CSE 326/426 (Fall 2022) Homework 9

Due on 11:55pm, Fri, Nov 11, 2022

Grading: All questions have the same points (10 each). Only questions with $*$ will be graded (undergraduates: do not need to answer the graduate-only question).

Submitting: Only electronic submissions in PDF format on Coursesite are accepted. Name your file as

<Your LIN>_HW9.pdf

A few ways to create a single PDF file. i) Use Microsoft Word and insert your writing or any figure taken, scanned, or plotted. Then choose "File→Print" in the main menu and you will find an option for outputting the file to a PDF file. ii) Use Latex to write your solution and include any figures. iii) Use the online Google Doc as an alternative of Word. It has sufficient features to combine multiple images and texts. Exporting to PDF files is similar as with Word.

Please DO NOT compress the PDF file as that will slow down the processing of your submission and the grading.

Questions:

1. $*$ Neural network architecture: given input feature vectors $\mathbf{x}^{(i)} \in \mathbb{R}^5$, we want to predict one out of three classes for each input. Design a neural network with 1 input layer, 1 hidden layer, and 1 output layer. Specify the following architecture parameters:
   - number of neurons of each layer (you have to select one for the hidden layer as well);
   - activation function $g^{[\ell]}$ for each layer $\ell = 1, 2$. Make sure to think about which activation function is suitable for predicting one out of more than two classes. You may need to refer back to the logistic regression note.
2. For the above network, list the weight and bias parameteters ($(W^{[\ell]}, b^{[\ell]})$ for $\ell = 1, 2$) necessary to define the neural network. Make sure you specify the number of rows and columns of each such parameter.
3. $*$ Given B training examples, feeding them through an MLP will generate B column activation vectors at the output layer, collected in the matrix $A \in \mathbb{R}^{n[L] \times B}$. Let $A^{[L-1]}$ be the matrix of B activation vectors at the penultimate (second last) layer. Let $Y \in \{0, 1\}^{n[L] \times B}$ be the ground truth label matrix, with the k-th column $\mathbf{y}^{(k)}$ being the one-hot label vector for the k-th training example. Prove that

$$\frac{\partial J}{\partial W^{[L]}} = \frac{1}{B}(A^{[L]} - Y)(A^{[L-1]})^\top \in \mathbb{R}^{n[L] \times n[L-1]} \qquad (1)$$

   is indeed the average of the gradients of the loss function J with respect to $W^{[L]}$ over the B training examples. Your answer should contain the gradients for each of the B training examples.

   (*Hints: first write down the (i, j)-th element of the matrix $\frac{\partial J}{\partial W^{[L]}}$ given above. Then for each training example, take the gradient of the loss function J evaluated on the k-th training example with respect to $W^{[L]}$ and inspect its (i, j)-th element. Finally, average the gradients over B training examples and check how that is the same as the (i, j)-th element of $\frac{\partial J}{\partial W^{[L]}}$. The backpropagation section of the lecture note will be useful.*)

4. MLP can also be used for regression on multiple target values. Given training examples $\{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}_{i=1}^m$ where $\mathbf{y}^{(i)} \in \mathbb{R}^k$ and $\mathbf{x}^{(i)} \in \mathbb{R}^n$. MSE loss will be appropriate for training an MLP to predict $\hat{\mathbf{y}}^{(i)}$, so that $\hat{\mathbf{y}}^{(i)}$ is as close as possible to $\mathbf{y}^{(i)}$ for the input $\mathbf{x}^{(i)}$. The MSE loss function will then be

$$J = \frac{1}{2m} \sum_{i=1}^{m} \|\hat{\mathbf{y}}^{(i)} - \mathbf{y}^{(i)}\|_2^2, \tag{2}$$

where $\|\mathbf{w}\|_2$ is the 2-norm of the vector $\mathbf{w}$. Find the partial derivative of J with respect to $\hat{\mathbf{y}}^{(i)}$. Is the softmax activation function appropriate for regression MLP? Justify your answer.

5. $*$ (Graduate only) This question is related to Project 3. In the function `explain` of the `NN` class, you will be asked to find $\partial z_c^{[L]}/\partial \mathbf{a}^{[0]}$, the gradient of $z_c^{[L]}$ with respect to the input data $\mathbf{a}^{[0]}$, where $z_c^{[L]} = W_c^{[L]} \mathbf{a}^{[L-1]}$ and $W_c^{[L]}$ is the c-th row of the parameter matrix $W^{[L]}$, and $\mathbf{a}^{[0]} \in \mathbb{R}^n$ is the input vector. Use the following hints to find $\partial z_c^{[L]}/\partial \mathbf{a}^{[0]}$.

   [*Hints:* first find the gradient $\partial z_c^{[L]}/\partial \mathbf{z}^{[L-1]}$, then follow the backprop algorithm to find the gradient $\partial z_c^{[L]}/\partial \mathbf{z}^{[\ell]}$ for all $1 \le \ell < L - 1$. Lastly, find $\partial z_c^{[L]}/\partial \mathbf{a}^{[0]}$ in terms of $\partial z_c^{[L]}/\partial \mathbf{z}^{[1]}$. Your final expression should be a chain of matrix multiplications and element-wise vector multiplications. To use your answer in Project 3, you will need to vectorize your final expression for multiple input vectors.]