Written Homework 2 : Support Vector Machine

Title : Discovering Supports Vector Models for Dwelling Ownership Prediction: A Case Study of Data on Washington State Houses

Abstract

This report explores the application of Support Vector Machines (SVMs) in predicting dwelling ownership in Washington State using data obtained from the US Census via IPUMS USA. The dataset includes a wide array of variables relating to individuals and their housing situations, ranging from demographic characteristics to housing attributes. The primary objective is to develop predictive models capable of distinguishing between owner-occupied and renter-occupied dwellings based on a combination of factors pertaining to the occupants and the housing itself. This study employs three different SVM kernels - linear, radial, and polynomial - to build and evaluate predictive models. There were seven predictors used: “NCOUPLES” (Number of couples in the household), “NFAMS”(Number of families in the household), “ROOMS”(Number of rooms), “BEDROOMS” (Number of bedrooms), “VEHICLES” (Number of vehicles), “AGE” (Age of the person), besides “MARST”(Marital status). Of which, two of these were for individual-level variables namely “AGE” and “MARST” while the rest are housing-level variables. The first model was SVM with kernel type Linear. This model yielded an testing accuracy of 86.22% initially which then was improved to 86.56% after tuning the C(0.1) parameter to an optimal value found through Cross Validation employing 5 K fold. The second model was SVM with kernel type Radial. Initially this model yielded an testing accuracy of 83.20% which was notably increased to 86.22% after building with optimised C(1) and gamma(0.1) value. The third model was SVM with polynomial kernel type which with default hyper parameters gave a testing accuracy of 80.29%. Then the model was improved using tuned C(10) and Degree(2), there was a drastic improvement in the testing accuracy from 80.29% to 86.78%.The knowledge gained from this analysis helps in understanding housing trends and encouraging equal housing opportunities and more sustainable urban development that leads to economic advancement.

Introduction and Overview:

Support Vector Machines (SVMs) are powerful supervised learning models widely used for classification tasks. In this report focus is on the housing prediction using SVMs by emphasizing dwellings’ ownership classification in Washington State. The dataset utilized in this study was obtained from the US Census via IPUMS USA and contains a huge range of variables pertaining to individuals and their housing situations. These variables include demographic factors such as age, income, and education level, alongside housing-related attributes like electricity cost, year of construction, and population density. The objective of this analysis is to predict whether a dwelling is occupied by owners or renters based on a combination of factors related to the occupants and the housing itself.

This study employs three different SVM kernels - linear, radial, and polynomial - to build and evaluate predictive models. There were seven predictors used: “NCOUPLES” (Number of couples in the household), “NFAMS”(Number of families in the household), “ROOMS”(Number of rooms), “BEDROOMS” (Number of bedrooms), “VEHICLES”

(Number of vehicles), “AGE” (Age of the person), besides “MARST”(Marital status). Of which, two of these were for individual-level variables namely “AGE” and “MARST” while the rest are housing-level variables. Through the analysis of these predictors, valuable insights into the determinants of housing ownership are uncovered, facilitating informed decision-making in the housing sector. By utilising these predictive models, the study aims to reveal insights on understanding housing dynamics, promoting equitable access to housing, and supporting sustainable urban development and economic growth.

Theoretical Background:

Classification:

