CS6923 Machine Learning, Fall 2023
Prof. Linda Sellie, NYU School of Engineering

Homework 3[1]

Submit on Gradescope by Friday, September 29 at 11:00 p.m.

You may work together with one other person on this homework. If you do that, *hand in JUST ONE homework for the two of you*, with both of your names on it. You may *discuss* this homework with other students but YOU MAY NOT SHARE WRITTEN ANSWERS OR CODE WITH ANYONE BUT YOUR PARTNER.

Part I: Written Exercises

1. A medical researcher wishes to evaluate a new diagnostic test for cancer. A clinical trial is conducted where the diagnostic measurement y of each patient is recorded along with attributes of a sample of cancerous tissue from the patient. Three possible models are considered for the diagnostic measurement:

   - Model 1: The diagnostic measurement y depends linearly only on the cancer volume.
   - Model 2: The diagnostic measurement y depends linearly on the cancer volume and the patient's age.
   - Model 3: The diagnostic measurement y depends linearly on the cancer volume and the patient's age, but the dependence (slope) on the cancer volume is different for two types of cancer – Type I and II. (Hint: Use a variable x_3 which is assigned the value 1 if the cancer is Type I, and x_3 has the value 0 if the cancer is of Type II.)

   (a) Define variables for the cancer volume, age and cancer type and write a linear model for the predicted value $\hat{y}$ in terms of these variables for models 1 & 2 above.

   (b) Do the same for model 3. For Model 3, you will want to use one-hot coding as mentioned above.

   (c) What are the number of parameters in model 1 & 2? Which model is the most complex?

   (d) Since the models in part (a) are linear, given training data, we should have $\hat{\mathbf{y}} = X\mathbf{w}$ where $\hat{\mathbf{y}}$ is the vector of predicted values on the training data, X is a design matrix (feature matrix) and $\mathbf{w}$ is the vector of parameters. To test the different models, data is collected from 100 patients. The records of the first three patients are shown below:

Patient ID	Measurement y	Cancer type	Cancer volume	Patient age
12	5	I	0.7	55
34	10	II	1.3	65
23	15	II	1.6	70
⋮	⋮	⋮	⋮	⋮

   For model 1 in part (a), based on this data, what are the first three rows of the matrix X?
   For model 2 in part (a), based on this data, what are the first three rows of the matrix X?

   For model 3 in part (a), based on this data, what are the first three rows of the matrix X?

   (e) To evaluate the models, 10-fold cross validation is used with the following results.

[1]Some of these are modified from Prof. Rangan's questions.

Model	training MSE	validation MSE
1	2.0	2.01
2	0.7	0.72
3	0.65	0.90

Which model should be selected?

2. Suppose you were interested in crop yields and you had collected data on the amount of rainfall, the amount of fertilizer, the average temperature, and the number of sunny days.

   How could you formalize this a as regression problem?

3. This data for this problem is from `https://stats.libretexts.org/Homework_Exercises/General_Statistics/Exercises%3A_OpenStax/12.E%3A_Linear_Regression_and_Correlation_(Exercises)`

	Mid-Career Salary (in thousands)	Yearly Tuition
Princeton	137	28,540
Harvey Mudd	135	40,133
CalTech	127	39,900
US Naval Academy	122	0
West Point	120	0
MIT	118	42,050
Lehigh University	118	43,220
NYU-Poly	117	39,565
Babson College	117	40,400
Stanford	114	54,506

   Suppose we want to predict the mid-career salary from a person based on the cost of the college they attended. To do this, apply linear regression to the above dataset, using the closed-form formula presented in class.[2] Then answer the following questions.

   (a) The closed form formula uses a matrix X. What is the value of the matrix X in this problem?

   (b) Give the equation for the linear function (line) produced using linear regression, in the form $g(x) = w_1 x + w_0$.

   (c) Create a scatter plot of the data, and plot the least squares line in your graph.

   (d) Compute the determination of correlation, R^2.

   (e) Using this linear function, what is the predicted the mid-career salary of a person whose yearly college tuition costs $40,000$?

   (f) Repeat the above steps where the outliers are removed (the outliers are the two service academies whose tuition is $0.00)

4. In this problem, each example has only one feature: $\mathbf{x^{(i)}} = x_1^{(i)}$ for $i = 1 \dots N$.

   Consider a linear model of the form,

   $$y \approx \mathbf{w}x,$$

   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 $(x^{(i)}, y^{(i)})$ for $i \in 1 \dots N$, 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 $\mathbf{w}$.

[2] You may use a computer/calculator to perform the steps - just make sure you know how to do this by hand.

(b) Taking the derivative with respect to $\mathbf{w}$, find the $\mathbf{w}$ that minimizes the RSS.

