DSC 40A - Homework 02

Due: Friday, January 24, 2020

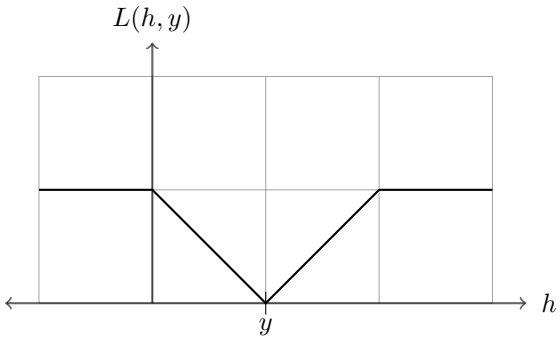
Write your solutions to the following problems by either typing them up or handwriting them on another piece of paper. Unless otherwise noted by the problem's instructions, show your work or provide some justification for your answer. Homeworks are due via Gradescope on Friday afternoon at 5:00 p.m.

Problem 1.

In this problem, consider the loss function

$$L(h, y) = \begin{cases} 1, & |y - h| > 1 \\ |y - h|, & |y - h| \leq 1 \end{cases}.$$

a) Consider y to be a fixed number. Plot $L(h, y)$ as a function of h.



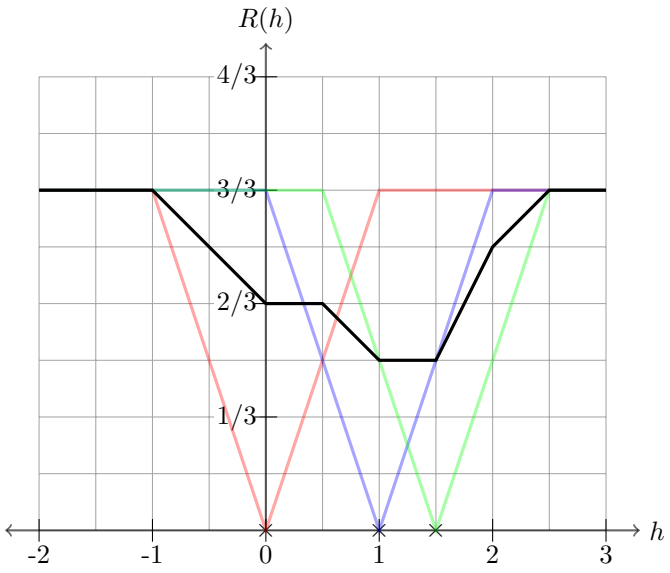
b) Suppose that we have the following data:

$$y_1 = 0$$
$$y_2 = 1$$
$$y_3 = 1.5$$

Plot the empirical risk

$$R(h) = \frac{1}{n} \sum_{i=1}^{n} L(h, y)$$

on the domain $[-2, 3]$. It might help to use the grid below; note that the vertical axis tick marks occur in increments of $1/3$ while the horizontal axis tick marks are in increments of 1.



Solution: The loss function, $L(h, y)$, is a piecewise linear function, and so the risk, $R(h)$, is a constant $(1/3)$ times sum of piecewise linear functions. Recall that the slope of the sum of piecewise linear functions at a particular point is the sum of the slopes of the pieces at that point.

For instance, consider $h = -1/2$. $R(h)$ is the sum of the red, blue, and green functions plotted above. At this point, the slope of the red function is -1, while the slopes of the other two functions are zero. So the slope of the risk, R, at this point is $\frac{1}{3}(-1 + 0 + 0) = -1/3$.

We can break the function into pieces where the slope doesn't change: these "break points" will occur at the data points, and at points which are 1 to the left or 1 to the right of a data point. We calculate the slope in each section. Starting from the left, we can draw R as a piecewise linear function using these slopes.

c) Suppose that we are interested in finding the typical price of an avocado using this loss function. To do so, we have gathered a data set of n avocado prices, $y_1, \ldots, y_n$, and we found the price h^* which minimized the empirical risk (a.k.a, average loss), $R(h) = \frac{1}{n}\sum L(h, y)$.

Unfortunately, a flat tax of c dollars has been imposed on avocados since we performed our analysis, increasing every price in our data set by c.

Is it true that $h^* + c$ is a minimizer of R when we use the new prices, $(y_1 + c), (y_2 + c), \ldots, (y_n + c)$? Explain why or why not by explaining how the graph of R changes.

