Introduction to Machine Learning
Homework 1: Multiple Linear Regression

Prof. Linda Sellie

1. Modified Question 2 on page 52 from "An Introduction to Statistical Learning."[1]

   "Explain whether each scenario is a classification or regression problem, and provide N and d (d is the number of features).

   (a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

   (b) We are considering launching a new product and wish to know whether it will be a *success* or *failure*. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

2. Question 4 on page 53 from "An Introduction to Statistical Learning."[2]

   Think of real-life applications for machine learning

   (a) Describe three real-life applications in which *classification* might be useful. Describe the target, as well as the features.

   (b) Describe three real-life applications in which *regression* might be useful. Describe the target, as well as the features.

3. A university admissions office wants to predict the success of students based on their application material. They have access to past student records as training data.

   (a) 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.

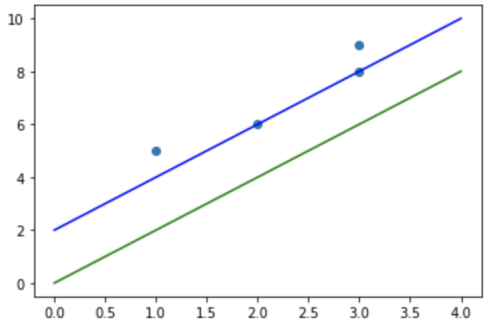
[1] I have slightly modified the information to use the notation we are using in our course.
[2] I have modified the questions.

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

(c) State at least one possible variable that can act as the feature (aka[3] predictor) for the target variable you chose in part (a).

(d) Before looking at the data, would a linear model for the data be reasonable? If so, what sign do you expect the slope to be?

4. Consider data $(1,5),(2,6),(3,8),(3,9)$ and regression lines: $y = 2x_1$ (the green line), $y = 2x_1 + 2$ (the blue line). (Note here $\mathbf{x} = [x_1]$.)



(a) What is the squared error of each of the points[4] with respect to the line $y = 2x_1$?

(b) The gradient of our cost function includes a sum over contributions of individual points. We could calculate the individual contributions separately. The gradient for a single $(\mathbf{x}^{(i)}, y^{(i)})$ point is:[5] $\begin{bmatrix} (w_0 + w_1 \cdot x_1^{(i)} - y^{(i)}) \\ (w_0 + w_1 \cdot x_1^{(i)} - y^{(i)}) x_1^{(i)} \end{bmatrix}$
For the line $y = 2x_1$ what is the gradient contribution for each of the four examples?[6]

(c) What is the squared error of each of the points with respect to the line $y = 2x_1 + 2$? (the blue line in the graph)[7]

(d) What is the gradient contribution for each of the four examples to the line $y = 2x_1 + 2$?

(e) Which line has a smaller RSS?

(f) Would it be possible for a different line to have a smaller RSS?

[3] 'aka' is 'also known as'

[4] For the first example, the answer would be $(2 \cdot 1 - 5)^2 = 9$.

[5] We are taking the gradient of $\frac{1}{2}(w_0 + w1 \cdot x_1^{(i)} - y^{(i)})^2$.

[6] For the first example, the answer is: $\begin{bmatrix} (0 + 2 - 5) \\ (0 + 2 - 5)1 \end{bmatrix} = \begin{bmatrix} -3 \\ -3 \end{bmatrix}$

[7] For the first example the answer would be $(2 \cdot 1 + 2 - 5)^2 = 1$.

5. Given[8] the following data $((x_1, x_2)^T, y)$: $((0,0)^T, 1), ((0,1)^T, 4), ((1,0)^T, 3), ((1,1)^T, 7)$

   (a) create the design matrix X (include the column of 1's)

   (b) create the target vector y

   (c) write out the closed form formula for computing $\mathbf{w}$ that minimizes RSS($\mathbf{w}$)

   (d) determine the w_0, w_1, w_2 that minimizes RSS($\mathbf{w}$)

   (e) compute RSS

   (f) compute TSS

   (g) compute R^2

   (h) what portion of the variance in y is explained by $\mathbf{x}$?

