Homework 01

ECE 469/568 – Machine Learning

SIU Southern Illinois University CARBONDALE

Date: 09/09/2024
Due data and time: 09/16/2023 by 11.59 PM
Section: Foundations of Machine Learning
Instructions: Solutions must include Matlab/Python codes, plots, and results
Submission format: Create a single PDF file for solutions/plots/descriptions
Upload instructions: Upload a single ZIP folder to folder Homework-1 in D2L
Codes: ZIP folder must also include all Matlab/Python codes as separate files.

[Q-1:] Select ALL correct choices. Every incorrect answer would earn a penalty of 1 point but the total marks of any multiple choice question will not be less than zero.

[1.1] Assume that the stochastic gradient descent algorithm is used to minimize the loss function $J(\mathbf{w})$, where $\mathbf{w}$ denotes the weights/parameters of the learning model. If α denotes the learning rate, m is the number of training data points, and $J_i(\mathbf{w})$ denotes the loss function for the ith training input, one performs the following update in each iteration:

(a) $\mathbf{w} \leftarrow \mathbf{w} - \alpha \nabla \sum_{i=1}^{m} J_i(\mathbf{w})$
(b) $\mathbf{w} \leftarrow \mathbf{w} - \alpha \nabla J_i(\mathbf{w})$
(c) $\mathbf{w} \leftarrow \mathbf{w} + \alpha \nabla J_i(\mathbf{w})$
(d) $\mathbf{w} \leftarrow \mathbf{w} + \alpha \nabla \sum_{i=1}^{m} J_i(\mathbf{w})$

[1.2] A certain regression learning model gives a large test error and small training error for a given data set. Which of the following would you prefer to use to resolve this issue?

(a) Use Adam optimizer to increase the efficiency of learning process.
(b) Use a learning model with a lower complexity.
(c) Use a learning model with a higher complexity.
(d) None of the above.

[1.3] Both the training and test errors for a certain regression learning model are high for a given data set. Which of the following would you prefer to use to resolve this issue?

(a) Use cross-validation.
(b) Use a learning model with a lower complexity.
(c) Use a learning model with a higher complexity.
(d) None of the above.

[1.4] Which statements are correct regarding optimizers in machine learning

(a) Batch gradient descent algorithm computes exact gradients before updating model parameters in each step.

(b) Adam optimizer uses both the momentum method and adaptive learning rate concept.

(c) AdaGrad does not use an adaptive learning rate

(d) Stochastic gradient descent computes gradients approximately during each iteration, and hence, there may be oscillations near the minimum.

[1.5] What are the techniques that can be used to prepare a data set to train a machine learning model?

(a) Remove redundant features and fill out missing features by using statistical methods.

(b) Use data standardization and normalization.

(c) Use one-hot encoding to convert numerical data into categorical format.

(d) Introduce outliers to improve the generalization of the machine learning model.

[Q-2:] A linear machine learning model can be written as

$$f(\mathbf{x}, \mathbf{w}) = \sum_{i=0}^{n} w_i x_i = \mathbf{w}^T \mathbf{x},$$

where $\mathbf{x} = [x_0, x_1, \cdots, x_n]^T$ is a vector of input features with $x_0 = 1$, and $\mathbf{w} = [w_0, w_1, \cdots, w_n]^T$ is a vector of weights or model parameters.

A loss function for such a linear model can be written as

$$J(\mathbf{w}) = \frac{1}{m} \sum_{i=1}^{m} \left[f(\mathbf{x}^{(i)}, \mathbf{w}) - y^{(i)} \right]^2,$$

where $\mathbf{x}^{(i)}$ is the feature vector of the ith example, and $y^{(i)}$ is the corresponding output of the ith example.

[2.1] Show analytically that the optimal weight vector that minimizes the above cost function $J(\mathbf{w})$ is

$$\mathbf{w}^* = \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{y}.$$

where $\mathbf{X}$ is a matrix obtained by concatenating $\mathbf{x}^{(i)}$ for $i \in \{1, \cdots, m\}$ as its rows, and $\mathbf{y} = [y^{(1)}, y^{(2)}, \cdots, y^{(m)}]^T$.

Discuss the drawbacks of this analytical solution when the size of the training data-set is very large.

[2.2] Develop a pseudo-codes for implementing the batch gradient descent, stochastic gradient descent, and mini-batch gradient descent algorithms to train the above linear model.

[2.3] Discuss the performance versus computational complexity of each of the above algorithms.

[Q-3:] In this question, you are required to use a software package that supports machine learning, preferably Keras, PyTorch and Scikit-Learn Python libraries/classes.

(a) Download the 'housing.csv' data-set from the following link and load it.
`https://github.com/ageron/data/tree/main/housing`

(b) Prepare the data by choosing the 'median_house_value' as the output and the rest as the input features.

(c) 'ocean_proximity' is a text attribute (categorical). You can either drop this feature or transform it into numerical values by using Scikit-Learn classes such as 'OneHotEncodee' or 'OrdinalEncoder'

(d) Clean the data by either dropping the missing values or replacing the missing values with the median. (hint: use SimpleImputer class in SciKit-Learn)

(e) Carry out feauture scaling either via normalization or standardization.

(f) Create a training data-set and a test data-set.