CSCI 720 Big Data Analytics

Assignment 6

In this assignment, you will try Linear Regression and Linear classifiers on the adult income dataset.

Data

Download the dataset in adult income dataset from https://archive.ics.uci.edu/ml/datasets/Adult. The data set consists of 32561 training samples and 16281 test samples of adult American's demographic information. The output labels are Income <=50K and income >50K.

Preprocessing

The input features consists of both categorical and continuous features. You have already looked at this data for association rules mining. But there you needed to convert continuous features into categories. Here you would need to convert the categorical features to numerical values for analyses. Look at statistics to decide which categorical states to keep and which ones to combine. An extremely sparse feature matrix may result in a poor model. Relabel the output features to -1 for <=50K and +1 for >50K.

Analysis

You will train both regression and classification models on the data. You may use scikit learn in python or similar package.

Regression:

For interpreting accuracy of your regression model, you may use residual error (sum of squares) or pseudo classification accuracy by assigning a label +1 if output >0 else assign label -1.

1. Calculate accuracy for linear regression (Ordinary Least Squares). Interpret the coefficients of the regression model.

2. Now consider the effect of regularization using ridge regression (L2/quadratic regularizer). Explore the effect of regularization parameter, and discuss your results.

3. Repeat the above using Lasso (L1 regularizer). How do the coefficients of the regression model for Lasso compare with Ridge Regression?

Classification:

You will train Logistic Regression, Perceptron and Linear Support Vector Machine classifiers.

4. Report your results for the classifiers, in terms of accuracy, precision, recall and F1-score.
5. Plot ROC curves for the classifiers, and calculate the AUC (Area under curve)
6. Discuss the performance of the classifiers.

Include your results (include plots to support your assertions) as well any code you developed with your submission.