   (i) predict the value of $\mathbf{x}^T = (0.5, 0.5)$ using the values of $\mathbf{w}$ computed in question 5d

6. For the following function: $f(w_0, w_1) = (w_0 + 2w_1 - 4)^2 + (w_0 + 3w_1 - 3)^2$

   (a) Determine the gradient $\nabla f(w_0, w_1)$

   (b) Run the gradient descent algorithm for `num_iters` $= 10$ iterations (you can use your computer to perform the calculations) where you try different learning rates. For each start with $(w_0, w_1) = (0, 0)$:

      i. learning rate of $\alpha = 0.06$

      ii. learning rate of $\alpha = 0.001$

      iii. learning rate of $\alpha = 0.03$

      Report the value of w_0, w_1 and $f(w_0, w_1)$ at the end of each step. On one graph, plot the points (w_0, w_1) at every iteration.

      Evaluate (briefly in one sentence) how each learning rate contributed or did not contribute to finding a new assignment to the parameters that decreased the value of the function.

7. Do not turn in the following question. A solution will be provided by the TAs.

[8] For this question when you need to calculate a value 1) write out the formula using the numbers given in the problem, 2) to compute the value you can use numpy or a calculator etc. Just make sure you could do this by hand if you need to in an exam. (Don't worry, I won't ask you to compute the inverse of a matrix.) These instructions are true for any future homework assignment unless otherwise specified.

Given the following data matrix:

$$X = \begin{bmatrix} 1 & small & Chevy & 130 \\ 1 & large & Buick & 165 \\ 1 & medium & Plymouth & 150 \\ 1 & medium & Ford & 140 \\ 1 & small & Ford & 198 \\ 1 & medium & Chevy & 150 \\ 1 & large & Buick & 225 \end{bmatrix}$$

Perform One-hot encoding on the second feature, and perform an ordinal encoding on the first feature. In your answer provide the transformed data matrix.

8. Do not turn in the following question.

   For linear regression, on a data set $X = \{(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(N)}, y^{(N)})\}$, if the ith feature for every example is scaled by a constant c , does $\mathbf{w}$ change? If it does change, describe how.

9. Do not turn in the following question.

   An online retailer like Amazon wants to determine which products to promote based on reviews. They only want to promote products that are likely to sell. For each product, they have past sales as well as reviews. The reviews have both a numeric score (from 1 to 5) and text.

   (a) To formulate this as a machine learning problem, suggest a target variable that the online retailer could use.

   (b) For the predictors of the target variable, a data scientist suggests combining the numeric score with the frequency of occurrence of words that convey judgment like "bad", "good," and "doesn't work." Describe a possible linear model for this relation.

   (c) Now, suppose that some reviews have a numeric score from 1 to 5 and others have a score from 1 to 10. How would change your features?

   (d) Now suppose the reviews have either (a) a score from 1 to 5; (b) a rating that is simply good or bad; or (c) no numeric rating at all. How would you change your features?

   (e) For the frequency of occurrence of a word such as "good", which variable would you suggest using as a feature (aka predictor): (a) total number of reviews with the word "good"; or (b) fraction of reviews with the word "good"?

10. Do not turn in this question.

The problem with the least square loss in the existence of outliers (i.e., when the noise term can be arbitrarily large).

In class we looked at two cost functions:

- Absolute value of the difference (i.e. sum of absolute difference $\sum_{i=1}^{N} |y^{(i)} - \hat{y}^{(i)}|$)
- Squared value of the difference (i.e residual sum of squares $\sum_{i=1}^{N} (y^{(i)} - \hat{y}^{(i)})^2$)

We choose to minimize the residual sum of squares instead of the sum of absolution residuals. Which of these two methods would most likely be affected by outliers?

11. Do not turn in this question

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

    $$z(t) \approx z_0 e^{-\alpha t}$$

    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 $(t_i, z(t_i))$, $i = 1, \ldots, N$. Unfortunately, the model (11) is non linear, so she can't directly apply the linear regression formula.

    (a) Taking logarithms, show that we can rewrite the model in a form where the parameters z_0 and α appear linearly.

    (b) Using the transform in part (a), write the least-squares solution for the best estimates of the parameters z_0 and α from the data.[9]

12. Do not turn in this question

    Consider a linear model of the form,

    $$y \approx wx,$$

    which is a linear model, but with the intercept forced to zero. This occurs in applications where we want to force the predicted value $\hat{y} = 0$ 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) Given data (x_i, y_i), write a cost function representing the residual sum of squares (RSS) between y_i and the predicted value $\hat{y}_i$ as a function of w.

    (b) Taking the derivative with respect to w, find the w that minimizes the RSS.

[9] Provide the formula. You are not applying this formula to a specific dataset.