COMP – 8740 Machine Learning & Pattern Recognition – Fall 2023
Assignment 2

Ferruccio Sisti	Muhammad Haseeb Ahmad	Guangrui Guo	Siddharth Samber
104807246	110123184	000047362	110124156

Pledge: "As a student of the University of Windsor, I pledge to pursue all endeavors with honor and integrity, and will not tolerate or engage in academic or personal dishonesty. I confirm that I have not received any unauthorized assistance in preparing for or writing this assignment. I acknowledge that a mark of 0 may be assigned for copied work." Muhammad Haseeb Ahmad 110123184

Section I: Introduction

In this assignment, we run three different classifiers using SVM with the following kernels and their parameters: (a) SVM-L: linear kernel; (b) SVM-P: polynomial on six datasets (circle0.3, moons1, spirial1, twogaussians33, twogaussians42 and halfkernal). We first run the three classifiers with default parameters using 10-fold cross-validation, obtaining, for each classifier, the averages of the five measures of efficiency seen in class: PPV, NPV, specificity, sensitivity, accuracy, where class 1 corresponds to "positive" and class 2 to "negative" kernel – degree 2; (c) SVM-R: RBF. We create three SVM classifiers with different kernels: (a) **svm_linear** uses a linear kernel with the kernel='linear' parameter. (b) **svm_poly** uses a polynomial kernel of degree 2 with the kernel='poly' and degree=2 parameters. (c) **svm_rbf** uses the radial basis function (RBF) kernel with the kernel='rbf' parameter. We fit each classifier to the training data and calculate their accuracy on the testing data. Then, for SVM-R, we plot the ROC curve and find the AUC for each dataset. Finally, we apply grid search to obtain the best parameters you can based on accuracy.

Section II: 10-fold cross-validation (a)SVM-L, (b) SVM-P, and (c)SVM-R (Default Parameters)

Results and Plots

Below are the results obtained from the SVM model using different parameters and kernels on different datasets.

Table 1: SVM-L: Accuracy, PPV, NPV, Specificity, and Sensitivity

	Circle0.3	Moons1	Spirial1	Twogaussian33	Twogaussian42	Halfkernal
Accuracy	0.624	0.87	0.76	0.995	0.925	0.745
PPV	0.578	0.886	0.741	0.996	0.980	0.819
NPV	0.807	0.87	0.741	0.988	0.879	0.683
Specificity	0.326	0.888	0.742	0.996	0.982	0.868
Sensitivity	0.922	0.874	0.740	0.988	0.864	0.733

Table 2: SVM-P: Accuracy, PPV, NPV, Specificity, and Sensitivity SVM-polynomial kernel.

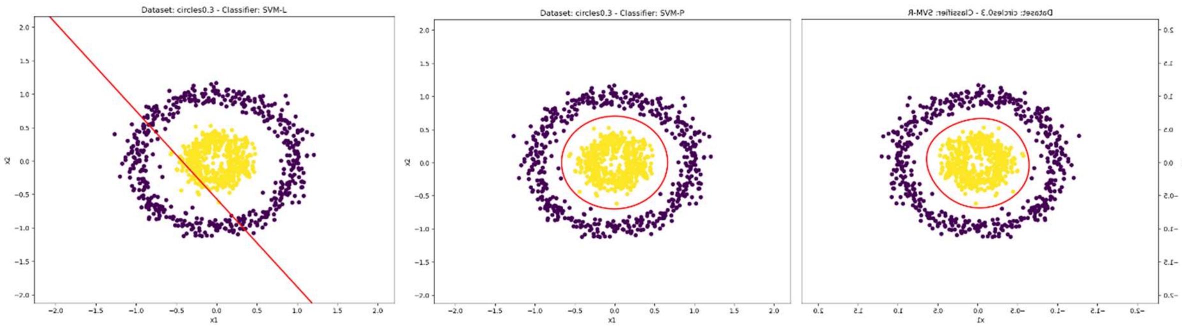
	Circle0.3	Moons1	Spirial1	Twogaussian33	Twogaussian42	Halfkernal
Accuracy	1.000	0.801	0.471	0.576	0.739	0.792
PPV	1.000	0.728	0.476	0.717	0.759	1.000
NPV	1.000	0.944	0.464	0.545	0.722	0.706
Specificity	1.000	0.640	0.376	0.898	0.778	1.000
Sensitivity	1.000	0.962	0.566	0.257	0.699	0.584

