DATA473: Assignment 1

Due: 4pm on Friday 3 April, 2020

(Extension granted to: 4pm on Friday 10 April, 2020)

- You may work on this assignment individually, or as a pair.
- Any code used for this assignment should be written in Python.
- There are 6 marks for presentation.
- There are a total of 35 marks for this assignment.

Submitting your assignment: There will be an assignment dropbox folder on the Learn page for DATA473 and you should submit you assignment electronically.

- Your main assignment should be submitted as a pdf.
- You should submit your code as a python notebook.

You must **write your name and student ID number** (or both names and both student ID numbers if working as a pair) on your assignment pdf and all submitted code.

1. **[3 marks]** Scientists uncovered the following matrices of weights describing a fully-connected neural network. Draw a diagram of this network, depict size of all layers and connections between units. Calculate the output of this network $f(x)$.
The weights for the hidden layer are given in the matrix:

$$W^{[1]} = \begin{bmatrix} 5 & 10 & 5 \\ 1 & 2 & 3 \end{bmatrix}$$

The biases for the hidden layer are given in the vector:

$$b^{[1]} = \begin{bmatrix} 5 \\ -15 \end{bmatrix}$$

The weights for the output layer are given in the vector:

$$W^{[2]} = \begin{bmatrix} 1 & 2 \end{bmatrix}$$

The bias for the output layer is $b^{[2]} = -34$.
The input X is given in the vector:

$$X = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}$$

The activation function for all units in the hidden layer is Relu: $g(z) = \max(0, z)$
The activation function for the output unit is sigmoid: $g(z) = \frac{1}{1+e^{-z}}$

2. **[7 marks]**

   (a) Draw the unit norm balls for the 1-, 2- and ∞-norms.

   (b) Give 3 different examples of a convex set (a diagram is sufficient).

   (c) Give 3 different examples of a nonconvex set (a diagram is sufficient).

   (d) Give an example of a convex function with a unique optimal solution (both a diagram and the corresponding mathematical function are needed).

3. **[7 marks]** Consider the function:

$$f(w_1, w_2) = 1.25(w_1 + 6)^2 + (w_2 - 8)^2. \tag{1}$$

(a) State the gradient $\nabla f(w_1, w_2)$ for the function in (1).

(b) State the Hessian $\nabla^2 f(w_1, w_2)$ for the function in (1).

(c) What are the necessary conditions for a minimizer of this function?

(d) What is the optimal solution to $\min\limits_{\mathbf{w}} f(w_1, w_2)$.

(e) This function has an L-Lipschitz continuous gradient. Compute L. [Hint: $f(w_1, w_2)$ can be expressed as $\frac{1}{2}\|X^T\mathbf{w} - \mathbf{y}\|_2^2$ for some X and $\mathbf{y}$.]

4. **[12 marks]** Consider the optimization problem

$$\min_{\mathbf{w}} f(w_1, w_2) \tag{2}$$

where $f : \mathbf{R}^2 \to \mathbf{R}$ is the function in (1). For this question you should write a Python implementation of the gradient descent algorithm to solve problem (2). For all questions, use the following parameters

- A starting point $\mathbf{w}^{(0)} = [-12, 16]^T$.
- Run your code for 50 iterations.

For the plots in parts (a) and (b), the 'y'-axis should be in log scale.

(a) On the same set of axes, plot $f(w_1, w_2)$ vs k, where k is the iteration counter, when using gradient descent to solve (2) using the learning rates:
(i) $\alpha^{(k)} = \frac{1}{2.5}$; (ii) $\alpha^{(k)} = 1$; and (i) $\alpha^{(k)} = 0.01$.

(b) On the same set of axes, plot $\|\nabla f(w_1, w_2)\|_2$ vs k, where k is the iteration counter, when using gradient descent to solve (2) using the learning rates:
(i) $\alpha^{(k)} = \frac{1}{2.5}$; (ii) $\alpha^{(k)} = 1$; and (i) $\alpha^{(k)} = 0.01$.

(c) Write a few sentences describing the behaviour of gradient descent on problem (2) using each of the different learning rates.

(d) In the lecture notes for Lecture 6 on Learn, contour plots with the trajectory of the iterates (i.e., the points $\mathbf{w}^{(0)}, \mathbf{w}^{(1)}, \mathbf{w}^{(2)}, \ldots, \mathbf{w}^{(50)}$) for gradient descent were presented. Reproduce 3 similar plots for problem (2), one for each of the learning rates:
(i) $\alpha^{(k)} = \frac{1}{2.5}$; (ii) $\alpha^{(k)} = 1$; and (i) $\alpha^{(k)} = 0.01$.