Classification and Evaluation with Support Vector Machine (SVM)

Yang Yang

STAT 745 Final Report

Abstract:

Classification with Support Vector Machine (SVM) is a supervised machine learning method. It is used to separate collections of labeled objects, which share similar features in each class, and assign new objects to defined groups. In this project, a classification method was built based on a wheat kernel dataset, consisting of 210 observation and 7 variables, to predict kernel variety from three classes. After comparing linear, polynomial and radial models, a linear kernel SVM training with two variable, length of kernel and length of kernel groove, yields reliable prediction with high accuracy.

1. Introduction

Classification is a multivariable techniques concerned with separating distinct sets of objects with labels and with allocating new objects to previously defined classes. The emphasis of classification is on deriving a rule with which to optimally sort new objects to the labeled classes.

Support Vector Machine (SVM) is a supervised machine learning algorithm which is mostly used for classification. It was first introduced in the early 1990s. However, due to the effective and versatile performance in high dimensional spaces, SVMs have led to a recent explosion of real-world applications and deepening theoretical analysis[1]. Currently, the broad range of application of SVMs lies in handwritten digit recognition[2], text categorisation, image classification, biosequences analysis[3], chemometrics[4] and others. Moreover, SVMs also has become one of the standard tools for machine learning and data-mining field.

This project is to develop a classification method to predict kernel variety of wheat based on a multivariate dataset, consisting of 210 instances and 7 attributes. The dataset was obtained from kernels of 3 different varieties of wheat (Kama, Rosa and Canadian), which are the label for classification. The images of internal kernel structure were detected using a soft X-ray technique. Seven geometric parameters of wheat kernels were measured, which are area A, perimeter P, compactness C = 4*pi*A/P^2, length of kernel, width of kernel, asymmetry coefficient, length of kernel groove. The dataset is from http://archive.ics.uci.edu/ml/datasets/seeds.

2. Method

2.1 Theory

In two-class classification, SVM constructs a maximal margin hyperplane (MMH) to separate two classes. MMH maximizes distance from hyperplane to support vectors which are the closest points of the training set belonging to different classes.

For linearly non-separable data set containing n spectra $X=\{(x_1,y_1),\ldots,(x_n,y_n)\}$ with $x_i \in R^n$, $y_i \in \{-1,+1\}$, $i = 1, \ldots, n$. The decision function of linear hyperplane is

$$f(x) = \langle w^T \cdot \mathrm{x} \rangle + b \tag{1}$$

The function involves the hyperplane parameters b and w, which are offset and weight vector. Points with corresponding *f(x)* > 1 belong to class '+1', otherwise to the class '−1'.

The weight vector *w* can be solved by minimizing the following problem:

$$\min\{0.5\| w \|^2 + \mathrm{C}\sum_{i=1}^{N} \xi_{\mathrm{i}}\} \tag{2}$$

subject to the constraints :

$$y_{\mathrm{i}} \left(\langle w^T \cdot \mathrm{x} \rangle + b\right) \geq 1 - \xi_{\mathrm{i}}, \text{for class } y_{\mathrm{i}} = \pm 1 \tag{3}$$

$\xi_i \geq 0$, for all i

Due to the overlapping of class distributions, the first term in Eq. (2) is to maximize the margin, while the second term minimizes training error by penalizing non-separable samples.

In cases where no linear hyper-plane exist between two classes, kernel trick is used to separate classes by transforming low dimensional input space to a higher dimensional space.

The decision function of nonlinear hyperplane is:

$$f(x) = \sum_{i=1,SV} \alpha_i \, y_i \, K(x_i, x_j) + b \tag{4}$$

Where α_i are the sought contribution factors for each training sample, y_i are their class labels, and $K(x_i, x_j)$ is *kernel* function.

Prototypical kernel functions are:

Gaussian radial basis function kernel (RBF),

$$K(\mathrm{x}, x') = \exp(-\| \mathrm{x} - \mathrm{x}' \|^2 / 2\sigma^2) \tag{5}$$

where σ is the variance of the Gaussian.

Polynomial kernel,

$$K(\mathrm{x}, x') = (x \cdot x' + 1)^d \tag{6}$$

where d is the degree of the polynomial.

Linear kernel,

$d = 0$ in Eq. (6).

2.2 Data analysis and visualization

The wheat kernel data was split into training set and test set with split ratio 0.8 so that there are 168 observations in training set and 42 in test set.

Three support vector machine (SVM) models (linear, polynomial and radial) were built based on all seven variables from training set, and then the test set was used to evaluate the accuracies of predictions to select the best model for this dataset.

