CAP 6610 Project 2

Group 8
Rachit Ranjan, Jun Deng, Zhitao Liu, Pu Fang, Ashutosh Garg

1. Member Roles

Table 1. Member Roles

	TrainMyClassifier + Classification Algorithms	MyCrossValidate	TestMy Classifier	MyConfusionMatrix
Rachit Ranjan	1			
Ashutosh Garg	2			
Zhitao Liu		3		
Pu Fang			4	
Jun Deng				5

2. Experiments

2.1 Settings

Follow the README.md in the submitted code(Klassifier.zip)

2.2 Confusion Matrices

Perform cross validation on a sample of size 5000. We get the following results. The size of each estimation set is 4000 and the size of each validation set is 1000.

Table 2. Confusion Matrices for SVM

	Class 1	Calss 2	Class 3	Class 4	Class 5	Unseen
Class 1	683	7	7	1	2	0
Class 2	3	736	2	15	0	0
Class 3	7	0	656	1	17	0
Class 4	4	13	4	741	6	0
Class 5	7	0	7	6	681	0
Unseen	0	0	0	0	0	1394

Table 3. Confusion Matrices for RVM

	Class 1	Calss 2	Class 3	Class 4	Class 5	Unseen
Class 1	560	11	67	8	54	0
Class 2	112	402	23	196	17	6
Class 3	61	26	388	101	103	2
Class 4	70	78	30	560	27	3
Class 5	48	0	36	105	504	8
Unseen	0	0	0	0	0	1394

Table 4. Confusion Matrices for GPR

	Class 1	Calss 2	Class 3	Class 4	Class 5	Unseen
Class 1	681	7	7	2	3	0
Class 2	7	715	2	30	1	1
Class 3	16	2	643	2	18	0
Class 4	12	27	7	712	9	1
Class 5	8	0	9	16	668	0
Unseen	0	0	0	0	0	1394

Table 5. Number of Support Vectors & Relevance Vectors

	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4
Support Vectors	2070	2083	2042	2042	2061
Relevance Vectors	6	6	6	6	6

Table 6. Classifier Rankings

Ranks	Classifier	Average Accuracy
1	Support Vector Machine	**0.97709**
2	Gaussian Process Regression	0.95438
3	Relevenve Vector Machines	0.70473

3. Obervations

In this project, we get the best results with SVM. The accuracy of Gaussian Process is slightly lower than SVM with RVM taking the last spot.

As we can see from the table 7, the complexity of RVM and GPR is relatively high, both cubic to the number of rows in the dataset. The number of features in the given dataset is 60. So for SVM, D is 60. The complexity of SVM is O(max(N,60) min (N,60)^2). When the size of training dataset or testing dataset is bigger than 60. SVM is better than RVM and GPR. When N grows bigger, The complexity is asymptotically similar to O(N). So for large dataset, SVM is significantly faster than RVM and GPR.

During cross validation, we give multiple candidates for each parameter, and we would choose the combination of parameters that give the best results. So the time for cross validation is linear to the number of permutations of parameters. The space required to build the model is also taken into consideration. The space taken by SVM and RVM models is acceptable. However, for Gaussian Process Regression, the size of the model is quadratic to the size of the training dataset, which makes it unsuitable for large dataset.

Table 7. Classifier Computational Complexity

Algorithm	SVM	RVM	GPR
Complexity	O(max(N,D) min (N,D)^2)	O(N^3)	O(N^3)

References

1. Chapelle, O. (2007). Training a support vector machine in the primal. Neural computation, 19(5), 1155-1178.

2. Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. Journal of machine learning research, 1(Jun), 211-244.

3. Rasmussen, C. E. (2004). Gaussian processes in machine learning. In Advanced lectures on machine learning (pp. 63-71). Springer, Berlin, Heidelberg.