Solution: Yes, it is. The same number of points are greater than distance one away from h^* after the shift, and the points within distance once from h^* are the same distance away. Hence the graph of L is simply shifted over by c units, and the minimizers are shifted, too.

d) Suppose that instead of a flat tax, a tax of α-percent has been imposed. That is, the new avocado prices are $(1+\alpha)y_1, (1+\alpha)y_2, \ldots, (1+\alpha)y_n)$. Is $(1+\alpha)h^*$ sill a minimizer of R when we use the new prices? Explain why or why not.

Solution: No, it is not. Consider, for instance, the data set $\{1/5, 1/5, 2/5, 3/5, 4/5\}$. Because all of the data points are within distance one of another, the risk function is simply:

$$L(h) = \frac{1}{5}\sum_{i=1}^{5} |x_i - h|.$$

This is minimized by the median: $2/5$. On the other hand, scaling the data by 10 gives the data set: $\{2, 2, 4, 6, 8\}$. Because no pair of data points is within distance 1 of each other (except for 2 and 2), the risk becomes:

$$R(h) = \frac{1}{5}\left(\sum_{|x_i - h| > 1} 1 + \sum_{|x_i - h| \leq 1} |x_i - h|\right)$$

This is minimized at the mode, 2. Since $2 \neq 2/5 \cdot 10 = 4$, the claim is not true in general.

e) Given avocado prices $\{1/4, 1/2, 3/4, 7/8, 9/8\}$, find a minimizer of R. Provide some justification for your answer.

Hint: you don't need to plot R or do any calculation to find the answer.

Solution: Because all of the data points are within distance one of another, the risk function is simply:

$$R(h) = \frac{1}{5}\sum_{i=1}^{5} |x_i - h|.$$

This is the mean (absolute) error, and it is minimized by the median: $3/4$.

Problem 2.

The *Huber Loss* is a mixture between the square loss and the absolute loss. It is is defined piecewise as follows:

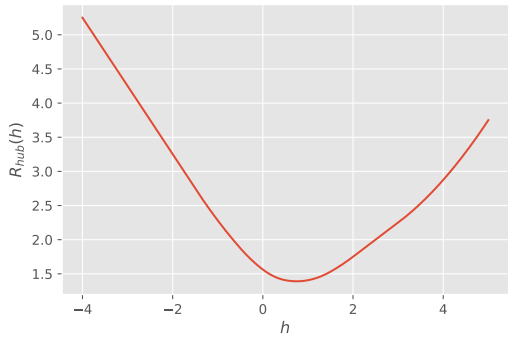
$$L_{\text{hub}}(h, y) = \begin{cases} |h - y|, & |h - y| > 1 \\ \frac{1}{2}(h - y)^2 + \frac{1}{2}, & |h - y| \leq 1 \end{cases}$$

a) What is the derivative of L_{hub} with respect to h? Your answer should also be a piecewise function.

Solution: We break L_{hub} into pieces and find the slope in each part. When $h - y < -1$, the function looks like $y - h$ and the slope is negative 1. Likewise, when $h - y > 1$, the function looks like $h - y$ and the slope is 1. In the middle, when $|h - y| \leq 1$, the function looks like $\frac{1}{2}(h-y)^2 + \frac{1}{2}$, and the slope is h. Hence

$$\frac{dL_{\text{hub}}}{dh}(h) = \begin{cases} -1, & h - y < -1 \\ 1, & h - y > 1 \\ h - y, & \text{otherwise} \end{cases}$$

b) Suppose $\{-\frac{1}{2}, \frac{1}{2}, 1, 4\}$ is a data set. The plot of the empirical risk, $R_{\text{hub}}(h) = \frac{1}{n}\sum L_{\text{hub}}(h, y)$ is shown below:



It is not possible to directly solve for the value of h which minimizes this function. Instead, run gradient descent by hand using an initial prediction of $h_0 = 5$ and a step size of $\alpha = 2$. Run the algorithm until it converges (it shouldn't take too many iterations). Please show your calculations. To help the graders track your progress, include a table with the value of h at each iteration, such as below:

$$\begin{aligned} h_0 &= 5 \\ h_1 &= ? \\ h_2 &= ? \\ h_3 &= ? \\ &\vdots \end{aligned}$$

