

# CIS6060 Assignment 2: Performance of Logistic Regression and Support Vector Machines Classifiers on Three Datasets

J Moggridge

2021-04-06

*I chose to add the popular Pima Indians diabetes dataset to this analysis (<https://www.kaggle.com/uciml/pima-indians-diabetes-database>). These data were previously obtained and additional features were generated in R using an existing script (`make_diabetes_data.R`).*

---

## Introduction

The performance of logistic regression and support vector machines (SVM) classifiers were compared for three datasets: abalone (n = 4177, 7 variables, 3 classes); cancer (n = 79, 7129 genes, binary outcome, PCA transform was applied); and the Pima Indians diabetes dataset with interaction and quadratic terms (n = 768, 45 predictors, binary outcome). Performance metrics of interest (accuracy, precision, recall, and f1 score) were collected for hypothesis testing with paired t-tests. Two evaluations were performed for each classifier with each of the three datasets: a single validation set approach (table 1) and a 5-fold cross validation (table 2, fig. 1).

## Results and Discussion

In terms of dataset-specific performance, both classifiers achieved near-perfect accuracy on the cancer data (except the SVM in the simple validation), both had good performance on the diabetes data (~0.8), and both struggled with the abalone data (~0.55; tables 1 & 2). In evaluations with a single test-train split specifically, both algorithms had similar performance with the various datasets except with the cancer dataset, where logistic regression predicted the hold-out set without any error (table 1).

---

Table 1: Performance estimates of logistic regression and support vector machines classifiers on the Abablone, Cancer, and Diabetes sets from a single train-test split.

data	Classifier	Accuracy	Precision	Recall	F1
Abalone	Logistic regression	0.54	0.53	0.53	0.47
Abalone	SVM	0.6	0.59	0.6	0.59
Cancer	Logistic regression	1	1	1	1
Cancer	SVM	0.83	1	0.5	0.67
Diabetes	Logistic regression	0.74	0.76	0.51	0.61
Diabetes	SVM	0.73	0.74	0.51	0.6

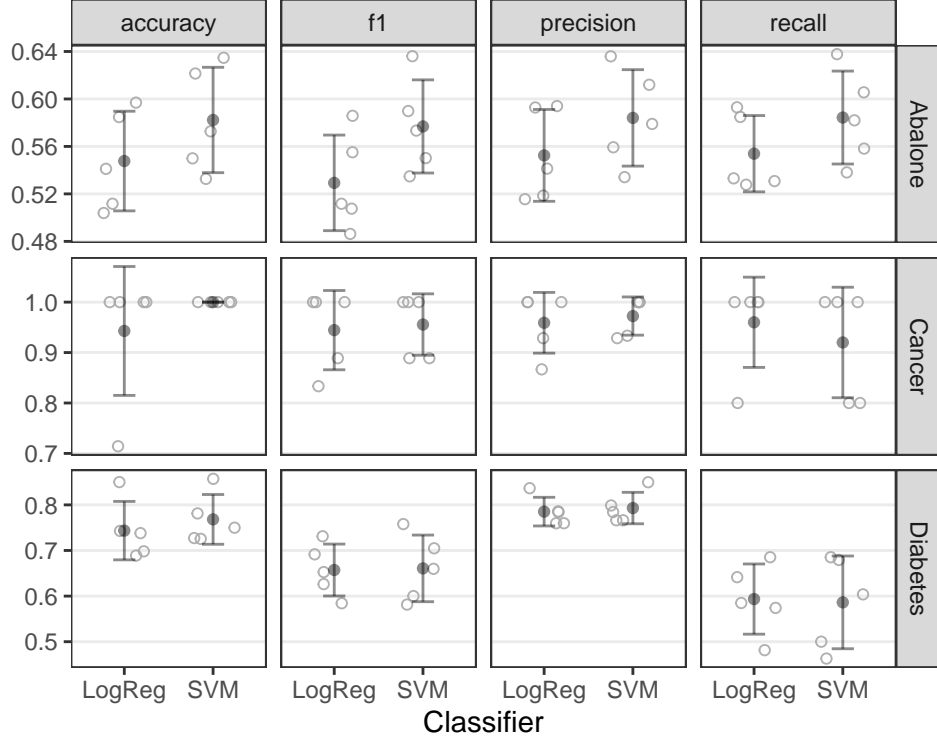


Figure 1: Performance metrics (accuracy, f1, precision, and recall) in 5-fold cross-validation of logistic regression and support vector classifiers for Abalone, Cancer, and Diabetes datasets. Means  $\pm$  sd are shown as black point and errorbars, with individual observations are shown as grey circles. Note that different scales are used for each dataset.

