FEM31002 & FEM21045 - Assignment 5

In this assignment, you are asked to estimate SVMs to predict credit card defaults. You will estimate these models using a preprocessed version of the "default of credit card clients data set", made public by Yeh and Lien (2009). Their dataset contains payment data of credit card holders of an 'important bank' in Taiwan in the year 2005. Each instance of this dataset consists of one binary response variable – default payment in October (Yes = 1, No = 0) - and a set of accompanying customer-specific variables. Descriptions of the customer-specific variables and the applied preprocessing can be found in table 1. The assignment consists of two parts outlined in the sections below.

1 The Main Task (12 pt)

For the first part of the assignment, we ask you to create three SVMs using different kernels of your choice that use customer-specific variables to predict whether or not an individual defaults in October 2005. Describe clearly and concisely how you performed the model estimation, what design choices you made, and why. In doing so, please adhere to the following requirements:

1. The three support vector machines you estimate should maximize the *accuracy* measure: the number of correctly labeled instances, both 0 and 1, over the total number of observations.[1]

2. When you are estimating and tuning your models, we ask you to optimize hyperparameters and evaluate the accuracy of the final models. This requires you to come up with a training, validation, and test strategy. It is up to you what strategy you choose, and as you have learned in the first week, there are many good choices. Note that you should apply your strategy to each of the three models separately!

3. Which or even how many hyperparameters to optimize is also up to you. What would be a reasonable number? Do you have reason to believe that some hyperparameters will affect the accuracy more than others?

4. You may assume that accurately predicting individuals that do not default is equally as important as accurately predicting individuals that do default.

[1] If this explanation is not clear, then a quick google search query should clarify any ambiguity!

Tip: the learning objective of this assignment is to get a better understanding of SVMs, and of the choices involved in applying machine learning algorithms in general. Your final accuracy is of less importance, so don't go overboard in terms of the number of evaluated hyperparameters and the complexity of your training, validation, and test strategy. We grade well-explained and argued approaches, conditional on your accuracy score not being unnecessarily low. Use your time efficiently!

2 Questions (8 pt)

For the second part of the assignment, we ask you to reflect on your results from the first part of the assignment by answering the following questions:

1. The estimation procedure requires you to maximize accuracy. Why is this a reasonable measure for the task in this assignment and the provided dataset? Give one example of altered assignment instructions, or dataset characteristics for which this would not be a reasonable measure to use.
2. What, if any, statistical evidence do you find to indicate a difference in the predictive performance among the three SVMs?
3. Can you explain the outcome of the previous question when you consider the data transformations that your kernel choices imply?

The report

In your report, we want you to discuss the two parts of the assignment separately.

1. *The main task*: describe your methodology and main findings. Please refer to the main task description to determine what information you should include.
2. *Answers to the questions.* Write your answers to the three questions separately in a numbered list. You can refer to tables or other results from the main task part of your report when answering the questions.

Please adhere to the following (formatting) rules:

1. Include a cover page with your team-number, member names and student numbers
2. Use a professionally looking easy-to-read font-type. Use font size 12.
3. Page margins should be 2.5cm (1 inch) on all sides.
4. No more than 2 pages, excluding only the cover page and the references list (if used).

Please upload your code to CodeGrade using the assignment in Canvas. You will not receive a grade for your code, we only use it to check for plagiarism.

Table 1: Individual variable and normalization descriptions.

Var.	Description	Month	Meaning / unit
X1	Credit Limit	-	x 1000 NT Dollar
X2	Gender	-	1 = male; 2 = female
X3	Education	-	1 = graduate school; 2 = university; 3 = high school; 4 = others.
X4	Marital status	-	1 = married; 2 = single; 3 = others.
X5	Age	-	year.
X6	Normalized payment status	September	-1 = pay duly; n = payment delay for n months;
X7		August	
X9		Juli	
X8		June	
X10		May	
X11		April	
X12	Normalized bill statement	September	Bill statements were normalized by dividing them by X1 x 1000. The normalized bill statement can therefore be interpreted as the fraction of the credit card's limit that is being used.
X13		August	
X14		Juli	
X15		June	
X16		May	
X17		April	
X18	Normalized payment	September	Montly payment (of the previous month) were normalized as above.
X19		August	
X20		Juli	
X21		June	
X22		May	
X23		April	

2.1 References

Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473-2480.