Table 3: SVM-R: Accuracy, PPV, NPV, Specificity, and Sensitivity

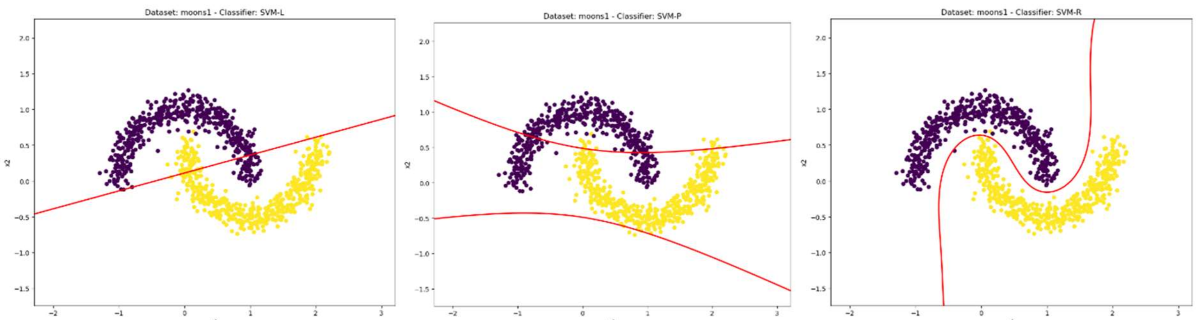
	Circle0.3	Moons1	Spirial1	Twogaussian33	Twogaussian42	Halfkernal
Accuracy	1.000	0.998	0.985	0.994	0.935	1.000
PPV	1.000	0.998	0.986	0.998	0.993	1.000
NPV	1.000	0.998	0.984	0.990	0.889	1.000
Specificity	1.000	0.998	0.986	0.998	0.994	1.000
Sensitivity	1.000	0.998	0.984	0.990	0.876	1.000

Below are the plots that contain the results of all experiments.