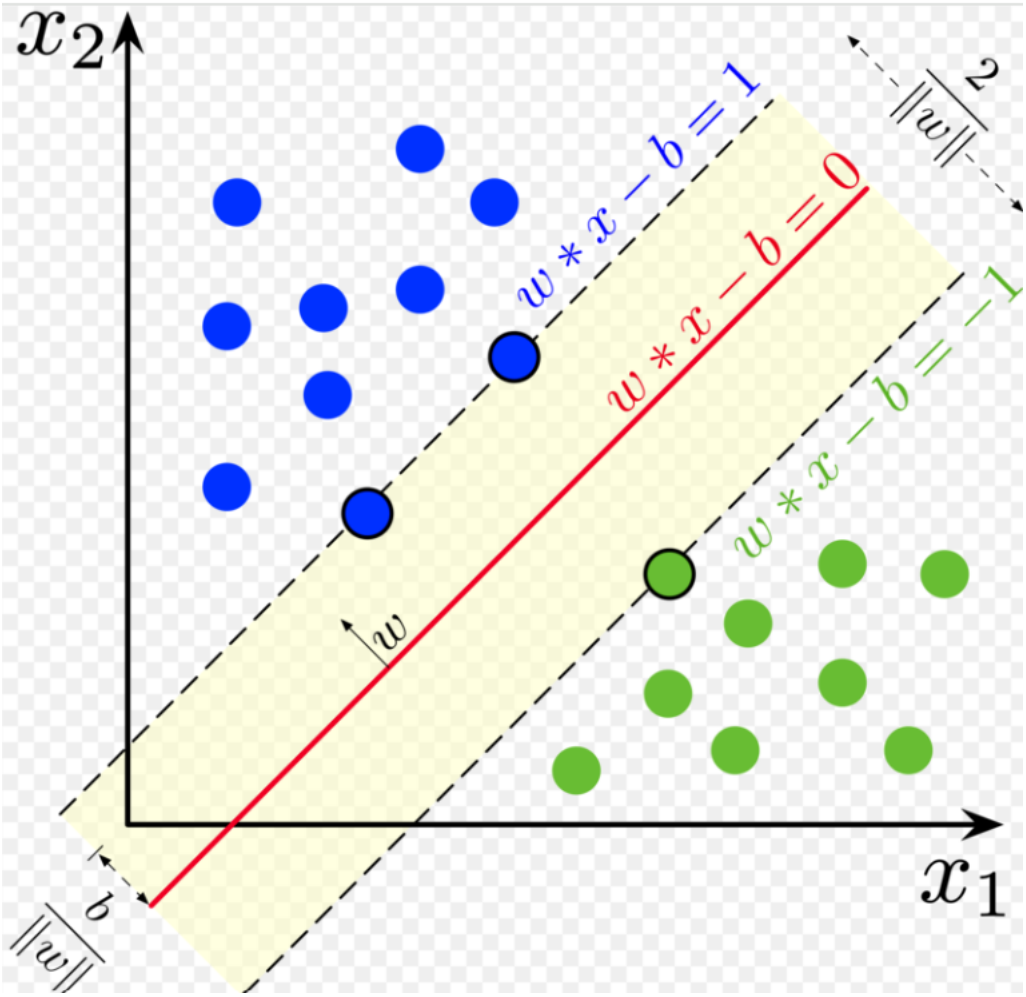
In artificial intelligence and machine learning, classification refers to the machine’s ability to assign the instances to their correct groups. This is done through learning a function that maps input features to a discrete target variable, with the target variable being predefined classes or categories.

Support Vector Machines:

SVM is a supervised machine learning algorithm that helps in classification or regression problems. It aims to find an optimal boundary between the possible outputs. SVM tries to find a line that maximizes the separation between a two-class data set of 2-dimensional space points. An objective of SVM is to find a hyperplane that maximizes the separation of the data points to their potential classes in an n-dimensional space. The data points with the minimum distance to the hyperplane (closest points) are called Support Vectors.

The hyperplane is a decision boundary that separates data points of one class from another.

[3]



Types of Kernels:

Various kernel types in the support vector machines (SVM) make it possible to map input data to higher-dimensional space resulting in better separability of the data. Below are three types of Kernels that were used in this analysis.

1. **Linear Kernel:**

Linear kernel SVMs are particularly effective when the data is linearly separable,.This is the simplest kernel, which computes the dot product between input feature vectors, suitable for linearly separable data. They are known for their high speed thus making them the most commonly used methods in more difficult classification tasks. It represents data in the same feature space as the original input data without any transformation.

Parameters(C):

C : C parameter is a regularization parameter that controls the trade-off between achieving training error and model complexity. It decides the smoothness of the decision boundary. When the value of C is large, the SVM optimization algorithm tries to minimize

the training error aggressively. Whereas, when the value of C is small, the SVM optimization algorithm prioritizes a simpler decision boundary that may not perfectly classify all training examples.

2. **Radial Basis Function (RBF) Kernel:**

This maps data into infinite-dimensional space by computing the similarity between data points based on their distance, suitable for non-linearly separable data. It is also called as Gaussian kernel. This SVM can capture complex relationships in the data by mapping it into a higher-dimensional space based on their Euclidean distance.

Parameters (C and gamma):

The RBF kernel has two main parameters.