Solution: First we must calculate the derivative of the risk. We have:

$$\begin{aligned} \frac{dR_{\text{hub}}}{dh}(h) &= \frac{d}{dh}\left[\frac{1}{n}\sum_{i=1}^{n} L_{\text{hub}}(h, y_i)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{dL_{\text{hub}}}{dh}(h, y_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \begin{cases} -1, & h - y_i < -1 \\ 1, & h - y_i > 1 \\ h - y, & \text{otherwise} \end{cases} \end{aligned}$$

We start the first iteration with $h_0 = 5$. To apply the gradient descent update rule, we first have to calculate the derivative of R_{hub} at $h_0 = 5$:

$$\begin{aligned}\frac{dR_{\text{hub}}}{dh}(5) &= \frac{1}{4}\sum_{i=1}^{n}\frac{dL_{\text{hub}}}{dh}(5, y_i) \\ &= \frac{1}{4}\left[\frac{dL_{\text{hub}}}{dh}\left(5, -\frac{1}{2}\right) + \frac{dL_{\text{hub}}}{dh}\left(5, \frac{1}{2}\right) + \frac{dL_{\text{hub}}}{dh}(5, 1) + \frac{dL_{\text{hub}}}{dh}(5, 4)\right]\end{aligned}$$

We now evaluate each derivative using the answer from the previous part.

$$\begin{aligned}&= \frac{1}{4}\left[1 + 1 + 1 + 1\right] \\ &= 1\end{aligned}$$

Applying the update rule, we find:

$$\begin{aligned}h_1 &= h_0 - \alpha\frac{dR_{\text{hub}}}{dh}(h_0) \\ &= 5 - 2 \cdot 1 \\ &= 3\end{aligned}$$

On to the second iteration. We star by calculating the slope at $h_1 = 3$:

$$\begin{aligned}\frac{dR_{\text{hub}}}{dh}(h_1) &= \frac{dR_{\text{hub}}}{dh}(3) \\ &= \frac{1}{4}\sum_{i=1}^{n}\frac{dL_{\text{hub}}}{dh}(3, y_i) \\ &= \frac{1}{4}\left[\frac{dL_{\text{hub}}}{dh}\left(3, -\frac{1}{2}\right) + \frac{dL_{\text{hub}}}{dh}\left(3, \frac{1}{2}\right) + \frac{dL_{\text{hub}}}{dh}(3, 1) + \frac{dL_{\text{hub}}}{dh}(3, 4)\right] \\ &= \frac{1}{4}\left[1 + 1 + 1 + (-1)\right] \\ &= \frac{1}{2}\end{aligned}$$

Applying the update rule, we find:

$$\begin{aligned}h_2 &= h_1 - \alpha\frac{dR_{\text{hub}}}{dh}(h_1) \\ &= 3 - 2 \cdot \frac{1}{2} \\ &= 2\end{aligned}$$

On the third iteration, we have:

$$\begin{aligned}
\frac{dR_{\text{hub}}}{dh}(h_2) &= \frac{dR_{\text{hub}}}{dh}(2) \\
&= \frac{1}{4}\sum_{i=1}^{n}\frac{dL_{\text{hub}}}{dh}(2, y_i) \\
&= \frac{1}{4}\left[\frac{dL_{\text{hub}}}{dh}\left(2, -\frac{1}{2}\right) + \frac{dL_{\text{hub}}}{dh}\left(2, \frac{1}{2}\right) + \frac{dL_{\text{hub}}}{dh}(2, 1) + \frac{dL_{\text{hub}}}{dh}(2, 4)\right] \\
&= \frac{1}{4}\left[1 + 1 + 1 + (-1)\right] \\
&= \frac{1}{2}
\end{aligned}$$

Applying the update rule, we find:

$$\begin{aligned}
h_3 &= h_2 - \alpha\frac{dR_{\text{hub}}}{dh}(h_2) \\
&= 2 - 2 \cdot \frac{1}{2} \\
&= 1
\end{aligned}$$

On the fourth iteration:

$$\begin{aligned}
\frac{dR_{\text{hub}}}{dh}(h_3) &= \frac{dR_{\text{hub}}}{dh}(1) \\
&= \frac{1}{4}\sum_{i=1}^{n}\frac{dL_{\text{hub}}}{dh}(1, y_i) \\
&= \frac{1}{4}\left[\frac{dL_{\text{hub}}}{dh}\left(1, -\frac{1}{2}\right) + \frac{dL_{\text{hub}}}{dh}\left(1, \frac{1}{2}\right) + \frac{dL_{\text{hub}}}{dh}(1, 1) + \frac{dL_{\text{hub}}}{dh}(1, 4)\right] \\
&= \frac{1}{4}\left[1 + \frac{1}{2} + 0 + (-1)\right] \\
&= \frac{1}{8}
\end{aligned}$$

