DSC 40A - Homework 2

due Friday, October 17th at 11:59 PM

Homeworks are due to Gradescope by 11:59PM on the due date.

You can use a slip day to extend the deadline by 24 hours; you have four slip days to use in total throughout the quarter.

Homework will be evaluated not only on the correctness of your answers, but on your ability to present your ideas clearly and logically. You should **always explain and justify** your conclusions, using sound reasoning. Your goal should be to convince the reader of your assertions. If a question does not require explanation, it will be explicitly stated.

Homeworks should be written up and turned in by each student individually. You may talk to other students in the class about the problems and discuss solution strategies, but you should not share any written communication and you should not check answers with classmates. You can tell someone how to do a homework problem, but you cannot show them how to do it. **Only handwritten solutions will be accepted (use of tablets is permitted). Do not typeset your homework (using LaTeXor any other software)**.

For each problem you submit, you should **cite your sources** by including a list of names of other students with whom you discussed the problem. Instructors do not need to be cited.

This homework will be graded out of **59 points**. The point value of each problem or sub-problem is indicated by the number of avocados shown.

Problem 1. Reflection and Feedback Form

🥑🥑 Make sure to fill out this Reflection and Feedback Form, linked here, for two points on this homework! This form is primarily for your benefit; research shows that reflecting and summarizing knowledge helps you understand and remember it.

Problem 2. Speed Measurement

A coastal research station records hourly surface current velocities

$$\vec{y}_i = (\vec{y}_i^{(1)}, \vec{y}_i^{(2)}) \in \mathbb{R}^2,$$

measured in meters per second, where $\vec{y}_i^{(1)}$ is the eastward component and $\vec{y}_i^{(2)}$ is the northward component. Oceanographers suspect that during the study period the motion is driven predominantly by a single tidal stream that always points in the direction

$$\vec{d} = \begin{bmatrix} 1 \\ 2 \end{bmatrix},$$

but whose *speed* (magnitude) is unknown. They therefore propose the vector-valued constant model

$$f(t) \ = \ h\,\vec{d}, \qquad h \in \mathbb{R}.$$

Six velocity measurements (in $\mathrm{m\,s^{-1}}$) and their timepoints (in hours from the start of the experiment) are listed below:

$$\{(t_i, \vec{y}_i)\}_{i=1}^{6} = \Big\{(0, (0.60, 1.21)),\ (1.1, (0.70, 1.47)),\ (1.9, (0.50, 0.93)),$$
$$(2.5, (0.65, 1.25)),\ (3.2, (0.55, 1.11)),\ (4.0, (0.72, 1.38))\Big\}.$$

a) 🥑🥑🥑 Identify the predictor and response variables, state the hypothesis function, and write down a formula for the squared loss $L_{\text{sq}}(h)$ in this setting.

b) 🥑🥑🥑🥑🥑 Write the empirical risk function

$$R(h) = \frac{1}{6}\sum_{i=1}^{6} L_{\text{sq}}(h)$$

for the data above and derive the critical point equation

$$h^* = \frac{\sum_{i=1}^{n} \vec{d}^\top \vec{y}_i}{\sum_{i=1}^{n} \|\vec{d}\|^2}.$$

c) 🥑🥑🥑 Verify that the second derivative $\frac{d^2 R}{dh^2}$ is positive and conclude that h^* is the unique global minimizer.

d) 🥑🥑 Evaluate the expressions in part (b) for the six data points and report the numerical value of h^*.

Problem 3. Pop Quiz

Complete the following two concept checks.

a) 🥑🥑🥑🥑 Determine whether each statement is true or false. Explain briefly.

(a) If the training features $x_1, \ldots, x_n$ for a simple linear regression problem have mean zero then the intercept of the optimal linear model will equal zero.

(b) The line of best fit is found by minimizing the empirical risk.

(c) The slope of the line of best fit is always positive.

(d) A simple linear regression model has exactly two parameters.

b) 🥑🥑 Match each term (a)–(e) with its correct definition (1)–(5) (no explanation required):

(a)	Empirical risk	Equations obtained by setting partial derivatives of empirical risk to zero.	(1)
(b)	Intercept	Average value of the loss function over the training data.	(2)
(c)	Square loss	Parameter indicating the amount that the predicted value changes if the feature increases by one unit.	(3)
(d)	Normal equations	Parameter indicating the predicted value when the feature is zero.	(4)
(e)	Slope	A loss function measuring the squared difference between prediction and actual value.	(5)

Problem 4. Potion Brewing

An alchemist measures gold created (in grams) against hours spent brewing philosopher's potions and obtains the following dataset.

Hours spent brewing	2	3	5	7	8	11
Gold created	0.5	0.75	1.2	1.7	2.1	2.9

a) 🥑🥑 Determine the slope and intercept of the simple linear model $y = w_1 x + w_0$ which minimizes mean squared error for the given dataset.

b) 🥑🥑 Predict the amount of gold the alchemist will create after 10 hours of brewing philosopher's potions.

Problem 5. Tour de France

Two bicyclists are training to compete in the Tour de France. One lives in San Diego, US and the other lives in Stuttgart, Germany. Each week, they record their training distances in miles and kilometers, respectively. For each week i, let

$$x_i = \text{Athlete A's distance (in miles)}, \qquad y_i = \text{Athlete B's distance (in kilometers)}.$$

The paired measurements are:

Athlete A (miles)	Athlete B (km)
186.4	300
199.5	320
211.3	340
223.7	360
236.1	380
248.5	400
260.9	420
273.4	440

Your goal is to use simple linear regression to model Athlete B's distance as a linear function of Athlete A's distance:

$$y = w_1 x + w_0.$$

You may also find it convenient to convert Athlete A's distances to kilometers using

$$1 \text{ mile} = 1.60 \text{ km} \quad \Rightarrow \quad x_{\text{km}} = 1.60\, x. \tag{$\star$}$$

a) 🥑🥑🥑🥑🥑 Your friend Zoe suggests two equivalent approaches:

Workflow A: Convert each x_i to $x_{i,\text{km}} = 1.60\, x_i$ and model y on x_{km} to get coefficients: $\widetilde{w}_1$ (in units of *Athlete B km per Athlete A km*), and $\widetilde{w}_0$ (in units of *km*).

