Computer assignment 5:1
Credit risk and machine learning

Introduction

In this assignment you will prepare and explore loan data and apply several machine learning models to predict probability of default and classify customers as good or bad. ML-algorithms:

- Logistic regression
- Decision trees
- Random Forest
- Neural networks

Data:

The data is in the file "loan_data.csv" and has 5960 observations and 13 columns:

BAD	LOAN	MORTDUE	VALUE	REASON	JOB	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC
1	1100	25860.0	39025.0	HomeImp	Other	10.5	0.0	0.0	94.366667	1.0	9.0	NaN
1	1300	70053.0	68400.0	HomeImp	Other	7.0	0.0	2.0	121.833333	0.0	14.0	NaN
1	1500	13500.0	16700.0	HomeImp	Other	4.0	0.0	0.0	149.466667	1.0	10.0	NaN
1	1500	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
0	1700	97800.0	112000.0	HomeImp	Office	3.0	0.0	0.0	93.333333	0.0	14.0	NaN

where "BAD" is the label column (1 = default), and:

LOAN: Loan amount

MORTDUE: Amount due on existing mortgage

VALUE: Value of the property

REASON: Reason for taking the loan

JOB: Occupational categories

YOJ: Years at present job

DEROG: Number of major derogatory reports

DELINQ: Number of delinquent credit lines

CLAGE: Age of oldest credit line in months

NINQ: Number of recent credit inquiries

CLNO: Number of credit lines

DEBTINC: Debt-to-income ratio

1. Explore the data*

For b-e), produce two bar plots, the first, a stacked bar plot, showing the percentages of non-defaulted/defaulted customers for each feature value (which for REASON are 'DebtCon' and 'HomeImp'), and the second, a regular bar plot, showing the number of customers corresponding to each value.

a) What is the default probability in the whole data set?
b) Type of job (JOB)
c) Reason for taking the loan (REASON)
d) Number of delinquent credit lines (DELINQ)
e) Number of major derogatory reports (DEROG)
f) Calculate the correlation between all numerical columns (including BAD) and plot it as a heatmap

Discussion: What conclusions do you draw from b-e? Based on f, which three features seem to be most important for predicting default?

2. Preprocess data and split into training and test sets*

a) Fill missing values:
   - For numerical columns (except YOJ, DEROG, DELINC, NINQ, CLNO): fill with the mean value of the column
   - For categorical columns and YOJ, DEROG, DELINC, NINQ, CLNO: fill with the most frequent value
b) Create dummy variables for the categorical columns (REASON and JOB)
c) Split the data into training and test sets by taking the first 4768 rows as training data and the rest as test data (an 80/20-split). Include the first 10 rows of each data set in the report.
d) Create normalized versions of the training and test sets to be used for logistic regression and the neural networks. Do not normalize the dummy columns. Include the first 10 rows of each data set in the report. To normalize you should subtract the mean and divide by the standard deviation (StandardScaler in Sci-kit learn). Note that normalization of the test set is to be done using the means and standard deviations from the training set.

** You can do this in Excel if you like*

3. Logistic regression

Fit a logistic regression classifier to the training set and report (for cut-off = 0.5):

a) For the training set: the accuracy, the ROC-curve, AUC and the confusion matrix
b) For the test set: the accuracy, the ROC-curve, AUC and the confusion matrix
c) Include the trained parameters and specify for each of the non-categorical features in 1.b-e) which effect it has on the probability of default, i.e. does it increase or decrease with the feature value? For the categorical features JOB and REASON, which job and which reason implies the lowest and highest predicted PD?

4. Decision trees

Fit a "medium"-sized decision tree classifier (in Matlab), in Python set "max_depth" = 4, to the unnormalized training set and report (for cut-off = 0.5):

a) For the training set: the accuracy, the ROC-curve (including logistic regression), AUC and the confusion matrix
b) For the test set: the accuracy, the ROC-curve (including logistic regression), AUC and the confusion matrix
c) Plot the decision tree (Matlab: in command window: `view(Model.ClassificationTree,'mode','graph'))`

Discussion: How does the decision tree compare to logistic regression?

5. Random Forest

Fit two random forest models (called "Bagged Trees" in Matlab's Classification Learner) to the unnormalized training set:

1. With 20 trees (in Matlab: set "Maximum number of learners" to 20 by clicking "Advanced" in the menu. In Python: also set "max_depth" = 8)
2. With 200 trees (In Python: also set "max_depth" = 8)

and report (for cut-off = 0.5):

a) For the training set: the accuracy, the ROC-curve (including the previous models), AUC and the confusion matrix
b) For the test set: the accuracy, the ROC-curve (including the previous models), AUC and the confusion matrix

Discussion: How do these two random forest models compare with each other, and the other models?

6. Neural networks

Fit two neural networks, both with a sigmoid output layer. Use a batch size of 1192 and determine the number of epochs to be used in final training in the following way:

- Split the training set into a training- and a validation set, where you use the first 3576 rows for training and the next 1192 rows for validation
- Train the model using 2000 epochs and find the number of epochs for which the validation loss attains its minimum
- Use the number of epochs, from above, to train the final model using the whole training set

Neural networks to train and evaluate (in Python: use the RMSprop optimizer with default settings):

1. With one hidden layer having 20 neurons and ReLU-activations
2. With two hidden layers, the first having 50 neurons and ReLu-activations, and the second having 20 neurons and ReLU-activations

Report (for cut-off = 0.5), for each of them:

a) Number of epochs used for training
b) For the training set: the accuracy, the ROC-curve (including the previous models), AUC and the confusion matrix
c) For the test set: the accuracy, the ROC-curve (including the previous models), AUC and the confusion matrix

Discussion: How do these neural networks compare with each other and to the other methods?

Grading

These are the requirements for the grades 3, 4 and 5:

Problem	3	4	5
1	All	All	All
2	All	All	All
3	All	All	All
4	All	All	All
5	-	one of 5 or 6	All
6	-	one of 5 or 6	All