Homework 2

AM 218 Machine Learning

Due on March 30 at 11:59AM (noon) on Canvas

Part I: Probabilistic view of Lasso regression (15 pts)

Suppose the response Y is given by a deterministic function and an additive Gaussian noise

$$Y = \boldsymbol{\beta}^\top \mathbf{x} + \epsilon, \epsilon \sim N(0, \sigma^2)$$

where the parameters $\boldsymbol{\beta}$ has the Laplace prior, $\beta_i \sim \text{Laplace}(0, 1/\tau)$, i.e., $p(\beta_i) = \frac{\tau}{2}\exp(-\tau|\beta|)$.

Suppose a data set $(\mathbf{X}, \mathbf{y})$ is generated from the process described as above. Please show that the MAP estimator of the data set is equivalent to the corresponding Lasso regression estimator.

Part II: Implementation of Logistic Regression

Problem setting

Loans default will cause huge loss for the banks, so they pay much attention on this issue and apply various methods to detect and predict default behaviors of their customers. Your client company, a commercial bank, asked you to design a predictive model for them to predict the default behaviors of their future customers. The client company provides a set of data that includes the information of their previous customers and whether they defaulted on their loans. Specifically, each record indicates the age, the education level, the length of employment, the address, the income, the debt ratio, the debt on credit card, and other debts of one customer. Based on the information of the new customer, the model that you provide to the client company should predict whether this customer would default on the bank loan.

Data

In the dataset (see "bankloan.xls") provided by the client company, the first eight columns (age, edu, empl, addr, income, debt ratio, credit debt, other debt) indicate the information of customers.

The last column indicates whether the customer defaulted on the bank loan. There are in total 700 instances in this dataset.

Approach

You recently started to learn the course Machine Learning and realized that this is a binary classification problem. Based on the feature of the customer (age, education, etc.,), you could predict the default behavior (0 or 1) using a machine learning algorithm. You determine to implement Logistic Regression to help the client company to predict the default behavior of the customers.

Programming assignment

In this assignment, you will implement Logistic Regression on the bankloan dataset to predict the default behavior using Python. You can either use the functions from libraries such as scikit-learn or write the code from scratch to implement the algorithm. After you install Python on PC and run the sample code (see, "lr_demo.py") successfully (this step is not required, but recommended), you should apply the Logistic Regression following the instructions as below.

1. Learn and apply the algorithm (25pts)

   a. For each % in [5, 15, 50, 75, 85]

      i. Randomly sample % of the data for training

      ii. Use the remaining (1-%) of the data for testing

      iii. Learn a model from the training data and apply it to test data

      iv. Measure the performance on both the training and test data using accuracy

      v. Repeat 100 times with different random samples (using the same %)

      vi. Record and report the average and the standard deviation of accuracy on both training and test data across the 100 trials

   b. Plot curves for the results (training set size vs avg., training set size vs std. dev.) for both training and test data. Discuss the results.

2. Explore the effect of the threshold for the labeling decision (20pts)

a. Randomly sample 70% of the data for training

b. Use the remaining 30% of the data for testing

c. Measure the performance on the test data using recall and precision with different thresholds $h \in [0, 1]$, and discuss the results.

d. Plot the ROC curves for the logistic regression model that you trained and a baseline model using random guessing, and discuss the results.

Hint

1. For part 1, you may use `sklearn.linear_model.score()` to evaluation the accuracy.

2. For part 2, you may use `sklearn.metrics.precision_recall_curve()` to compute precision and recall. You can also use `sklearn.metrics.roc_curve()` to compute false positive rate and true positive rate and plot the ROC curves.

Submission

1. The source code in python.

   Name your file as "lr.py". You should include the code that you write for the above assignments. In addition, you should prepare the code such that TA can run without additional modifications and the output should report the results for the question 1 on the test data like this:

   `$python lr.py`

   `The average accuracy on test data:`

   `[0.977 0.394 0.102 0.304 0.904 0.811]`

   `The standard deviation of accuracy on test data:`

   `[0.053 0.026 0.021 0.018 0.015 0.032]`

   (Your output can be quite different from the above; this doesn't mean that your solution is wrong.)

2. Your evaluation & analysis in .pdf format.

   Note that your analysis should include the graphs as well as a discussion of results.

Useful references and libraries for Python

1. Andreas C. Müller, Sarah Guido, "Introduction to Machine Learning with Python: A Guide for Data Scientists", O'Reilly Media, 2016

2. Python tutorial: docs.python.org/3.7/tutorial

3. scikit-learn: scikit-learn.org

4. matplotlib: matplotlib.org

5. pandas: pandas.pydata.org

6. numpy: numpy.org