Applying the update rule, we find:

$$\begin{aligned}
h_4 &= h_3 - \alpha\frac{dR_{\text{hub}}}{dh}(h_3) \\
&= 1 - 2 \cdot \frac{1}{8} \\
&= \frac{3}{4}
\end{aligned}$$

On the fifth iteration:

$$\begin{aligned}
\frac{dR_{\text{hub}}}{dh}(h_4) &= \frac{dR_{\text{hub}}}{dh}(3/4) \\
&= \frac{1}{4}\sum_{i=1}^{n}\frac{dL_{\text{hub}}}{dh}(3/4, y_i) \\
&= \frac{1}{4}\left[\frac{dL_{\text{hub}}}{dh}\left(3/4, -\frac{1}{2}\right) + \frac{dL_{\text{hub}}}{dh}\left(3/4, \frac{1}{2}\right) + \frac{dL_{\text{hub}}}{dh}(3/4, 1) + \frac{dL_{\text{hub}}}{dh}(3/4, 4)\right] \\
&= \frac{1}{4}\left[1 + \frac{1}{4} + (-\frac{1}{4}) + (-1)\right] \\
&= 0
\end{aligned}$$

Applying the update rule, we find:

$$\begin{aligned}
h_5 &= h_4 - \alpha\frac{dR_{\text{hub}}}{dh}(h_4) \\
&= \frac{3}{4} - 2 \cdot 0 \\
&= \frac{3}{4}
\end{aligned}$$

We have converged to $\frac{3}{4}$ after five steps. The progress of the algorithm was:

$$\begin{aligned}
h_0 &= 5 \\
h_1 &= 3 \\
h_2 &= 2 \\
h_3 &= 1 \\
h_4 &= 3/4 \\
h_5 &= 3/4
\end{aligned}$$

Problem 3.

We have so far been concerned with predicting numerical quantities, like salaries. Now suppose we want to predict which college an incoming UCSD student will be assigned to (e.g., Warren, Sixth, Muir, etc.). Predicting a discrete category (as opposed to a number) is an important machine learning task called *classification*.

We can use empirical risk minimization to make classifications, too. Suppose we have gathered a data set of n previous students and their colleges:

Warren
Sixth
Warren
Muir
Marshall
Warren
Warren
$\vdots$

The first step is to choose a number which will uniquely represent each college. For instance:

$$\begin{aligned} \text{Revelle} &\to 1 \\ \text{Muir} &\to 2 \\ \text{Marshall} &\to 3 \\ \text{Warren} &\to 4 \\ \text{Sixth} &\to 5 \end{aligned}$$

We then map each instance of a college to its corresponding number, giving us a new data set of numbers. For instance, the data above becomes:

$$\begin{matrix} 4 \\ 5 \\ 4 \\ 2 \\ 3 \\ 4 \\ 4 \\ \vdots \end{matrix}$$

a) Now that we have converted converted the data to a list of numbers, we can make a prediction by minimizing the mean absolute loss. Explain why this is not a good idea.

Solution: The minimizer of the mean absolute loss is the median, but the median is not a meaningful prediction here. For instance, suppose we had a data set of size 7: three students in Revelle, one student in Muir, and three students in Warren. After mapping the college names to numbers, the data set becomes $\{1, 1, 1, 2, 3, 3, 3\}$. The median is 2, which corresponds to Muir, despite this being the least popular college.

b) The *zero-one* loss is defined as follows:

$$L_{01}(h, y) = \begin{cases} 0, & h = y, \\ 1, & h \neq y. \end{cases}$$

As usual, define the risk to be:

