**Machine Learning PS2: Regression**

Template Write-Up

**NOTES:**

* Due to the nature of Q1 & Q2 (it is difficult to write mathematical expressions in MS Word), it is fine to submit a physical (handwritten) copy of your answers, if that is what you prefer. **However, it must be scanned and attached along with your code on NYU Classes, and the answers to the appropriate questions must be clearly labeled**. Write “see attached document [document name]” for every question you do this for in this template document.
* Note that Q1, part 1 has a typo:
  + ~~J(~~**~~w~~**~~) = J(w1, w2, w3)~~ 🡪 J(**w**) = J(w0, w1, w2)
* Note that for Q2B part (c), J(**w**) decreases with each pass of gradient descent, but will never actually converge to zero. Find a workaround – either implement a pass limit, or stop doing passes if the loss function J(**w**) isn’t changing much from pass to pass (define a threshold).

**PROBLEM 1 (20 points)**

Q1-1. What is the cost function J(**w**) = J(w0, w1, w2) for the 4-item dataset in Problem 1?

|  |
| --- |
|  |

Q1-2. What are the values of (w0, w1, w2) after one iteration of gradient descent? Comment on whether you normalized the data or not for this pass by hand.

|  |
| --- |
|  |

**PROBLEM 2 (60 points)**

\* Make sure that all your relevant code / implementation is attached. You must implement the relevant functions (mean, standard deviation, partial derivatives, the loss function, vanilla and stochastic gradient descent, etc.) yourself.

Q2A-a. Given a set of numbers x1,…,xm, write down the equations for the mean and the standard deviation of these numbers. **USE POPULATION STDEV** (***not*** simple stdev)

|  |
| --- |
| The mean  The standard deviation |

Q2B-a. Write down the formula for the loss function J(**w**) using the sum of the squared errors. Be sure to include a 1/2m term. Is it different from your answer to Q1-1?

|  |
| --- |
| They appear different but the answer to Q1-1 is simply an expansion of the formula above using the provided data. |

Q2B-b. Derive the partial derivatives of J(**w**) with respect to w0, w1, and w2.

|  |
| --- |
|  |

Q2B-c. For learning rates α = 0.01, 0.1 and 0.3. Plot the J(**w**) values for [10,80] for each of the three learning rates. Comment on which gives the best result.

|  |
| --- |
| Learning rate of 0.01 |

Q2B-d. For learning rates α = 0.05 and 0.3. Comment on which of the three learning rates gives the best result.

|  |
| --- |
|  |

Q2C. Use the **w** obtained in Q2B to predict the housing prices. Predict the price of a house with 1650 sq. ft. area and 3 bedrooms. Don’t forget to normalize the features when you make this prediction!

|  |
| --- |
|  |

Q2D. Use learning rate α = 0.1 and compare the result of vanilla gradient descent using 80 passes over the data – that is, the w0, w1, and w2 values – to the result of stochastic gradient descent with 3 full passes over the data. Provide the value of J(**w**) after each pass of SGD.

|  |
| --- |
|  |

**PROBLEM 3 (20 points)**

Q3. Prove that the perceptron algorithm will have at most R2/γ2 mistakes. Justify all steps in your proof.

|  |
| --- |
|  |