SVM classifier was then built using the best model with every two variables. Combinations of variables with high accuracy were selected for classifier visualization.

3. Result and discussion

The SVM classifiers were constructed with linear, polynomial or radial model based on training sets built in repeated times. From results by several times of training, predictions with radial and linear models have similar accuracy and are both better than those with polynomial model. A represented result was shown in Table. 1. The model type is given in first cell. The diagonal values represent correct classification where the predicted classes are the same as labels. The off-diagonal values mean that the predicted results did not match their labels. The cell in second row and second column shows calculated prediction accuracy.

linear		predict		
test	100%	Canadian	Kama	Rosa
	Canadian	14		
	Kama		14	
	Rosa			14

Radial		predict		
test	97.6%	Canadian	Kama	Rosa
	Canadian	14		
	Kama		13	1
	Rosa			14

Polynomial		predict		
test	90.5%	Canadian	Kama	Rosa
	Canadian	12	2	
	Kama		14	
	Rosa		2	12

Table. 1: Prediction results with 7 variables using linear, polynomial or radial models.

Since linear and radial models performed similarly in the dataset, linear model was selected for further comparisons among predictions with two variable combinations. The accuracy of prediction from each combination of two variable is given in Fig. 1 and Fig. 2 is corresponding heat map. The diagonal values represent predictions from one variable. From Fig.1 and Fig.2, prediction with X1 or X2 has higher accuracy compared with other one variable predictions. While results predicted with X3, X6 or X7 have the least accuracy. However, combination of X7 with X2 or X4 yields highest accuracy, which are 100%.

Classification Accuracy of SVM from Two Variables

	X1	X2	X3	X4	X5	X6	X7
X1	0.952	0.976	0.952	0.929	0.929	0.929	0.952
X2	0.976	0.976	0.952	0.952	0.952	0.929	1
X3	0.952	0.952	0.548	0.881	0.929	0.738	0.905
X4	0.929	0.952	0.881	0.857	0.905	0.881	1
X5	0.929	0.952	0.929	0.905	0.881	0.929	0.929
X6	0.929	0.929	0.738	0.881	0.929	0.5	0.81
X7	0.952	1	0.905	1	0.929	0.81	0.643

Figure. 1: Classification accuracy of linear SVM predicted from two variables.

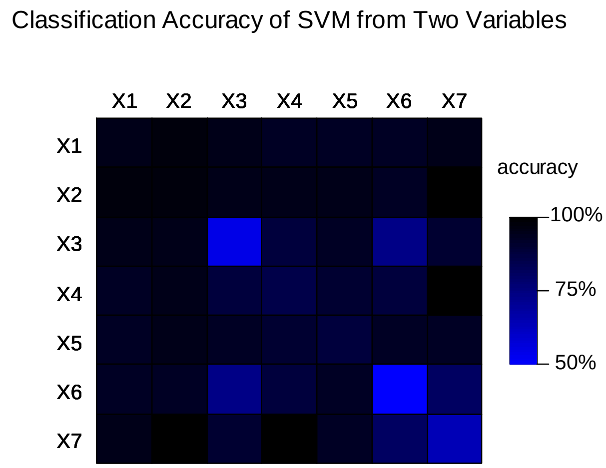

Figure. 2: Heat map of classification accuracy of linear SVM predicted from two variables.

Figure. 3: Pairwise scatter plots and correlations of seven variables.

Pairwise scatter plots were constructed to compare each two variables in Fig.3. Correlation of every two variables are shown in the upper part of Fig. 3. Diagonal plots are histogram displaying the distributions of each variable. According to correlation results, X1 has linear relationship with X2, X4 and X5 whose correlations are 0.995, 0.952 and 0.974 respectively, all close to 1. From histograms, distributions of X1 and X2 almost separate three classes, which are similar to results in Fig. 1 and 2. Moreover, the conclusion from Fig1 and 2, which prediction using X7 combined with X2 or X4 leads to highest accuracy, is verified in pairwise scatter plots of X7-X2 and X7-X4 where fewer points with different colors are mixed.

To visualize the classifier in two dimension plot, variable X4 and X7 were chosen for prediction. Classifiers of linear, polynomial and radial model were visualized in Fig.4. Highlighted points indicate misclassification. The linear model has the best separation and lowest error compared with polynomial and radial. This conclusion can also be drawn by the linear boundaries of three classes in X4-X7 pairwise scatter plot in Fig.3.

