- So we can choose the best m by just making sure it lies simultaneously in the intervals $[y_1, y_n], [y_2, y_{n-1}], \dots, [y_{n/2}, y_{n/}$ One such choice is clearly the median.

(d) (2pts) In a few short sentences, discuss when you might prefer each of the three losses above. Is the median typically considered a more "robust" measure of central tendency than the mean? Why?

No real wrong or right answers here. Although they should be able to identify that the median is **more robust.** If you add a large outlier to a data set, the median will barely change, but the mean could move arbitrarily far.

- You might prefer the mean because it's unique and has concrete meaning: i.e. if you look at the mean income of all people in NYC and multiply by the number of people you get the total gross wages of the city.
- You might prefer the second loss function if you're trying to institute some fairness: e.g. if $y_1, \dots, y_n$ are locations of where a group of friends lives, and they want to select a meeting place that minimizes the furthest distance anyone needs to walk to get there.
- You might prefer the median in a setting where you have outliers.

The median is typically

Problem 4: Practice with Non-linear Transformations. (10pts – I forgot to write this)

(**Hint:** Take a look at the example at the end of my Lecture 2 notes)

A medical researcher wants to model, $f(t)$, the concentration of some chemical in the blood over time. She believes the concentration should decay exponentially in that

$$f(t) = z_0 e^{-\alpha t}, \tag{1}$$

for some parameters z_0 and α. To confirm this model, and to estimate the parameters z_0, α, she collects a large number of time-stamped samples (t_i, c_i), $i = 1, \ldots, n$, where c_i is the measured concentration at time t_i. Unfortunately, the model (1) is non-linear, so she can't directly apply the linear regression formula to estimate z_0 and α.

(a) (4pts) Taking logarithms, show that we can transform our training data so that the conjectured relationship between predictor and target variables is in fact linear.

$$\log(f(t)) = \log(z_0) - \alpha t.$$

which is linear in t.

(b) (6pts) Write pseudocode (or actual Python) for how you might estimate z_0 and α using this transformation.

- Let $\tilde{c}_i = \log(c_i)$ for all $i = 1, \ldots, n$.
- Let $\bar{c} = \frac{1}{n}\sum_{i=1}^{n} \tilde{c}_i$ and $\bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i$.
- Compute $\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(\tilde{c}_i - \bar{c})(t_i - \bar{t})$ and $\sigma_{x^2} = \frac{1}{n}\sum_{i=1}^{n}(t_i - \bar{t})^2$.
- Compute $\beta_1 = \sigma_{xy}/\sigma_{x^2}$ and $\beta_0\bar{c} - \beta_1\bar{t}$.
- Set parameters $z_0 = e^{\beta_0}$ and $\alpha = -\beta_1$

It's okay if they just say something like "find β_0 and β_1 for inputs $t_1, \ldots, t_n$ and $\tilde{c}_1, \ldots, \tilde{c}_n$ using the equations from class". Then they would only need to get the first and last step of what I wrote above.