C : This is same parameter as discussed in Linear SVM.

Gamma : Gamma determines the spread of the kernel and hence the influence of each training example on the decision boundary. A smaller gamma value leads to a smoother decision boundary with a larger margin, wheras a large gamma value results in a more complex decision boundary that closely fits the training data.This is also called as bandwidth parameter. This parameter controls the flexibility of the decision boundary.

3. **Polynomial Kernel:**

This maps data into higher-dimensional space using polynomial functions, allowing for non-linear decision boundaries. This SVM is also a flexible kernel function that can capture non-linear relationships in the data.

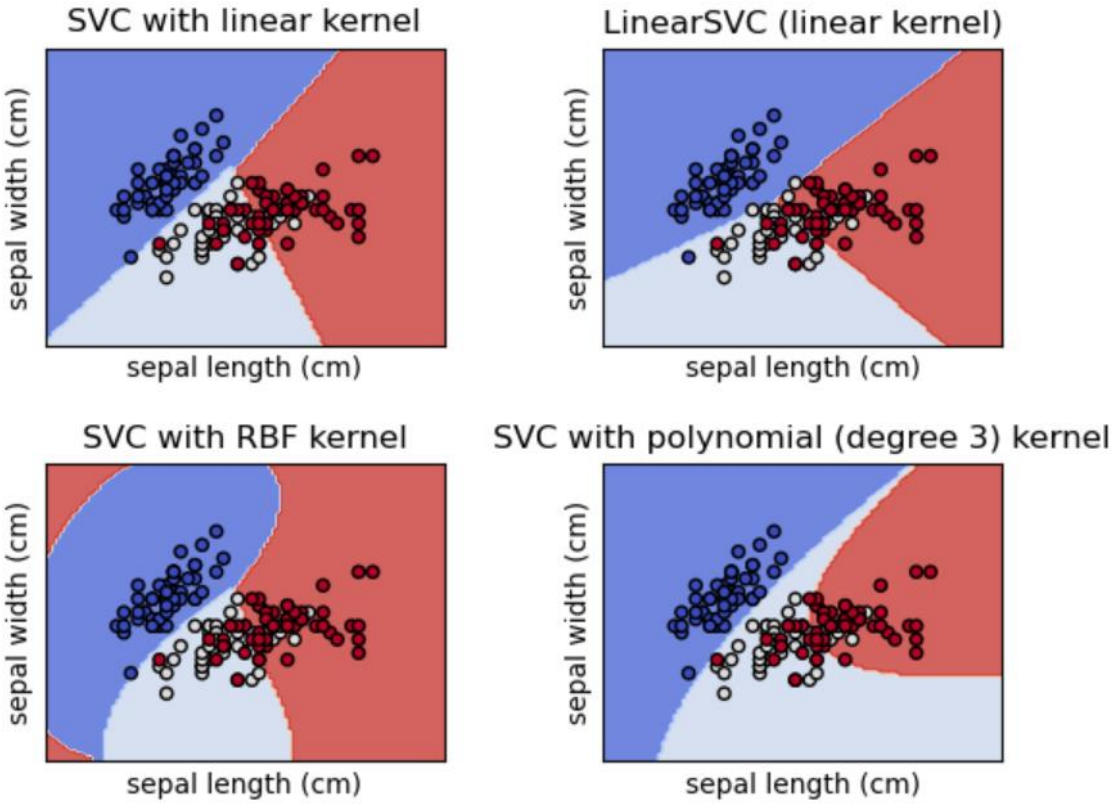
Parameters (C and Degree):

The Poly kernel has two main parameters.

C : This is same parameter as discussed in Linear SVM.

Degree: Degree parameter determines the degree of the polynomial function used to map the input data into higher-dimensional space. A higher degree value allows the model to capture more complex relationships in the data, potentially leading to overfitting if not properly tuned.

[2]



Parameter Tuning:

When parameters are tuned for SVM models, adjustments are made to maximize the model's effectiveness. It's similar to tuning a musical instrument to achieve the best possible sound. By tweaking settings such as the SVM hyper parameters like C , gamma and degree, the accuracy and applicability of the model can be enhanced. Parameter tuning ensures that the SVM model performs at its best for the specific dataset and issue at hand.Cross validation using K fold(5) was used in this analysis.