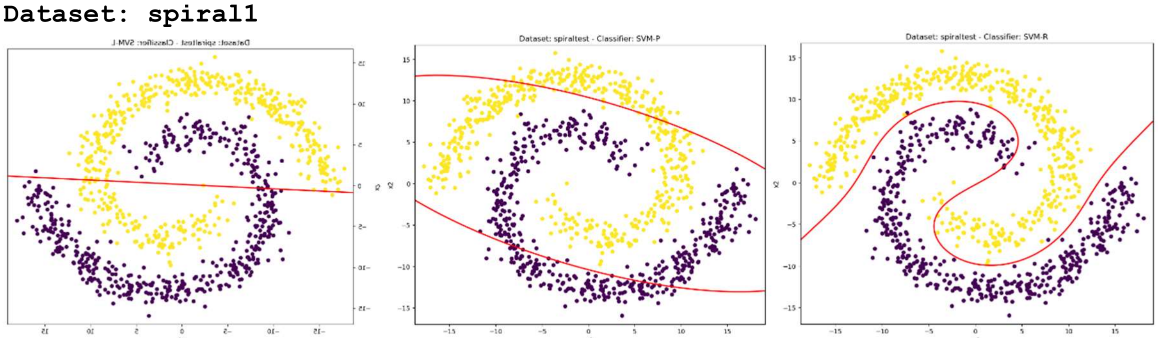
Dataset: circles0.3



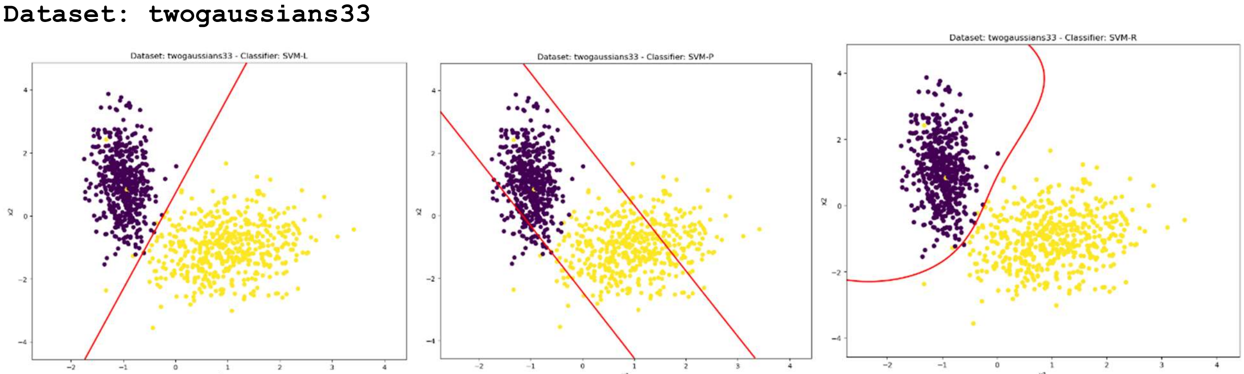
Dataset: moons1



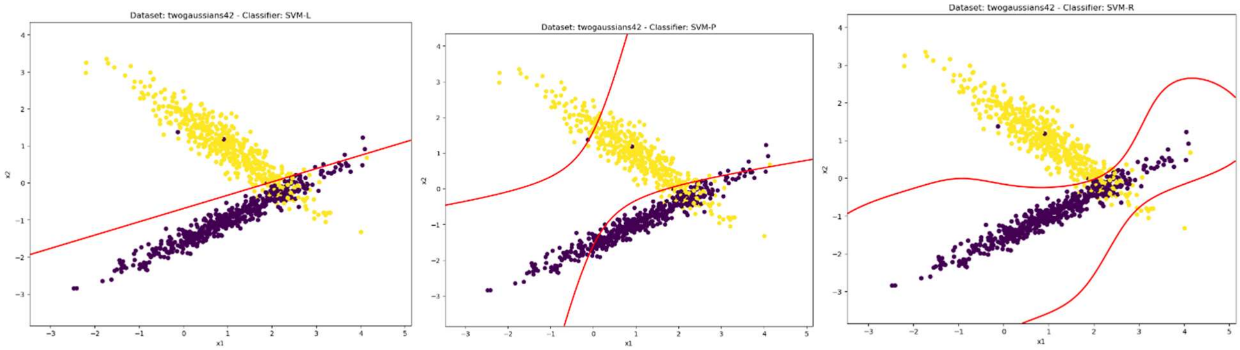
Dataset: spiral1



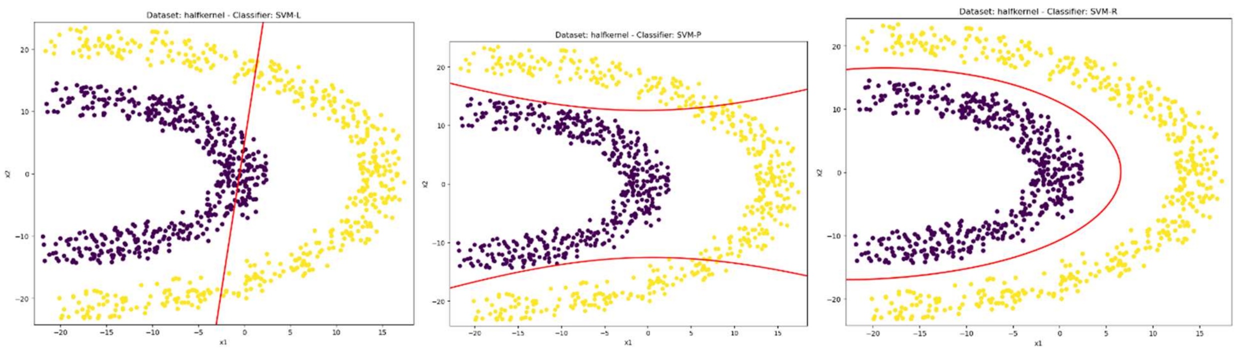
Dataset: twogaussians33



Dataset: twogaussians42



Dataset: halfkernel



Comparison, Comments, and Reasons:

Dataset: circles0.3: This dataset consists of circular clusters, and the decision boundary between the two classes is nonlinear. SVM-L struggles to capture this nonlinear boundary effectively, resulting in lower performance metrics. On the other hand, SVM-R and SVM-P can better fit nonlinear data, leading to perfect classification in this case.

Dataset: moons1: The "moons1" dataset consists of two crescent-shaped clusters, and the decision boundary between the two classes is nonlinear. SVM-R performs exceptionally well on this dataset due to its ability to handle complex nonlinear decision boundaries. SVM with Linear Kernel (SVM-L) also performs well because the linear boundary approximates the shape of the data reasonably well, while SVM with Polynomial Kernel (SVM-P) is less suitable for this dataset.

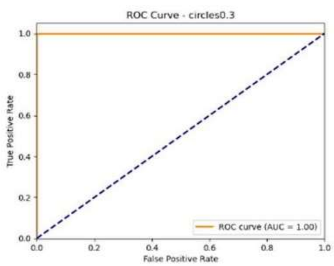
Dataset: spiraltest: The "spiraltest" dataset contains data points arranged in a spiral pattern, making it a complex nonlinear classification problem. SVM-R is well-suited for such complex datasets, as it can capture nonlinear decision boundaries effectively. This is reflected in its high-performance metrics. SVM-L can still perform reasonably well because it attempts to fit a linear boundary to the data, which, in this case, approximates the shape of the spiral to some extent. However, it may not capture the details as effectively as SVM-R. SVM-P may not be suitable for this dataset since it tries to fit a polynomial boundary to the data, which does not align well with the spiral pattern. This results in lower performance metrics, particularly in terms of specificity and accuracy.