Workflow B: Model y on x (in *miles*) to get coefficients: w_1 (in units of *Athlete B km per Athlete A mi*), and w_0 (in units of *km*); then convert w_1 (possibly shifting w_0 as needed) using ($\star$).

Is Zoe correct that both workflows lead to the same fitted line in kilometers? Show your calculations for computing (w_1, w_0) and $(\widetilde{w}_1, \widetilde{w}_0)$.

Note: In both workflows, the units of w_0 are always in km *since they should match the output of the model. In workflow B we do not convert them directly after training the model, but it is possible that they shift by a constant* (see part (b)).

b) 🥑🥑🥑🥑🥑 More generally, consider the simple linear regression fit of a response y on features x is given by $y = w_1 x + w_0$. Now replace x by a linear transformation $z = ax + b$ (with constants $a, b \in \mathbb{R}$, $a \neq 0$) and refit y on z.

Express the new slope and intercept $(\hat{w}_1, \hat{w}_0)$ *in terms of* w_1, w_0, a, b. Prove your formulas (e.g., rewrite the original fitted line using z and match coefficients).

c) 🥑🥑🥑🥑 For any data set $(x_i, y_i)_{i=1}^n$, define $z_i = ax_i + b$. Show that the *minimum* mean squared error (MSE) from regressing y on x equals the minimum MSE from regressing y on z.

d) 🥑🥑 Determine if the following is true or false by either proving the statement or finding a counterexample.

The simple linear regression line always passes through $(\bar{x}, \bar{y})$.

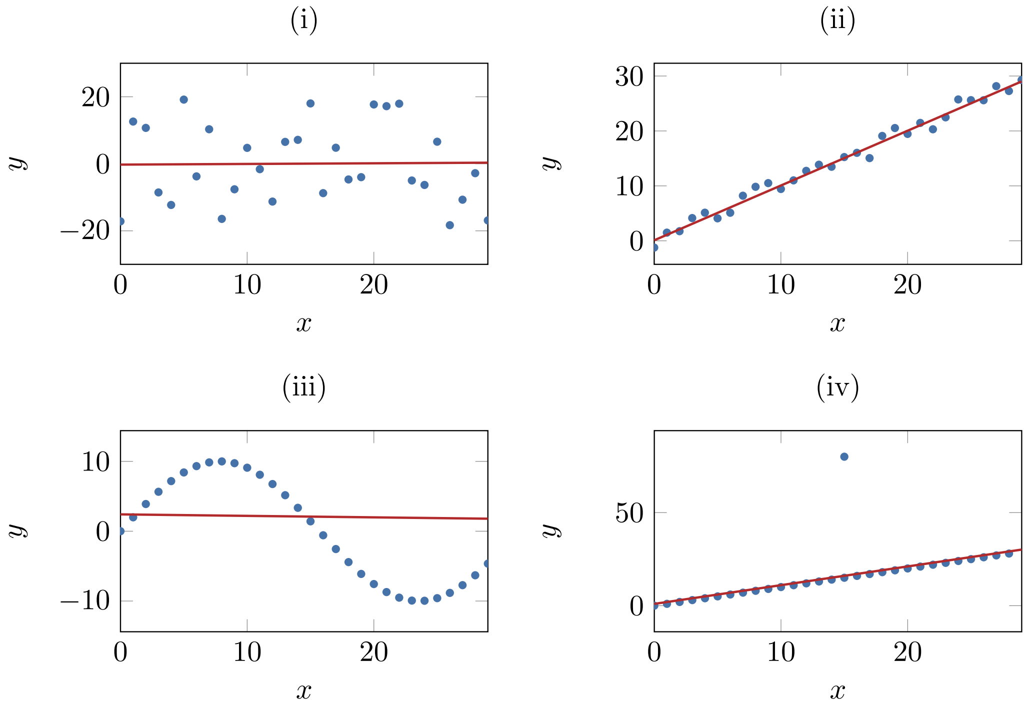
As you attempt this problem, it might be helpful to review Example 2.1.4 in the course notes.

Problem 6. Explain The Plot

The four plots below contain a training dataset $\{(x_i, y_i)\}_{i=1}^{30}$ of thirty points in the xy-plane (blue), along with the line of best fit for the associated simple linear regression problem (red).

a) 🥑🥑 For each plot below, determine whether the simple linear model is appropriate. Write a short explanation of your reasoning.

b) 🥑🥑🥑 For each plot, consider the value $R(w_0^*, w_1^*)$ of the empirical risk associated to the optimal parameters for the simple linear model and given training dataset. Rank the plots from least to greatest value of $R(w_0^*, w_1^*)$.



Problem 7. Does more data help (for linear regression)?

In this problem we will generate a dataset and implement simple linear regression.

The question is provided at this supplementary Jupyter Notebook. The code that you write in that notebook will **not** be graded and you do not need to submit the code on Gradescope. You **do** need to add the plots you generate in the notebook below and answer the two questions.

a) 🥑🥑🥑🥑🥑 Add the plots you generated for fixed δ and varying n.

b) 🥑🥑🥑🥑 Add the plots you generated for fixed n and varying δ.

c) 🥑🥑 How does adding more points to the dataset improve the fit of the model?

d) 🥑🥑 How does increasing noise impact the model's accuracy?