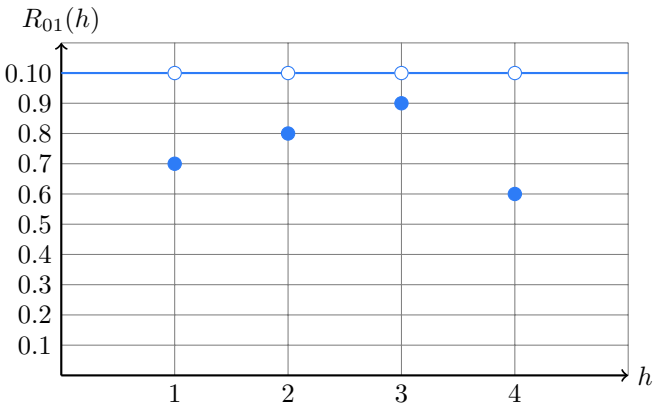
$$R_{01}(h) = \frac{1}{n} \sum_{i=1}^{n} L_{01}(h, y)$$

Notice that $R_{01}(h)$ can be interpreted as the misclassification rate. That is, if $R_{01}(h) = .7$, then predicting h would result in the wrong answer for 70% of the data points. Given the data set $\{1, 1, 1, 2, 2, 3, 4, 4, 4, 4\}$, plot the empirical risk $R_{01}(h)$ for $h \in [0, 5]$.

Hint: the function should have point discontinuities.

Solution:

At any h that does not coincide with a data point, the risk is one. If h does coincide with a data point, the risk is the fraction of data points not equal to h.



c) We have seen that the median minimizes the risk when the absolute loss is used, and that the mean minimizes the risk when the square loss is used. What quantity minimizes the risk when the zero-one loss is used?

Solution: The mode minimizes the risk with zero-one loss. This is because the lowest value of the risk occurs when h is equal to as many data points as possible.

d) Is gradient descent useful for minimizing the risk with zero-one loss? Why or why not? Make reference to your plot of the risk in your answer.

Hint: the risk is indeed non-convex, but gradient descent can still be useful for minimizing non-convex functions. Is there some other reason?

Solution: Gradient descent cannot be used to minimize the risk with zero-one loss. One reason for this is that the risk has point discontinuities, and is therefore not differentiable. Another reason is that the slope of the risk is almost everywhere zero, and so the derivative is not useful.

Problem 4.

The gradient descent update rule for minimizing a function $R(h)$ is:

$$h_{\text{next}} = h_{\text{prev}} - \alpha \frac{dR}{dh}(h_{\text{prev}}).$$

We said in class that the sign of dR/dh is meaningful: if it is positive we should move to the left, and if it negative we should move to the right.

Why is the *magnitude* of the derivative useful, too? That is, what is wrong with using the update rule:

$$h_{\text{next}} = h_{\text{prev}} - \alpha \cdot \text{sign}\left(\frac{dR}{dh}(h_{\text{prev}})\right),$$

where $\text{sign}(\cdot)$ returns the sign of its argument as either zero or one. For instance, $\text{sign}(-4) = -1$ and $\text{sign}(42) = 1$.

Solution: The magnitude of the derivative is useful in telling us when to stop the gradient descent procedure. If we did not use the magnitude, every step would be of size α and we would not be able to tell when we have converged. On the other hand, the magnitude of the derivative tells use the steepness of the function at the current position; if it is near zero, we may be near a local minimum.

Problem 5.

In class, we saw that convex risk functions are nice because they are relatively easy to minimize using gradient descent. But how do we determine if our risk function is convex? One way is to show that it is built from simpler convex functions.

Suppose that $f_1(x)$ and $f_2(x)$ are convex functions defined on all real numbers. We wish to show that their sum, $f(x) = f_1(x) + f_2(x)$, is also a convex function.

One way to show that f is convex is to prove that it satisfies the definition. That is, it is sufficient to show that for any real numbers a and b and for all $t \in (0, 1)$,

$$f(ta + (1-t)b) \leq tf(a) + (1-t)f(b).$$

Prove that this holds by using a chain of inequalities.

Hint: when proving something like this, first identify the special things that we know about the important entities in the problem. In this case, we know that 1) f is the sum of f_1 and f_2; and 2) f_1 and f_2 are convex functions. We will need to use both pieces of information in our proof. Which should we use first? If you get stuck, ask yourself: have I used all of these pieces of information yet?

Solution: We want to show that, for any $a, b \in \mathbb{R}$ and $t \in (0, 1)$,

$$f(ta + (1-t)b) \leq tf(a) + (1-t)f(b).$$

We will start with $f(ta + (1-t)b)$ and try to make it look like the right hand side using a chain of inequalities.

