

IE406: Machine Learning Lab

Assignment 3

(Date: 12/09/2021)

1. In this question we will use the data given in file “Social_Network_Ads.csv” which is a categorical dataset to determine whether a user purchased a product or not by using three features to determine user’s decision. Visualize the data by 3D plotting features using different colors for label 0 and 1. Use data in files “Social_Network_Ads.csv” to perform logistic regression by implementing logistic function and with available library function and compare your results. Use 90% data points from each set for training and remaining 10% for testing the accuracy of classification. Using confusion matrix find accuracy, precision, F1 score and recall.
2. You will work with a widely used Iris dataset. The Iris Dataset contains four features (sepal length, sepal width, petal length, and petal width) of 50 samples of three species of Iris (Iris setosa, Iris virginica, and Iris versicolor). Plot features’ histogram. Compute pdf and compare it with histogram. perform the exploratory data analysis by plotting the basic statistics like mean, median, min, and max value of each feature (sepal and petal lengths and widths) for each of the three classes (setosa, virginica, and versicolor).
3. Visualize the data in the Iris Dataset by considering maximum combinations of two features in a 2D plot. Use red, green, and blue colors for labeling the three classes: Iris setosa, Iris virginica, and Iris versicolor, respectively. Comment on whether any two classes among the three can be separated by a line? Report your observations for each case.
4. Perform logistic regression on IRIS Dataset and plot confusion matrix. Using confusion matrix find accuracy, precision, F1 score and recall.

5. Imbalanced dataset typically refers to a dataset where the classes are not represented equally. Classification problems having multiple classes with imbalanced dataset present a different challenge than a binary classification problem. The skewed distribution makes the machine learning algorithms less effective, especially in predicting minority class examples.

In this question you will perform logistic regression for multiclass classification on the [20 News groups dataset](#). Since this dataset is a balanced one, you will perform the pre-processing to create an imbalanced version of the dataset (by removing some news articles from some groups). One example is given below. Perform multiclass classification using logistic regression on both the balanced and the imbalanced version of the dataset. Compare the performance in each case by obtaining the confusion matrix and accuracy. Report your observations at the end.

You can refer to [this](#) article for a better understanding of multiclass classification using logistic regression.

rec.sport.hockey	688	rec.sport.hockey	688
soc.religion.christian	599	rec.motorcycles	598
rec.motorcycles	598	rec.sport.baseball	597
rec.sport.baseball	597	rec.autos	594
sci.crypt	595	talk.politics.guns	546
rec.autos	594	talk.religion.misc	377
sci.med	594	sci.med	287
sci.space	593	sci.electronics	285
comp.windows.x	593	sci.space	197
comp.os.ms-windows.misc	591	sci.crypt	183
sci.electronics	591	misc.forsale	171
comp.sys.ibm.pc.hardware	590	comp.os.ms-windows.misc	151
misc.forsale	585	comp.graphics	146
comp.graphics	584	comp.sys.ibm.pc.hardware	137
comp.sys.mac.hardware	578	comp.windows.x	136
talk.politics.mideast	564	comp.sys.mac.hardware	131
talk.politics.guns	546	soc.religion.christian	86
alt.atheism	480	talk.politics.mideast	67
talk.politics.misc	465	alt.atheism	63
talk.religion.misc	377	talk.politics.misc	55

Balanced

Imbalanced

Note:

Submission Deadline: 23:59 hrs, Saturday, 18th September 2021

Strictly follow the submission guidelines given in the classroom.