COMP3009 Machine Learning Coursework 3

Support Vector Machine

Runzhuo Li	psyrl6@nottingham.ac.uk
Xixuan Wang	scyxw3@nottingham.ac.uk
Zike Li	scyzl2@nottingham.ac.uk
Yu Zhang	scyyz4@nottingham.ac.uk
Pinyuan Feng	scypf1@nottingham.ac.uk

School of Computer Science
University of Nottingham

December 18th, 2020

Abstract

This is the third coursework of Machine Learning COMP3009 on the topic of Support Vector Machine (SVM). We trained, evaluated, optimised SVM models on Iris dataset[1] and Beijing PM2.5 dataset[2]. Besides, the performances of Artificial Neural Network (ANN), Decision Tree (DT) and SVMs with 3 kernel functions were compared based on a statistical approach. In this report, the details of the comparison will be demonstrated, including the experiment configurations and result analysis. All the experiment results are shown in the appendix.

SVM Training

The Initial Attempt

In this part, we tried different values of 'Epsilon' for SVM-linear with box constraint = 1 to see what would happen in terms of regression model performance. We found that with the increase of 'Epsilon', less support vectors would be selected, which resulted in lower performance but faster computation time. Besides, relatively small 'Epsilon' might lead to overfitting. Overall, the value of epsilon determines the level of accuracy of the approximated function.

Model Optimisation

We have implemented SVMs using three kernels (linear kernel, Gaussian RBF kernel, Polynomial kernel) for classification and regression tasks. To optimising SVM models, the optimal hyper-parameters were found based on inner-cross-fold validation. The results are shown in the Table 1 and Table 2. Also, how many support vectors were selected for each model are presented both in absolute terms (# of SVs) and the rate of the training data available (% of TD). Next, using inner cross validation to obtain optimal hyper-parameters, the models with the largest accuracy score value and the least RMSE value were selected for the classification and regression task, respectively.

The optimal hyperparameter combination for classification problem for each kernel machine:

- Kernel Function: linear, Box Constraint: 0.5132
- Kernel Function: rbf, Box Constraint: 0.9212, Kernel Scale: 110
- Kernel Function: polynomial, Box Constraint: 0.5132, Polynomial Order: 0.7077

The optimal hyperparameter combination for regression problem for each kernel machine:

- Kernel Function: linear, Box Constraint: 0.3077, Epsilon: 0.0579
- Kernel Function: rbf, Box Constraint: 0.3077, Epsilon: 0.0424, Kernel Scale: 213
- Kernel Function: polynomial, Box Constraint: 0.4219, Epsilon: 0.0531, Polynomial Order: 0.9309

Methodology Comparison

10-fold cross-validation was implemented to compare ANN, Decision Tree, and SVM, where accuracy and RMSE are the metrics for the classification and regression problems respectively. The built-in function TTEST2 in MATLAB was leveraged to see whether those approaches are significantly different from each other in a statistical manner. The comparison results are presented in Table 5 and Table 6.

Given that the P-Value of 0.0899 (> 0.05), it indicates that the results of Decision Tree and ANN are not significantly different, which is not the case for SVM-linear & ANN, SVM-linear & Decision Tree, SVM-RBF & ANN, SVM-RBF & Decision Tree, SVM-Poly & ANN, SVM-Poly & Decision Tree, whose P-Values are 3.329e-04, 3.0843e-05, 8.1750e-06, 6.9558e-07, 3.3279e-04 and 3.0843e-05, respectively.

In the regression task, it can be seen that the decision tree is not significantly different from ANN, since the P-Value (0.0773) is larger than the default significance level 0.05. On the contrary, the SVM-RBF & ANN, SVM-RBF & Decision tree pairs are proved to be significantly different from each other, whose P-values are larger than significance level (0.05).

Answers to Questions on the coursework sheet

Answer to Question 1:

To understand the role of the kernel parameter (sigma) in the Gaussian RBF kernel, it is necessary to observe its relationship with other variables in the equation that defines the kernel:

$$k(x, x') = \exp\,(-\frac{\|x - x'\|^2}{2\sigma^2})$$

Like the classical Gaussian distribution, the sigma represents the standard deviation, which determines the "width" of the distribution. In the case of the Gaussian kernel, this affects the type of decision boundary defined by SVM. More specifically, the sigma value determines the ‘influence’ or ‘reach’ of a single training example in the setting of the decision boundary.

The smaller the sigma is, the smaller the range will be, where only points close to the decision boundary will affect its shape. Conversely, as the sigma value increases, points farther from the boundary will also impact on the boundary itself. Taking the data points into account for larger sigma values far from the boundary will produce a smoother, more linear boundary, thus giving a more general classifier. This means that for overlapping data, the misclassification rate will increase, but overfitting will be avoided.

Answer to Question 2:

When a hard-margin SVM is applied to a dataset of two classes with overlapping features, its performance will be lower than expected. The hard-margin SVM aims to separate the data in a restricted manner. Because the features overlap, the approach will set a complicated (unsmooth) decision boundary to separate the training data. Although the fitting error is minimized based on training set, the generalization ability on testing set performs worse than expected.

Parameter C is a coefficient that punishes misclassified terms to minimize the data fitting error. If we need an SVM with hard margin, we will need to set this parameter as large as possible.