5. Consider a linear regression model $y^{(i)} = \mathbf{w}^T\mathbf{x}^{(i)} + \epsilon^{(i)}$, where instead of assuming the noise $\epsilon^{(i)} \sim N(0, \sigma^2)$ we assume $\epsilon^{(i)} \sim \text{Laplace}(0, b) = \frac{1}{2b}\exp\left(-\frac{|\epsilon^{(i)}|}{b}\right)$

   - Show that the maximum likelihood estimate $\mathbf{w}$ is the one that minimizes $\sum_{i=1}^{N}|y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)}|$.
   - Informally discuss why this model will be more robust to noise as compared to the model where we assume the noise $\epsilon^{(i)} \sim N(0, \sigma^2)$.

6. (Do not turn in)

   Suppose when performing linear regression, we cared more about getting some examples correct than on getting other examples correct. To model this we could *weigh* the cost of some mistakes more than others.

   Our goal is now to find:

$$\hat{\mathbf{w}}_{WLS} = \arg\min_{\mathbf{w}\in\mathbb{R}^d}\left(\sum_{i=1}^{N} w^{(i)}(y^{(i)} - \mathbf{x}^{(i)T}\mathbf{w})^2\right)$$

   Find a closed form solution to this problem. (Hint: You can create a matrix $\Omega \in \mathbb{R}^{N\times N}$ which is a diagonal matrix where $\Omega_{i,i} = w^{(i)}$. Now $\hat{\mathbf{w}}_{WLS} = \arg\min_{\mathbf{w}\in\mathbb{R}^d}(\mathbf{y} - X\mathbf{w})^T\Omega(\mathbf{y} - X\mathbf{w})$.)

7. (Do not turn in)

   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.

   .

8. (Do not turn in)

   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.

Part II: Programming Exercise

In this programming experiment you will implement k nearest neighbors (k-NN) to classify an Iris as *Iris Setosa*, *Iris Versicolour*, or *Iris Virginica*.

The dataset can be found at `https://archive.ics.uci.edu/ml/datasets/iris`

Each line of the file consists of 4 measurements (*sepal length in cm, sepal width in cm, petal length in cm, and petal width in cm*) and the type of iris these measurements came from.

In your algorithm, you will only use two of these features: *sepal length* and *petal width*. (We are only using these two features since this is an assignment for you to experiment with different values of k and different distance measurements. With all 4 features, it is easy to classify the Iris perfectly; try this when you have completed your assignment)

Implement the k-Nearest Neighbor algorithm to classify the test examples, using an *euclidean-distance* function[3].

Run your algorithm with $k = 1$, $k = 3$ and with $k = 5$.

For $k > 1$, you can simply sort the training examples by distance from the test example, smallest to largest, to find the k nearest training examples. (You can also use a more clever algorithm to find the k nearest training examples.)

Answer the following questions in a pdf file called `proganswers.pdf`:

1. For $k = 1$, which examples were not correctly classified?
2. Report the accuracy on the test set for $k = 1$.
3. For $k = 3$, which examples were not correctly classified?
4. Report the accuracy on the test set for $k = 3$.
5. For $k = 5$, which examples were not correctly classified?
6. Report the accuracy on the test set for $k = 5$.
7. Suppose we used the very simple Zero-R classifier on this dataset, rather than k-NN. That is, we classify all examples in the test set as belonging to the class that is most common in the training set. What is the resulting accuracy?
8. To learn about different distance metrics, read section 2.8 in `https://sebastianraschka.com/pdf/lecture-notes/stat479fs18/02_knn_notes.pdf`. Try running your k-NN algorithm with a different the distance metric.

   When you run k-NN with your new distance function, use the same training and test sets, to classify the examples in the test set.

   I don't expect you to achieve higher accuracy (for $k = 1$, $k = 3$, or $k = 5$) than the first distance function. What is a possible reason that it didn't achieve a higher accuracy?[4]
9. (extra credit) Implement 5-fold cross-validation on the training set to determine which of the following values of k works better in k-NN: $3, 7, 9$.

   To implement the 5-fold cross-validation, divide the training set into 5 sets of equal size. In practice you may want to randomly permute the data before dividing it into 5 sets, but for this assignment, just take the first 1/5 of the examples listed in the file, then the second 1/5, etc. (and DON'T PERMUTE) the examples first.

   Then for each of the 3 values of k, do the following. (1) For each of the 5 sets, train on the examples in the other four sets, and test on the examples in the 5th set. The result is that a prediction has been made on each example in the training set. (2) Calculate the percentage of these predictions that were correct. This is the cross-validation accuracy (for this value of k).

[3]To make the grading easier, if there is a tie involving 'Iris-setosa', then return 'Iris-setosa'. Otherwise if there is a tie between 'Iris-versicolor' and 'Iris-virginica', then return 'Iris-versicolor'.

[4]If we use the "test" set to choose which distance function and value of k was best - we should be calling this set a "validation" set and not a "test" set.