Dustin McAfee

Project 5 Report

Fall 2018

1 Objective

There are three datasets for use with (soft-margin) Support Vector Machines (SVMs) in classification. The first is a binary classification problem: predicting 'good' or 'bad' interactions of radar signals with electrons in the ionosphere. This dataset has 351 instances and 34 attributes. The second classification is multi-class with 11 classes: predicting vowels independent of the speaker. This dataset has 528 instances and 10 attributes. The last classification problem is also multi-class with 7 classes: predicting terrain type from multi-spectral values of pixels in 3 by 3 neighborhoods in satellite images. This dataset has 4435 training observations and 2000 testing instances with 36 features. The objective is to perform grid searches on SVMs with different kernels (linear, polynomial, radial basis), each with different hyper-parameters in order to find an optimal SVM (with optimal hyper-parameters) for each dataset.

2 Methods

A Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification or regression. In this scope, SVM is used for classification, and in this respect, (soft-max) SVM minimizes the error of classification by creating boundaries (discriminants) between classes, with specified slack to the boundary. The slack variables allow for non-separable categories of data: If the data is non-separable, then the slack variables need to be increased, which allows for categorization of observations that are just passed the discriminant hyperplane. The kernel of this algorithm is, essentially, the shape of the hyperplane that defines the discriminant. With a linear kernel, the hyperplane is linear. Polynomial hyperplanes (polynomial kernels) take a hyper-parameter that is the degree of the polynomial. The Radial Basis Kernel is a distance measure of the observation to a center point. In this paper, the radial basis kernel used is Gaussian, and so the distances are measured in a Gaussian probability curve, which acts as a similarity measure. The objective of the SVM is to maximize the distance of each point from each discriminant hyper-plane, while minimizing the error. Increasing the complexity factor (hyper-parameter for each kernel) decreases the margin for the discriminant hyper-plane, making it more specific to the current training dataset, potentially over-fitting the data. On the other hand, a small complexity factor results in large margins for the discriminant hyper-plane, which allows for more noise (higher slack variables). The mathematics for this is shown below.

A linear discriminant hyperplane can be written as the set of points $\mathbf{x}$ satisfying

$$\mathbf{w} \cdot \mathbf{x}^t + w_0 = 0$$

The boundaries of this hyperplane (with a distance of 1 from the discriminant) can be written as the set of points $\mathbf{x}$ satisfying

$$\mathbf{w} \cdot \mathbf{x}^t + w_0 = 1 \text{ and } \mathbf{w} \cdot \mathbf{x}^t + w_0 = -1$$

To maximize these margins, $\frac{2}{||\mathbf{w}||^2}$ needs to be maximized, which means $||\mathbf{w}||$ needs to be minimized. To minimize error, we need to minimize the number of data points that fall past the margin of the discriminant. If $r^t(\mathbf{w} \cdot \mathbf{x}^t + w_0) \geq 1$ is a 'hard' constraint (no observations are allowed past the margins), then $r^t(\mathbf{w} \cdot \mathbf{x}^t + w_0) \geq 1 - \xi^t$ is a 'soft' margin, in which ξ^t represents the slack variable corresponding to observation $\mathbf{x}^t$. To this end, the hinge loss function is introduced:

$$\xi^t = \max(0, 1 - r^t(\mathbf{w} \cdot \mathbf{x}^t + w_0)),$$

which is zero if the observation obeys the hard constraint. For observations that do not obey the hard margin, the value is proportional to the distance from the margin. The goal, then, is to

$$\text{minimize: } \frac{1}{2}||\mathbf{w}||^2 + C\sum_t \xi^t$$
$$\text{subject to: } r^t(\mathbf{w} \cdot \mathbf{x}^t + w_0) \geq 1 - \xi^t,$$

where C is the complexity factor (otherwise known as penalty). This is also known as the primal problem, and by solving for the Lagrangian dual of this problem obtains the simpler problem:

$$\text{maximize: } \sum_t \alpha^t - \frac{1}{2}\sum_t \sum_s \alpha^t \alpha^s r^t r^s (\mathbf{x}^t \cdot \mathbf{x}^s)$$
$$\text{subject to: } \sum_t \alpha^t r^t = 0,$$
$$\text{and } 0 \leq \alpha^t \leq \frac{1}{n}$$

The dot product between the two observations, $\mathbf{x}^t$ and $\mathbf{x}^s$ is called a linear kernel, because it describes a linear discriminant hyper-plane. A polynomial kernel expands the feature-set to include higher order terms before making the dot product, and the radial basis kernel uses the function, $e^{-\gamma||\mathbf{x}^t - \mathbf{x}^s||^2}$, where $\gamma > 0$ is a hyper-parameter for this specific kernel. A small gamma means a Gaussian discriminant hyper-plane with a large variance, which means the influence of $\mathbf{x}^s$ is large (vice versa for large gamma).