Answer to Question 3:

The result will be less significant in both situations. Assume that the data is shuffled so that there is no significant difference between any two folds. Under this assumption, the accuracy and RMSE are both aggregations of individual predictions. As the standard deviation of the normal distribution will be given $\frac{\sigma}{\sqrt{n}}$, where σ is the standard deviation of the prior distribution and n is the sample size, the sample size will reduce whenever sub-sampling occurs. And then, the standard deviation of the normal distribution will increase, resulting in a higher probability of observing an outlier. Specifically, even if the sample's mean does not change, the accuracy and RMSE probability of different models are more likely to come from the same distribution.

Conclusion

This report presents details about how to set hyper-parameters for different kernel machines using inner-cross-fold validation. The performances of ANN, decision tree, and SVMs were also compared, where we found that the decision tree performed the best for both of the classification and regression tasks.

Reference

[1] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[2] Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., Huang, H. and Chen, S. X. (2015). Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC and winter heating. Proceedings of the Royal Society A, 471, 20150257.

Appendix

Table 1. Binary classification results using Linear, Gaussian RBF and Polynomial Kernels

	Linear Kernel		***RBF Kernel***		***Polynomial Kernel***	
ID	***# of SVs***	***% of TD***	***# of SVs***	***% of TD***	***# of SVs***	***% of TD***
1	84	62.22%	97	71.85%	84	62.22%
2	88	65.19%	92	68.15%	88	65.19%
3	88	65.19%	88	65.19%	88	65.19%
4	89	65.93%	88	65.19%	89	65.93%
5	85	62.96%	90	66.67%	85	62.96%
6	89	65.93%	88	65.19%	89	65.93%
7	88	65.19%	90	66.67%	88	65.19%
8	87	64.44%	86	63.70%	87	64.44%
9	87	64.44%	86	63.70%	87	64.44%
10	96	71.11%	96	71.11%	96	71.11%
AVR	88.1	65.26%	90.1	66.74%	88.1	65.26%

Table 2. Regression results using Linear, Gaussian RBF and Polynomial Kernels

	Linear Kernel		***RBF Kernel***		***Polynomial Kernel***	
ID	***# of SVs***	***% of TD***	***# of SVs***	***% of TD***	***# of SVs***	***% of TD***
1	35386	89.77%	35497	90.05%	35393	89.78%
2	36526	92.66%	36532	92.67%	36509	92.62%
3	35930	91.15%	36042	91.43%	35932	91.15%
4	35012	88.82%	34816	88.32%	35014	88.82%
5	32987	83.68%	33096	83.96%	32980	83.66%
6	31604	80.17%	31755	80.56%	31602	80.17%
7	34493	86.23%	34368	87.18%	34505	87.53%
8	33993	87.50%	33927	86.07%	33992	86.23%
9	36347	92.20%	36419	92.39%	36342	92.19%
10	35440	89.90%	35356	89.69%	35429	89.88%
AVR	34772	88.21%	34781	88.23%	34770	88.20%

Table 3. Classification Accuracy of Binary Classification

ID	SVM-Linear	SVM-RBF	SVM-Poly	ANN	Decision Tree
1	0.800000	0.733333	0.800000	1.000000	1.000000
2	0.733333	0.533333	0.733333	1.000000	1.000000
3	0.800000	0.666667	0.800000	0.916667	1.000000
4	0.866667	0.600000	0.866667	0.916667	1.000000
5	0.666667	0.600000	0.666667	0.916667	1.000000
6	0.666667	0.733333	0.666667	0.833333	0.933333
7	0.733333	0.800000	0.733333	1.000000	1.000000
8	0.533333	0.466667	0.533333	1.000000	1.000000
9	0.733333	0.666667	0.733333	0.916667	0.933333
10	0.866667	0.866667	0.866667	1.000000	1.000000
AVR	0.740000	0.666667	0.740000	0.950000	0.986667

Table 4. RMSE scores of Regression

ID	SVM-Linear	SVM-RBF	SVM-Poly	ANN	Decision Tree
1	0.460163	0.283275	0.330789	0.270540	0.220671
2	0.429722	0.280633	0.314720	0.257165	0.220607
3	0.390977	0.257008	0.279465	0.241124	0.208532
4	0.392631	0.282797	0.274842	0.211496	0.216336
5	0.366106	0.280248	0.299106	0.243783	0.220289
6	0.375666	0.291544	0.308745	0.230122	0.256271
7	0.484488	0.378012	0.374870	0.238475	0.283582
8	0.432731	0.314075	0.341080	0.278366	0.236332
9	0.353438	0.258331	0.328315	0.266229	0.196732
10	0.410577	0.298332	0.322476	0.268360	0.248983
AVR	0.409650	0.292425	0.317441	0.250566	0.230833

Table 5. T-Test comparison for binary classification task

	ANN		***Decision Tree***	
	Significantly Different?	***P-Value***	***Significantly Different?***	***P-Value***
Decision Tree	No	0.0899	NaN	
SVM-Linear	Yes	3.3279e-04	Yes	3.0843e-05
SVM-RBF	Yes	8.1750e-06	Yes	6.9558e-07
SVM-Poly	Yes	3.3279e-04	Yes	3.0843e-05

Table 6. T-Test comparison for regression task

	ANN		***Decision Tree***	
	Significantly Different?	***P-Value***	***Significantly Different?***	***P-Value***
Decision Tree	No	0.0773	NaN	
SVM-Linear	Yes	3.1799e-09	Yes	1.0568e-09
SVM-RBF	Yes	0.0042	Yes	2.6325e-04
SVM-Poly	Yes	3.1799e-09	Yes	1.0568e-09