Artificial Neural Networks

Prof. Dr. Sen Cheng

Nov 04, 2019

Problem Set 5: Logistic Regression/ Classification

Tutors: Olya Hakobyan (olya.hakobyan@rub.de), Vinita Samarasinghe (samarasinghe@ini.rub.de)

Further Reading: Hands-on Machine Learning with Scikit-Learn and TensorFlow, Ch. 4

1. Derive the gradient of the loss function of logistic regression using paper and pencil.
2. Implement logistic regression model using only elementary programming operations. For your guidance, see the steps below:
   (a) Load the file '05_log_regression_data.npy' in numpy. It contains a hypothetical dataset, where the first two columns reflect the features: length of current residency and yearly income, and the last column contains the labels, i.e., whether a bank loan was granted to an individual or not.
   (b) Because the scale of the two features differs considerably, it is advised to standardize them before fitting. Import the *zscore* function from *scipy.stats* to standardize each of the features.
   (c) Using the *train_test_split* function from *sklearn.model_selection* and a training to validation ratio of 80:20, split the data into training and validation sets.
   (d) Use a scatter plot to visualize the training data set.
   (e) Implement the gradient descent to minimize the loss function.
   - Set the initial parameters $\theta_0 = 0$ or to a small random number.
   - Run the fitting process for 15,000 epochs and a learning rate $\eta = 0.001$.
3. Once the gradient descent algorithm is completed, plot the decision boundary together with the data. Remember that the decision boundary is given by $\hat{p}(x) = \sigma(\theta^T x) = p_0$. If $p_0 = \frac{1}{2}$, the previous condition is equivalent to
$$\theta^T x = \theta_0 + \theta_1 x_1 + \theta_2 x_2 = 0 \tag{1}$$
4. Plot the training loss vs the epochs. Why does the loss saturate? Why is the asymptotic value not zero?
5. The decision boundary can be shifted by adjusting p_0 to trade off the likelihood of the different errors. Apply different classification thresholds p_0 to obtain the precision-recall curve, F_1 score and ROC curve for the validation set. Based on these measures assess the model performance and determine what would be a reasonable classification criterion.