Methodology

Data clean up or pre-processing:

The dataset "Housing.csv" was read and loaded into the DataFrame "df". A round of data validation was conducted by checking its size, shape, and inspecting some rows using the head command. There were duplicates in the SERIAL column denoting each member of the same House. This lead to consider only the rows with the highest age for each serial number. The resulting Data frame was stored. For the assignment, individuals with an educational attainment equivalent to a Master's Degree was consoidered. So, the dataset was filtered to include only those rows where the Educational Attainment (EDUCD) column matched the code 114, as given in the code book.

Predictors and Target Variables:

There were seven predictors used: “NCOUPLES” (Number of couples in the household), “NFAMS”(Number of families in the household), “ROOMS”(Number of rooms), “BEDROOMS” (Number of bedrooms), “VEHICLES” (Number of vehicles), “AGE” (Age of the person), besides “MARST”(Marital status). Of which, two of these were for individual-level variables namely “AGE” and “MARST” while the rest are housing-level variables. Few variables like MARST were converted to Categorial using one hot encoding.

Train and Test Data set Splitting:

For all the models, the data was divided into a training set of 75% and test set of 25%.

Models:

In this analysi, three SVM models (linear, polynomial, and RBF) were developed to predict house ownership using a subset of predictor variables as below.

Initially, a SVM model with a linear kernel type was established, utilizing the default C value of 1. The number of support vectors was recorded, and both training and testing accuracy and error rates were documented. Additionally, confusion tables for both training and testing errors were plotted. To enhance performance, parameter tuning for the C value was conducted using 5-fold cross-validation. Subsequently, a new model incorporating the optimal C(0.01) value derived from the cross-validation process was constructed, and the test and train error rates and accuracy were compared with the initial results. A slight improvement in performance was observed following these adjustments.

Secondly, a SVM model with RBF kernel type was created, utilizing default C and gamma values. The number of support vectors was noted, and training and testing accuracy, along with error rates, were recorded. Additionally, confusion tables for both training and testing errors were plotted. To enhance performance, parameter tuning for the combination of C and gamma values was conducted using 5-fold cross-validation. Subsequently, a new model incorporating the optimal C(1) and gamma(0.1) values derived from the cross-validation process was constructed, and the test and train error rates and accuracy were compared with the initial results. A notable improvement in performance was observed after parameter tuning.

A third SVM model with a polynomial kernel type was created, employing default C and degree values. The number of support vectors was noted, and training and testing accuracy, as well as error rates, were documented. Additionally, confusion tables for both training and testing errors were generated. To optimize performance, parameter tuning for the combination of C and degree values was conducted using 5-fold cross-validation. Subsequently, a new model incorporating the optimal C (10) and degree(2) values obtained from the cross-validation process was constructed, and the test and train error rates and accuracy were compared with the initial results. A notable improvement in performance was observed after parameter tuning bring down both training and testing error rates.

Computational Results

The computation results are tabulated for each model as below.

1. **Linear Kernel Model :**

Prediction	SVM Model with Default Hyper parameters			SVM Model with Tuned Hyper parameters (C = 0.01)		
	Number of Support Vectors for each class	Accuracy		Number of Support Vectors for each class	Accuracy	
		Training	Testing		Training	Testing
Predicting the Dwelling Ownership using SVM Linear Kernel	[472 472]	85.69%	86.22%	[496 493]	85.73%	86.56%

2. **Radial Kernel Model :**

Prediction	SVM Model with Default Hyper parameters			SVM Model with Tuned Hyperparameters (C = 1, gamma = 0.1)		
	Number of Support Vectors for each class	Accuracy		Number of Support Vectors for each class	Accuracy	
		Training	Testing		Training	Testing
Predicting the Dwelling Ownership using SVM Radial Kernel	[552 549]	84.27%	83.20%	[711 457]	88.34%	86.22%

3. **Polynomial Kernel Model:**

Prediction	SVM Model with Default Hyper parameters			SVM Model with Tuned Hyperparameters (C = 10, degree= 2)		
	Number of Support Vectors for each class	Accuracy		Number of Support Vectors for each class	Accuracy	
		Training	Testing		Training	Testing
Predicting the Dwelling Ownership using SVM Polynomial Kernel	[588 586]	82.03%	80.29%	[531 530]	85.91%	86.78%

