Project 2 Report

Team 10: Bhavesh Poddar, Nanda Kishore Sunki, Vineeth Chennapalli, Chuji Luo

1 Member Roles

	TrainMyClassifier			MyCrossValidate	TestMyClassifier			MyConfusionMatrix
	RVM	SVM	GPR		RVM	SVM	GPR	
B. Poddar		1				1		
N. K. Sunki			1				1	
V. Chennapalli	1				1			
C. Luo				1				1

2 Observations

All the three algorithms were trained using all-pairs method. Because of time limit, we couldn't train GPR on entire data. While we observed on 5000 samples subset to compare the performances of all classifiers, we trained SVM and RVM on entire data. NaNs in the observation tables represent no training sample from unseen class.

2.1 Support Vector Machines

Table 1: Overall Confusion Matrix (5000, 25000 samples respectively)

	C_1	C_2	C_3	C_4	C_5	Unseen		C1	C2	C3	C4	C5	Unseen
C_1	0.99	0	0	0	0	0		0.99	0	0	0	0	0
C_2	0.01	0.96	0	0.02	0.00	0.01		0	0.97	0	0.02	0	0
C_3	0	0	0.98	0	0	0.1		0.01	0	0.98	0	0.01	0
C_4	0	0.02	0	0.96	0.01	0.01		0	0.01	0	0.97	0.01	0.01
C_5	0	0	0	0.01	0.98	0		0	0	0	0	0.99	0
Unseen	NaN	NaN	NaN	NaN	NaN	NaN		NaN	Nan	Nan	Nan	Nan	Nan

Overall Accuracy: 97.29% (5000 samples), 98.18% (25000 samples)

Average Support Vectors for each fold: 360 (5000 samples), 1178 (2500 samples)

Observation: 1. This classifier trained quickest among all the 3 classifiers without compromising on quality.

2.2 Relevance Vector Machines

Number of Support Vectors per fold averaged over all 10 pair classifiers: 1600 (5000 training set), 8000 (25000 training set).

Observations: 1. Pretty robust performance even after training on noise/random data.

2. Utilizing PCA was hampering performance of this algorithm. It was vital to use full data.

3. As the training dataset size increased, the main class probabilities fell down around 5-8%.

Table 2: Overall Confusion Matrix for RVM (5000, 25000 respectively)

	C_1	C_2	C_3	C_4	C_5	Unseen		C_1	C_2	C_3	C_4	C_5	Unseen
C_1	0.97	0.01	0.007	0	0.005	0.002		0.97	0.011	0.008	0	0.003	0.007
C_2	0.01	0.93	0.005	0.05	0	0.001		0.008	0.86	0.003	0.12	0.001	0.007
C_3	0.008	0.002	0.94	0.007	0.036	0.004		0.015	0.002	0.83	0.004	0.14	0.008
C_4	0.004	0.035	0.008	0.93	0.02	0.004		0.009	0.08	0.007	0.85	0.03	0.02
C_5	0.005	0	0.022	0.027	0.94	0.004		0.012	0.001	0.037	0.029	0.91	0.009
Unseen	NaN	NaN	NaN	NaN	NaN	NaN		NaN	NaN	NaN	NaN	NaN	NaN

Overall Accuracy: 94.71% (5000 samples), 88.42% (25000 samples)

2.3 Gaussian Process Regression

Table 3: Overall Confusion Matrix for GPR.

	C_1	C_2	C_3	C_4	C_5	Unseen
C_1	1	0	0	0	0	0
C_2	0	0.96	0.01	0.03	0	0
C_3	0.04	0	0.89	0	0.07	0
C_4	0	0.01	0	0.95	0.03	0
C_5	0.01	0	0	0	0.98	0
Unseen	NaN	NaN	NaN	NaN	NaN	NaN

Overall Accuracy: 95.63% (5000 samples)

3 Conclusion

1. SVM was best in terms of accuracy and fastest among all. GPR is slowest.
2. Based on experiments performed on RVM, we understood that it was pretty robust.