3 Data Preprocessing

Each dataset is z-normalized, and the attributes scaled in the range of [0,1]. There are no missing values in the three datasets, but the categorical attribute for the ionosphere dataset (binary) is converted to 0/1 and separated into a testing dataset (20% or 70 observations), and a training dataset (281 observations). K-fold cross-validation is performed with 3 folds on the training dataset of the ionospheric data. For the vowel dataset, 20% is separated for the testing dataset (198 observations), and for the satellite image dataset, the testing dataset comes pre-separated with about 31% of the data (2000 observations). For the satellite and vowel dataset, K-fold cross-validation is performed with 6 folds.

4 Training Results

K-fold cross-validation is used for validation of hyper-parameters. For each classification problem, a coarse grid search is performed for the hyper-parameters of the polynomial kernel (degree and C complexity factor) and for the hyper-parameters of the radial basis kernel (γ and C). A fine grid search is then performed for the kernel than performs better, and the optimal hyper-parameters are chosen from the fine grid searches. The fine grid search is ran in the 'neighborhoods' of the optimal hyper-parameters chosen from the coarse grid search. The kernel that yields the best results is chosen to run with its optimal hyper-parameters on the testing dataset. For coarse grid searches, the complexity factor, C, is tested from the range of 0.0001 to 100000, by factors of 10, and gamma is tested (in the radial basis kernel) in the same range of C. For the polynomial kernel, the degree hyper-parameter is tested from 1 to 5.

4.1 Ionospheric Dataset

For this binary classification problem, the discriminant hyper-plane is determined from the only two possible classes (one vs one only). The K-fold cross-validation mean accuracy and precision measurements from the coarse grid search using a polynomial kernel (See file "output/ionosphere/ionosphere_train_poly.txt") are shown below, in Figure 1. The 95% confidence interval is also given for these performance measurements.

Hyper-Parameter C	Hyper-Parameter degree	Mean Accuracy	Mean Precision
$\leq$0.1	-	0.637 $\pm$ 0.002	0.637 $\pm$ 0.002
1	1	0.797 $\pm$ 0.023	0.759 $\pm$ 0.021
1	2	0.829 $\pm$ 0.053	0.791 $\pm$ 0.054
1	3	0.699 $\pm$ 0.127	0.683 $\pm$ 0.096
1	4	0.637 $\pm$ 0.002	0.637 $\pm$ 0.002
1	5	0.637 $\pm$ 0.002	0.637 $\pm$ 0.002
10	1	0.922 $\pm$ 0.018	0.905 $\pm$ 0.026
10	2	0.929 $\pm$ 0.010	0.919 $\pm$ 0.018
10	3	0.929 $\pm$ 0.018	0.921 $\pm$ 0.024
10	4	0.927 $\pm$ 0.014	0.921 $\pm$ 0.024
10	5	0.907 $\pm$ 0.022	0.917 $\pm$ 0.032
100	1	0.964 $\pm$ 0.005	0.954 $\pm$ 0.019
100	2	0.979 $\pm$ 0.018	0.976 $\pm$ 0.034
100	3	0.974 $\pm$ 0.015	0.978 $\pm$ 0.028
100	4	0.968 $\pm$ 0.023	0.978 $\pm$ 0.028
100	5	0.961 $\pm$ 0.013	0.975 $\pm$ 0.014
1000	1	0.989 $\pm$ 0.005	0.986 $\pm$ 0.015
1000	2	0.996 $\pm$ 0.005	0.994 $\pm$ 0.008
1000	3	0.996 $\pm$ 0.005	0.994 $\pm$ 0.008
1000	4	0.991 $\pm$ 0.013	0.992 $\pm$ 0.013
1000	5	0.984 $\pm$ 0.015	0.992 $\pm$ 0.014
10000	1	0.998 $\pm$ 0.005	0.997 $\pm$ 0.008
$\geq$10000	$\geq$2	1 $\pm$ 0	1 $\pm$ 0

Figure 1: Polynomial SVM Performance with Ionospheric Data

These same performance measurements are shown, again, for a radial basis kernel SVM, below, in Figure 2.