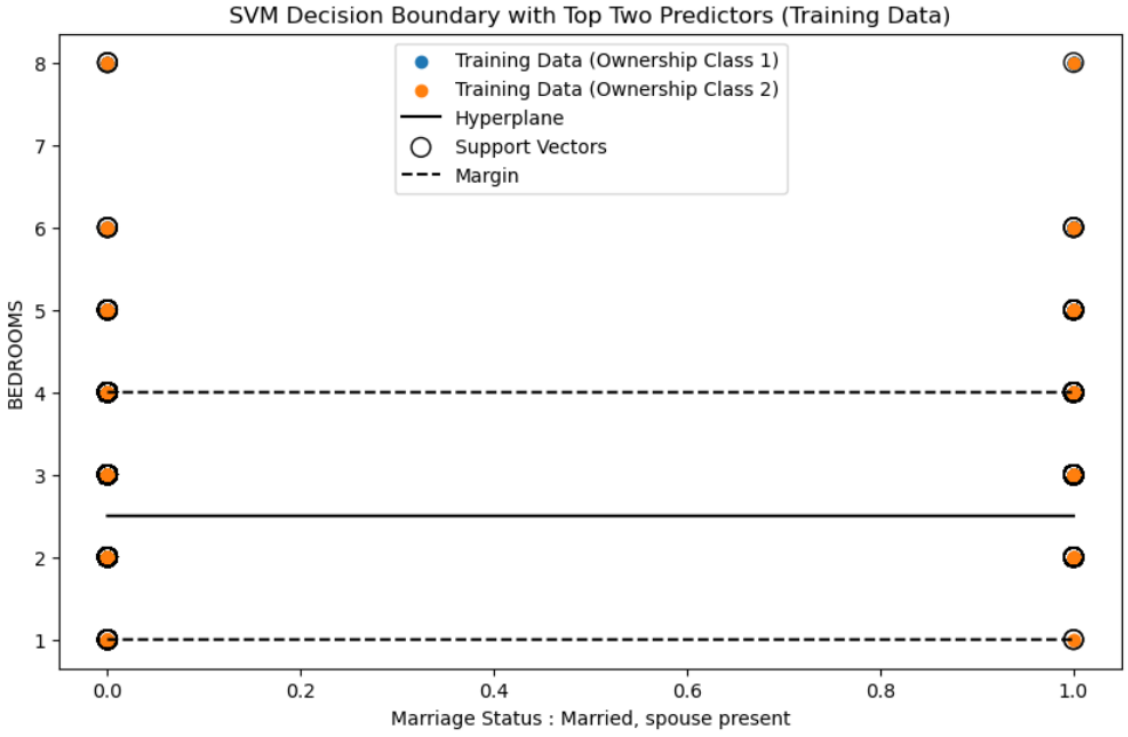
Discussion

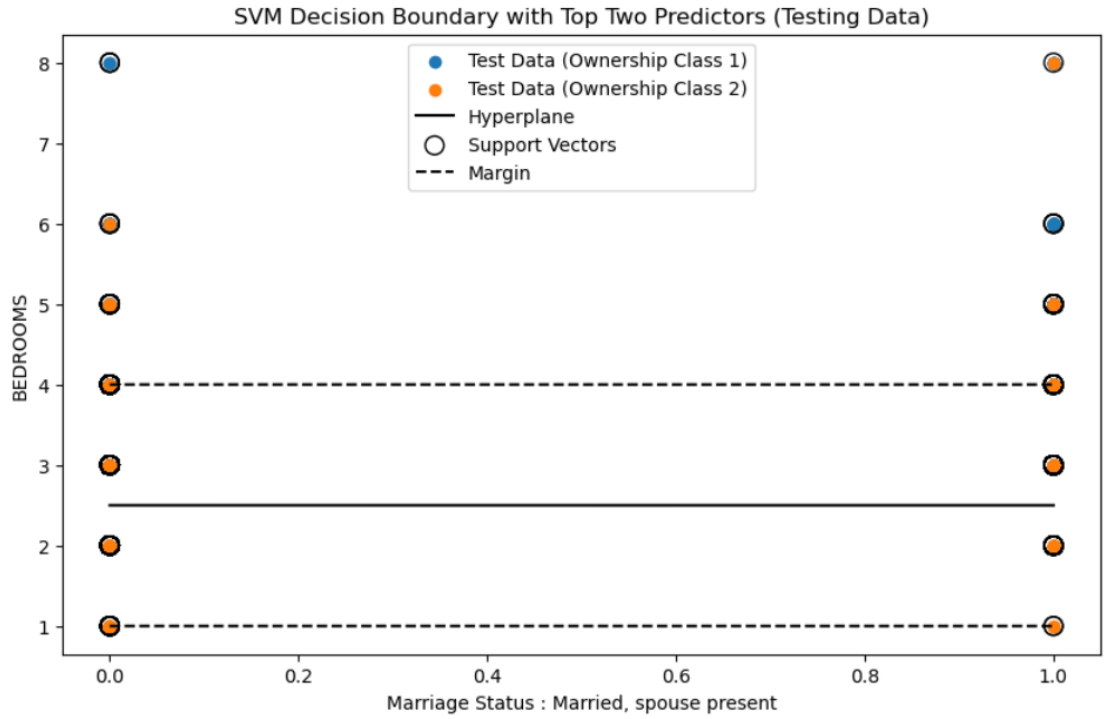
In this report, the world of housing is being explored to understand the factors that influence whether homes are owned or rented. SVM models will be utilized to assist in making sense of the data and determining the factors driving these decisions.