In evaluations with 5-fold cross-validation, I found that the support vector machine classifiers generally performed slightly better across datasets and metrics than the logistic regression classifier (fig. 1). However, according to paired t-tests the differences in performance metrics were only significant for the abalone dataset and in the case of precision on the diabetes data ( $p < 0.05$ ; table 1). Perhaps if we evaluated models with 10-fold cross validation, we may have found other significant results. The cancer dataset has only a few observations, making 10-fold cross-validation unfeasible, but bootstrapping is a possible alternative for re-sampling this dataset.

Interestingly, the optimal model hyperparameters found by grid-search cross-validation often varied among folds of a given dataset (see end of document for list). For the abalone data, the best SVM classifiers always used the radial basis kernel, suggesting that the classes are not separated well in linear space. With the cancer and diabetes datasets, the best SVM classifiers always used the linear kernel with the smallest regularization penalty ( $C = 1$ ), suggesting that these problems are more easily solved in linear space. The best logistic regression classifiers used the l1 regularization for the diabetes data, though a mixture of l1 and l2 norms were selected with cancer and abalone data. The logistic regression penalty hyper-parameter ( $C$ ) varied between nearly all folds, but usually only by a small amount. Conversely, the SVM models always performed better with strong regularization ( $C = 1$  or  $10$ ) suggesting that models can easily be overfit to these datasets.

Table 2: Performance of logistic regression and SVM classifiers applied to 3 datasets in 5-fold cross-validation. Data are expressed as mean (standard deviation)

Dataset	Metric	Logistic Regression	SVM	p_value	sig
Abalone	accuracy	0.548 (0.042)	0.582 (0.044)	0.033	*
Abalone	f1	0.529 (0.04)	0.577 (0.039)	0.012	*
Abalone	precision	0.552 (0.039)	0.584 (0.041)	0.028	*
Abalone	recall	0.554 (0.032)	0.584 (0.039)	0.016	*
Cancer	accuracy	0.943 (0.128)	1 (0)	0.374	
Cancer	f1	0.944 (0.079)	0.956 (0.061)	0.374	
Cancer	precision	0.959 (0.06)	0.972 (0.038)	0.374	
Cancer	recall	0.96 (0.089)	0.92 (0.11)	0.374	
Diabetes	accuracy	0.744 (0.064)	0.768 (0.055)	0.119	
Diabetes	f1	0.657 (0.057)	0.661 (0.073)	0.707	
Diabetes	precision	0.785 (0.031)	0.793 (0.034)	0.033	*
Diabetes	recall	0.593 (0.077)	0.586 (0.102)	0.726	

## Conclusion

In this work, I compared the performance of logistic regression and SVM classifiers on the cancer, abalone, and diabetes datasets. The abalone was the most challenging problem of the three and the cancer problem the easier. Logistic regression and SVM generally had similar performance on each dataset, with small but significant differences between the two on the abalone dataset, where the radial basis function kernel was slightly superior to logistic regression.

## Model hyperparameters selected in evaluations

### Abalone test train

-----

Logistic Regression: 'C': 0.0695, 'penalty': 'l1'  
SVN: 'C': 10, 'gamma': 0.3, 'kernel': 'rbf'

### Abalone 5 fold:

-----

Logistic Regression: 'C': 4.83, 'penalty': 'l2' (x 2)  
Logistic Regression: 'C': 0.021, 'penalty': 'l2'  
Logistic Regression: 'C': 1.438, 'penalty': 'l1'  
Logistic Regression: 'C': 0.038, 'penalty': 'l1'  
SVN: 'C': 10, 'gamma': 0.3, 'kernel': 'rbf' (x 4)  
SVN: 'C': 1, 'gamma': 0.1, 'kernel': 'rbf'

### Cancer test train

-----

Logistic Regression: 'C': 0.0207, 'penalty': 'l2'  
SVN: 'C': 1, 'kernel': 'linear'

### Cancer 5 fold:

-----

Logistic Regression: 'C': 4.833, 'penalty': 'l2'  
Logistic Regression: 'C': 0.0018, 'penalty': 'l2'  
Logistic Regression: 'C': 0.0036, 'penalty': 'l2'  
Logistic Regression: 'C': 100.0, 'penalty': 'l1'  
Logistic Regression: 'C': 0.001, 'penalty': 'l2'  
SVN: 'C': 1, 'kernel': 'linear' (all 5)

### Diabetes test-train

-----

Logistic Regression: 'C': 0.785, 'penalty': 'l1'  
SVN: 'C': 1, 'kernel': 'linear'

### Diabetes 5 fold:

-----

Logistic Regression: all l1 penalty, variable C (0.2-2.63)  
SVN: 'C': 1, 'kernel': 'linear' (all 5)