20-ECES-6037/5137

Instructor: Anca Ralescu  
Machine Learning   
Fall 2021

**Homework #1**  
**Assigned: August 31st, 2021**  
**Due on Canvas, September 7, 2021, by 11:59PM**

Your answers to this assignment must all be uploaded on Canvas.

For problems 1 and 2, you are to consider the questions they pose **conceptually**.  I call these problems “pencil-and-paper” problems. That is, no programming is required.

**DO NOT WRITE MORE** than ONE page per problem. Provide clear answers typed.

1. **points) Problem 1**   
   Consider the problem of designing an algorithm to play the game of Tic-Tac-Toe.  Set up the problem following the example of the checkers problem.  That is, carefully decide the board features which you want to use for defining the estimate V' as a linear combination of these features.

You must address all the issues in order to set up the problem for programming, but **you are not to implement this**. 

**(15 points) Problem 2**    
A good "straw man" (intuitive) learning algorithm is as follows:

1. Create a table of all training examples.
2. Determine which output occurs most often among the training examples and call this output **d**.
3. Then, given an input do the following
   1. If the input is NOT in the table, just return **d**.
   2. If the input is in the table, return the corresponding output from the table (if there are more than one output for an input return the most frequent one).

It can be argued that this algorithm gives an idea of the *baseline* for the problem, that is, it is the minimal performance that a learning algorithm should have.

Consider again the Tic-Tac-Toe example and consider giving such a training table.  Intuitively discuss the performance of the algorithm:

1. as a **function of the size of the training set**, that is, for example how accurate would expect the algorithm to be given the training set (table) size;
2. as a **function of the size of a data point** (i.e., number of features)**.**

**You are not to implement this problem.**

**(25 points) Problem 3.**This is a programming problem. Implement linear regression for the data set **insurance.csv** supplied with this assignment. The data set is a table: columns 1-6 describe the attributes of an individual - age, sex, bmi, number of children, smoker, region. Column 7 shows the insurance charge.

Note that some attributes are real-valued (e.g., age, bmi), others are integers greater than or equal to 0 (e.g., number of children), others are categorical (gender, region, smoker).

Eliminate the categorical attributes. For example, the rows

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **age** | gender | bmi | No. of children | smoker | region | Premium |
| **19** | female | 27.9 | 0 | yes | southwest | 16884.924 |
| **18** | male | 33.77 | 1 | no | southeast | 1725.5523 |
| **28** | male | 33 | 3 | no | southeast | 4449.462 |
| **33** | male | 22.705 | 0 | no | northwest | 21984.47061 |
| **32** | male | 28.88 | 0 | no | northwest | 3866.8552 |
| **31** | female | 25.74 | 0 | no | southeast | 3756.6216 |
| **46** | female | 33.44 | 1 | no | southeast | 8240.5896 |
| **37** | female | 27.74 | 3 | no | northwest | 7281.5056 |
| **37** | male | 29.83 | 2 | no | northeast | 6406.4107 |
| **60** | female | 25.84 | 0 | no | northwest | 28923.13692 |

**Are converted into**

|  |  |  |  |
| --- | --- | --- | --- |
| **age** | bmi | No. of children | Premium |
| **19** | 27.9 | 0 | 16884.924 |
| **18** | 33.77 | 1 | 1725.5523 |
| **28** | 33 | 3 | 4449.462 |
| **33** | 22.705 | 0 | 21984.47061 |
| **32** | 28.88 | 0 | 3866.8552 |
| **31** | 25.74 | 0 | 3756.6216 |
| **46** | 33.44 | 1 | 8240.5896 |
| **37** | 27.74 | 3 | 7281.5056 |
| **37** | 29.83 | 2 | 6406.4107 |
| **60** | 25.84 | 0 | 28923.13692 |

And therefore, X, the data matrix for this would be

|  |  |  |
| --- | --- | --- |
| **19** | 27.9 | 0 |
| **18** | 33.77 | 1 |
| **28** | 33 | 3 |
| **33** | 22.705 | 0 |
| **32** | 28.88 | 0 |
| **31** | 25.74 | 0 |
| **46** | 33.44 | 1 |
| **37** | 27.74 | 3 |
| **37** | 29.83 | 2 |
| **60** | 25.84 | 0 |

While the corresponding output (let us call it y) is

|  |
| --- |
| 16884.924 |
| 1725.5523 |
| 4449.462 |
| 21984.47061 |
| 3866.8552 |
| 3756.6216 |
| 8240.5896 |
| 7281.5056 |
| 6406.4107 |
| 28923.13692 |

Implement linear regression using the **Normal Equation**.

Divide the data into (1) a training set, and (2) a test set.

Choose the training set randomly and let the remaining data form the test set. Use the data matrix obtained from the training set to estimate the parameter q.

Then for each data point use q to obtain the model output for that data point.

For each choice of the training and test data points, compute two errors using the MSE:

1. **The error on the training set**. This gives us an idea of how well the model captures the training data. We call this the *modeling power* of the model.
2. **The error on the test set**. This gives us an idea how good is the model on the test data. We call this the *generalization power* of the model.

Repeat the above procedure on training sets of increasing size from 20% to 80% of the whole data, in increments of 10%, that is

|  |  |
| --- | --- |
| Training set size | Test set size |
| 20% | 80% |
| 30% | 70% |
| 40% | 60% |
| …. | … |
| 80% | 20% |

For each choice of the training set

1. Plot the regression line as function of the bmi
2. Plot the regression line as function of age
3. Plot the regression line as function of number of children

Plot the modeling and generalization errors as functions of the training sets size

Turn in:

1. your complete code (include the team names at the top of program, and carefully comment the code).
2. The results (including all figures) with an analysis. Write your report clearly and very carefully.