For each kernel type (linear, radial, and polynomial), two SVM models were built – one with default hyperparameters and the other with tuned hyperparameters obtained through GridSearchCV using 5 fold cross validation. All types of kernels showed that models with adjusted hyperparameters performed better on training and testing datasets than their counterparts. This demonstrates how important hyperparameter adjustment is for maximizing SVM models' predictive capabilities.

When comparing training errors across all kernel types and tuned parameters, it was observed that the radial kernel exhibited the highest training accuracy at approximately 88.34%. However, its corresponding testing accuracy (86.22%) was lower compared to the polynomial kernel (86.78%). This could be because of overfitting issue in the Radial Kernel model. This model has the highest number of support vectors when compared to the other kernel models. A larger number of support vectors typically indicates a more complex decision boundary, which may suggest that the model has overfit the training data. On the other hand, a smaller number of support vectors may indicate a simpler decision boundary, which could mean that the model has generalized well to unseen data. So the radial model was performing better in the training data set but was not in testing data set.

For the Linear SVM, the top two predictors were identified based on their coefficients from the model and plotted them individually for training and testing Data set as below.





The top two predictor variables were Number of bedrooms and the Martial Status (married, spouse present) for this tuned model. From the plots, it appears that the number of bedrooms is a more important predictor of home ownership than marital status (married, spouse present) since the decision boundary angles more steeply in the direction of the bedroom axis. Having more bedrooms tends to favour ownership (upper blue data points), although marriage status have less of an impact. According to the above model, those who live in houses with many bedrooms and are married couples who living with spouses are more likely to be homeowners.

Conclusion

In order to estimate dwelling ownership in Washington State, this analysis uses Support Vector Machines (SVMs) with a large dataset obtained from the US Census through IPUMS USA. With the use of seven predictors comprising both person and housing-level characteristics, and SVMs with three distinct kernels (linear, radial, and polynomial), this analysis obtains encouraging results in distinguishing individuals with Master's Degree owner-occupied and renter-occupied houses.

After adjusting the hyperparameters, the results indicate noteworthy testing accuracies from 86.22% to 86.78%, demonstrating the effectiveness of SVMs in this classification problem. This analysis adds to a better knowledge of the factors that influence housing ownership based on variables including age, marital status, number of vehicles, rooms, bedrooms, and families.

Overall, it is evident that polynomial kernel type SVM works better for this data set based on the accuracy on testing data. This analysis will not only help forecast which people will buy a home, but they also lay the groundwork for promoting equitable housing possibilities and moving Washington State and beyond toward more prosperous and sustainable communities.

Bibliography/References:

References

1. Retrieved from https://www.ipums.org/projects/ipums-usa/d010.V13.0

2. Retrieved from https://www.baeldung.com/cs/svm-multiclass-classification

3. Retrieved from
   https://en.wikipedia.org/wiki/Support_vector_machine#/media/File:SVM_margin.png

4. Retrieved from https://scikit-learn.org/stable/modules/svm.html

Appendix Code: