Erasmus University Rotterdam

Erasmus School of Economics
Quantitative Finance

Machine Learning - Assignment 5

Students:
Gerben van der Schaaf (416661)
Adnaan Willson (428043)
Casper Witlox (426233)

Group: 7

Instructors:
Prof. Dr. Dick van Dijk
Prof. Dr. Ilker Birbil
MSc. Utku Karaca
MSc. Karel de Wit

October 4, 2020

Main Task

We created three SVMs with three types of kernels: linear, polynomial and radial. We estimate the model with the help of an SVM package. By implementing a k-fold cross-validation, we try to avoid overfitting and obtain a more accurate estimate of the prediction performance of the models. The k is set equal to five, since there were only minor to no alterations when added more folds in terms of results. Furthermore, the computation time is quite significant, by reducing the number of folds the cross-validation takes less time to compute. The large data set is randomly divided into two parts. The first is used as an initial training data set, and the latter is used as test data. The initial training data is further divided into k equal sized subsamples. By using $k-1$ of these subsamples as training data in the cross-validation and 1 subsample as validation, we end up with estimating our hyperparameters k times per technique. Averaging over these results produce a fairly accurate estimate of the prediction accuracy over the original training data. We choose the model with corresponding hyperparameters values that generates the best accuracy score. Then, this model is fitted on the entire initial training data set and eventually evaluated with the original test data. This last part produces the predictions that we will use to obtain the accuracy score of a model.

Our first model is a SVM with a linear kernel, which has the following structure: $K(x_i, x_j) = x_i^T x_j$. In this environment, we choose to only variate the regularisation parameter C. This parameter determines to what extent the created hyperplane correctly separates the instances. Where as a small C generates a hyperplane with a large minimum margin. This gives a trade-off regarding the value of C. We find an optimal value of $C = 0.16$. Because a linear kernel is quite straightforward, we did not alter any different parameters.

The second model is a SVM with a polynomial kernel, which has the function: $K(x_i, x_j) = (1 + \gamma x_i^T x_j)^d$. We observe the new parameters γ and d, where the latter can only be an integer, on top of the original regularisation parameter C. The optimal values for these parameters are $C = 2.85$, $d = 3$ and $\gamma = 1.00$. The value for d is also the default value for an SVM with a polynomial kernel. Alterations of this hyperparameter causes large changes to the value K. Therefore, we assume that the degree of the polynomial kernel is generally optimal when it is set equal to three. The recommended value for γ is $\frac{1}{n_{feat} * var(X_{TR})} \approx 0.991$, so there is only a minor alteration here. Large values for γ and d were not incorporated due to time-management reasons. This causes the spectrum of possible optimal values to be limited.

The third and final model is a SVM with a radial kernel, which has the following structure: $K(x_i, x_j) = e^{-\gamma ||x_i - x_j||^2}$. By altering the hyperparameters C and γ, we find optimal values $C = 2.51$ and $\gamma = \frac{1}{n_{feat} * var(X_{TR})} \approx 0.99$.

Across our models, we optimise the hyperparameters C, d and γ. Since our data set is standardised, we do not alter the $coef_0$ hyperparameter in any of our models. For our linear kernel, we state that only the regularisation parameter C influences the accuracy. For the polynomial kernel, we argue that the hyperparameter d affects the accuracy more than γ and C. Especially, since γ functions similar to a multiplier, while d determines the degree of the polynomial function. The latter influences the K function value of the kernel

much more than the former. And for the radial kernel, we argue that γ influences the accuracy more than C. We noticed that small changes of γ had large influences on the accuracy, while fluctuations in the value of C resulted in minor changes to the accuracy.

Furthermore, optimising too many hyperparameters can take an substantial amount of computational time such that it is no longer beneficial to the optimisation problem. It is difficult to state a reasonable number since it is quite problem, and especially model, dependent. Generally, in the field of SVMs involving kernels, it could be best to optimise a maximum of two or three hyperparameters. The accuracy's, MSE and McNemar's test McNemar (1947) can be observed in Table 1.

Table 1: Performance characteristics of the SVMs with various types of kernels.

	Linear	Polynomial	Radial
Accuracy	0.683	0.690	0.697
MSE	0.317	0.310	0.303

Questions

1. Maximising accuracy is a reasonable measure because we deal with a classification problem regarding credit default where the loan either defaults or not. We follow the definition of default used in Yeh and Lien (2009). Banks are, more than anything else, interested whether a credit card defaults or not. Since defaults means that they lose money. By predicting these defaults, they can anticipate the consequences. Therefore, they need the predictions to be as accurate as possible, which makes it the main task in this assignment.

The data set of defaults is binary, one when a default payment occurs and zero when it does not. This type of data set is quite inviting towards an prediction because a prediction is either a yes or a no. And the most straightforward prediction evaluation is an accuracy measure.

When the data set of defaults would be a nominal variable with several categories. For instance, when there is a probability of default present and various groups are categorised based on their default risk. The bank should then assign each category its own protocol which must differ from each other. A basic binary accuracy measure is in that environment not correct.

2. We introduce the hypothesis that the predictions performances of the three models do not differ from each other: $H_0 : \hat{y}_{LIN} = \hat{y}_{POL} = \hat{y}_{RAD}$. The McNemar's test of McNemar (1947) is a non-parametric statistical test for paired comparisons that can be applied to compare the performance of two machine learning classifiers. We find a p-value of 0.000 when comparing models one and two, 0.000 when comparing models one and three and 0.246 when comparing models two and three. This gives us a reason to assume that a SVM with a linear kernel has a significantly different predictions performance than the other two models. Furthermore, we find

no statistical evidence that a SVM has different prediction performance with a polynomial kernel or a radial kernel, even at the 10% significance level.

3. ....

References

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Yeh, I.-C. and 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.