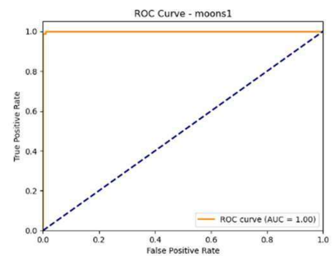
Dataset: twogaussians33: The "twogaussians33" dataset consists of two Gaussian distributions with well-separated means, making it relatively easy to separate the two classes. SVM-R and SVM-L are both suitable for this dataset, as they can effectively create linear decision boundaries to separate the clusters. SVM with Polynomial Kernel (SVM-P) may not be appropriate in this case because it attempts to fit a polynomial boundary to the data. This results in a less accurate separation between the clusters, especially in terms of sensitivity (recall), leading to lower overall performance.

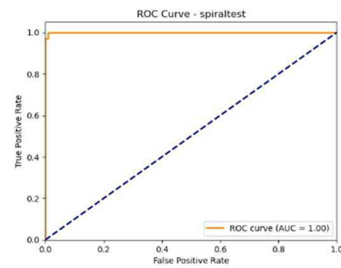
Dataset: twogaussians42: The "twogaussians42" dataset consists of two distributions with partial overlap, making it challenging to separate the two classes. SVM-R and SVM-L are suitable for this dataset because they can create decision boundaries to separate the Gaussian clusters effectively. SVM-P may not perform as well because it attempts to fit a polynomial boundary to the data, which is less accurate in separating the partially overlapping clusters, leading to lower performance in all metrics.

Dataset: halfkernel : The "halfkernel" dataset consists of two half-moon-shaped clusters, making it a relatively separable dataset**.** SVM-R and SVM -P are well-suited for this dataset because they can create curved decision boundaries to effectively separate the half-moon clusters. SVM-L may not perform as well because it attempts to create a linear decision boundary, which is less appropriate for this curved dataset, leading to slightly lower performance in all metrics compared to the other two classifiers.

Section III: SVM-R, ROC Curve Plots and the AUC

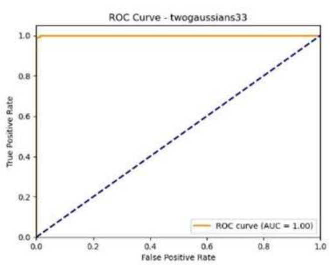


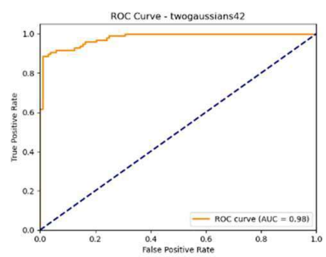


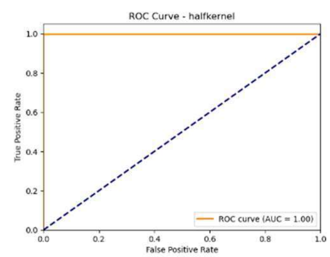


AUC for circles0.3: 1.00 AUC for moons1: 1.00 AUC for spiraltest: 1.00

Figure 1: Circle0.3 Figure 2: Moons1 Figure 3: Spiral1







AUC for twogaussians33: 1.00 AUC for twogaussians42: 0.98 AUC for halfkernel: 1.00

Twogaussain42: This one contains more overlapping of the two Gaussian distributions than Twogaussians33 does. As a result, AUC for twogaussians42: 0.98 instead of 1.00.

Section IV: Best Parameters

To find the best parameters for SVM based on accuracy, we can use grid search. Grid search involves trying a range of parameter values and selecting the combination that yields the highest accuracy. We use the **GridSearchCV** function from scikit-learn to perform this task.SVM. Below are the results.

	circles0.3	moons1	Spiral1	twogaussians33	twogaussians42	halfkernel
Best Kernel	RBF	RBF	RBF	RBF	RBF	RBF
Best Parameters:	'C': 0.1, 'gamma': 1	'C': 1, 'gamma': 1	'C': 0.1, 'gamma': 0.1	'C': 0.1, 'gamma': 1	'C': 10, 'gamma': 1	'C': 0.1, 'gamma': 0.01
Accuracy	1.00	0.99	0.99	0.99	0.94	1.00

Section V: Conclusion

SVM-R is the best for all six datasets because the RBF kernel is highly flexible and can capture complex, nonlinear relationships in the data, making it a suitable choice for a variety of dataset shapes and patterns.