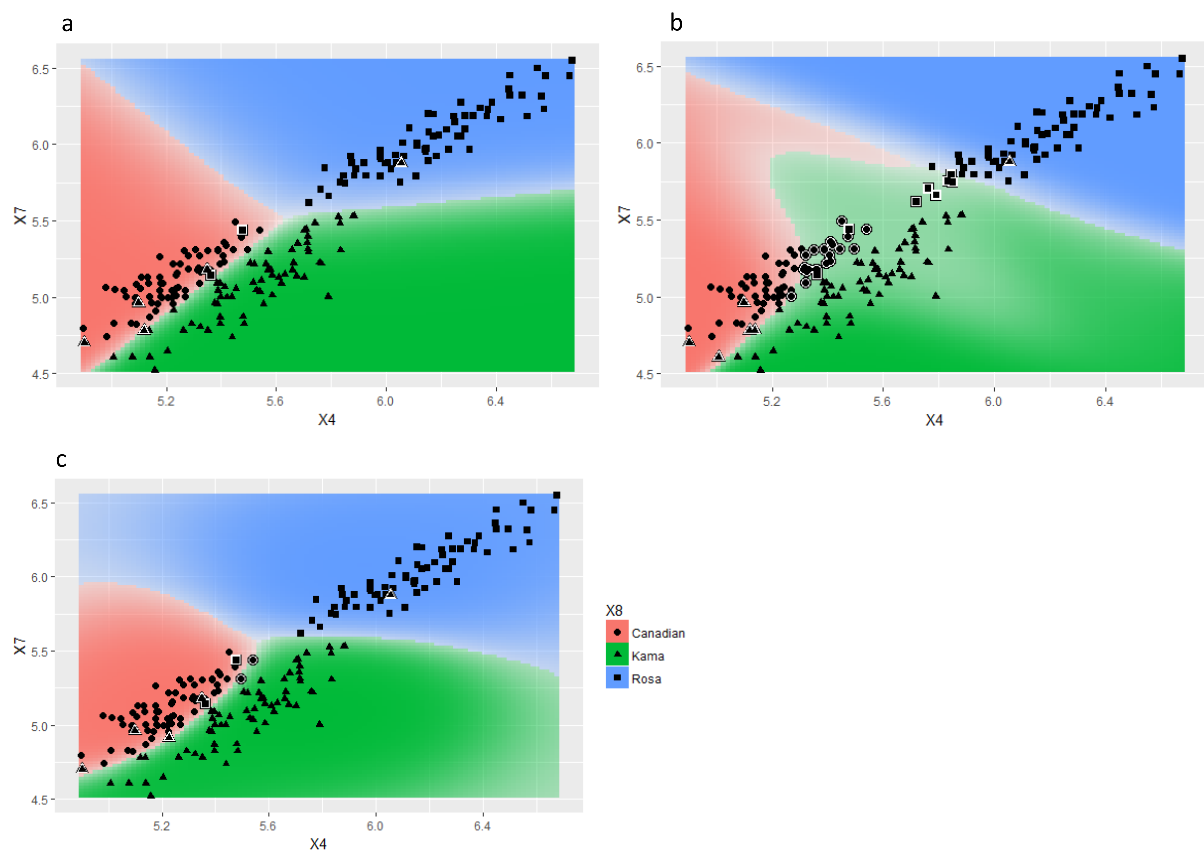

Figure. 4 Visualization of linear (a), polynomial (b) and radial (c) classifiers with variable X4 and X7.

4. Conclusion

Classification with SVM is a versatile tool by providing different algorithms to achieve the best separation. In this wheat kernel dataset, the linear and radial SVM models performed than polynomial models. By calculating the correlation, area of kernel (X1) was proved to have linear relationship with perimeter (X2), length (X4) and width (X5). Moreover, linear SVM model using length of kernel (X4) and length of kernel groove (X7) yields reliable prediction with high accuracy.

References:

1. NPSER BE. A training algorithm for optimal margin classifiers. *Proc.5th annual ACM workshop on Computation Learning Theory, 1992*. 1992:144-152. https://ci.nii.ac.jp/naid/10030527281/en/.
2. Bottou L, Cortes C, Denker J, et al. *Comparison of classifier methods: A case study in handwritten digit recognition.* ; 1994:82 vol.2. 10.1109/ICPR.1994.576879.
3. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*. 2000;16(10):906-914. http://dx.doi.org/10.1093/bioinformatics/16.10.906.
4. Masoum S, Malabat C, Jalali-Heravi M, Guillou C, Rezzi S, Rutledge DN. Application of support vector machines to 1H NMR data of fish oils: Methodology for the confirmation of wild and farmed salmon and their origins. *Analytical and Bioanalytical Chemistry*. 2007;387(4):1499-1510. https://doi.org/10.1007/s00216-006-1025-x. doi: 10.1007/s00216-006-1025-x.