We have two pieces of information that we know must be used in the proof: 1) f is the sum of f_1 and f_2, and 2) f_1 and f_2 are convex. We can't use the second piece of information yet, since we don't see f_1 or f_2, but we can use the first:

$$f(ta + (1-t)b) = f_1(ta + (1-t)b) + f_2(ta + (1-t)b)$$

Now we can use the second piece of information. Since f_1 is convex, we know that $f_1(ta + (1-t)b) \leq tf_1(a) + (1-t)f_1(b)$; a similar statement holds for f_2. Therefore:

$$\leq [tf_1(a) + (1-t)f_1(b)] + [tf_2(a) + (1-t)f_2(b)]$$

It helps to remember where we are trying to go. We want to make this look like $tf(a) + (1-t)f(b)$. We currently have an expression involving f_1 and f_2, however. We need to apply the first piece of information – that f is the sum of f_1 and f_2 – to make progress. First, we'll regroup the terms:

$$= [tf_1(a) + tf_2(a)] + [(1-t)f_1(b) + (1-t)f_2(b)]$$

Taking out the common factors:

$$= t\left[f_1(a) + f_2(a)\right] + (1-t)\left[f_1(b) + f_2(b)\right]$$

We recognize $f_1(a) + f_2(a)$ as $f(a)$ and $f_1(b) + f_2(b)$ as $f(b)$:

$$= tf(a) + (1-t)f(b)$$

This is what we wanted to prove. Therefore, f is indeed a convex function.

Problem 6.

Remember that there are several ways of showing that a function is convex:

1. From the definition.
2. Use the second derivative test for convexity.
3. Show that the function is built from other convex functions using one or more of the properties mentioned in lecture (i.e., it is the sum, composition, or pointwise maximum of convex functions).

In each of the problems below, use one of the above justifications to prove that the function is convex.

a) $f(x) = x$

> **Solution:** The second derivative test gives $f''(x) = 0 \geq 0$, so the function is convex.

b) $f(x) = e^x$

> **Solution:** $f(x)$ is the composition of a non-decreasing convex function and a convex function, and is therefore convex.

c) $f(x) = |x|$

> **Solution:** $|x|$ is the pointwise maximum of two convex functions: $-x$ (which is convex due to the second derivative test) and x. As a result, it is also convex.

d) The mean absolute error (as a function of h; consider the y_i as fixed):

$$R_{\text{abs}}(h) = \frac{1}{n}\sum_{i=1}^{n} |h - y_i|$$

> **Solution:** The function $|h - y_i|$ is convex because it is the pointwise maximum of $h - y_i$ and $y_i - h$. Hence R_{abs} is the sum of convex functions of h, and so it is also convex.

Problem 7.

Linear algebra plays an important role in machine learning, and we will be using it soon. These questions will help you remember some of the basics from your course on linear algebra.

a) Define the matrix X and the vector $\vec{a}$ as below:

$$X = \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{pmatrix}, \qquad \vec{a} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}.$$

One way to interpret the result of matrix-vector multiplication is that it is a mixture of the columns of X. To see this, show that $X\vec{a} = a_1\vec{x}^{(1)} + a_2\vec{x}^{(2)} + a_3\vec{x}^{(3)}$, where

$$\vec{x}^{(1)} = \begin{pmatrix} x_{11} \\ x_{21} \\ x_{31} \end{pmatrix}, \qquad \vec{x}^{(2)} = \begin{pmatrix} x_{12} \\ x_{22} \\ x_{32} \end{pmatrix}, \qquad \vec{x}^{(3)} = \begin{pmatrix} x_{13} \\ x_{23} \\ x_{33} \end{pmatrix}$$

are the columns of X.

b) Recall that the dot product of two vectors $\vec{u} = (u_1, u_2, u_3)^\intercal$ and $\vec{v} = (v_1, v_2, v_3)^\intercal$, written $\vec{u} \cdot \vec{v}$, is the number $u_1v_1 + u_2v_2 + u_3v_3$.

Let $\vec{u}$ and $\vec{v}$ be as above, and let $\vec{w} = (w_1, w_2, w_3)^\intercal$. Show that

$$\vec{u} \cdot (\vec{v} + \vec{w}) = \vec{u} \cdot \vec{v} + \vec{u} \cdot \vec{w}$$

by writing out the left hand side and the right hand side in terms of the elements of each vector and showing that they are equal.