Hyper-Parameter C	Hyper-Parameter γ	Mean Accuracy	Mean Precision
0.1	$\leq$0.01	0.637 $\pm$ 0.002	0.637 $\pm$ 0.002
0.1	0.1	0.653 $\pm$ 0.022	0.647 $\pm$ 0.014
0.1	1	0.847 $\pm$ 0.245	0.824 $\pm$ 0.230
0.1	$\geq$10	0.637 $\pm$ 0.002	0.637 $\pm$ 0.002
1	$\leq$0.001	0.637 $\pm$ 0.002	0.637 $\pm$ 0.002
1	0.01	0.714 $\pm$ 0.024	0.690 $\pm$ 0.017
1	0.1	0.943 $\pm$ 0.018	0.918 $\pm$ 0.024
1	1	0.991 $\pm$ 0.010	0.992 $\pm$ 0.013
$\geq$1	$\geq$10	1 $\pm$ 0	1 $\pm$ 0

Figure 2: Radial Basis Function SVM Performance with Ionospheric Data

The hyper-parameter selected should give the best generalization of the data. Of course, high complexity factor C means less generalization properties of the classifying discriminant hyper-plane.

In other words, with large C, the SVM tries to fit the data as best as possible with little slack. To this end, the optimal hyper-parameter intervals chosen for the fine grid search are from the radial basis SVM, which has better generalization than the polynomial SVM. A finer grid search is run for the chosen $C = 1$ hyper-parameter, and $\gamma \in \{0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$, and the average accuracy and precision for the K-fold cross-validation is shown, below, in Figure 3.

Hyper-Parameter C	Hyper-Parameter γ	Mean Accuracy	Mean Precision
1	0.05	0.915 ± 0.027	0.882 ± 0.033
1	0.1	0.943 ± 0.018	0.918 ± 0.024
1	0.15	0.952 ± 0.009	0.934 ± 0.018
1	0.2	0.966 ± 0.010	0.955 ± 0.019
1	0.3	0.975 ± 0.005	0.970 ± 0.008
1	0.4	0.980 ± 0.005	0.978 ± 0.008
1	0.5	0.984 ± 0.009	0.983 ± 0.013
1	0.6	0.984 ± 0.009	0.983 ± 0.013
1	0.7	0.986 ± 0.005	0.986 ± 0.008
1	0.8	0.989 ± 0.009	0.989 ± 0.008
1	0.9	0.991 ± 0.010	0.992 ± 0.013
1	1	0.991 ± 0.010	0.992 ± 0.013

Figure 3: Radial Basis Function SVM Performance with Ionospheric Data

A decent elbow point for both the mean accuracy and mean precision of the K-fold cross-validation seems to be at $\gamma = 0.7$ from $C = 1$. These are run on the testing dataset, with the confusion matrix shown below, in Figure 4.

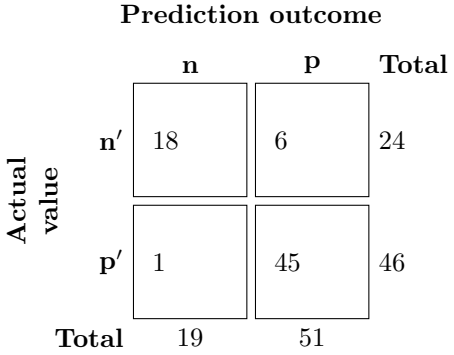

Figure 4: Confusion Matrix for Testing Dataset: Radial Basis SVM; $\gamma = 0.7$; $C = 1$

This yields 90% accuracy and 88.2% precision. The precision is lower than expected (higher false positive rate), seeing as though the training set had an average of 0.986% precision and accuracy.

4.2 Vowel Dataset

For this multi-class classification problem, the discriminant hyper-plane is determined from the multiple classes using OVA (one versus all). The K-fold cross-validation mean AUC ROC (Area Under Curve Receiver Operating Characteristics) and precision measurements from the coarse grid search using a polynomial kernel (See file "output/vowels/vowels_train_poly.txt") are shown below, in Figure 5. The 95% confidence interval is also given for these performance measurements. The ROC is a probability curve, and AUC represents a measurement of separability.

Hyper-Parameter C	Hyper-Parameter degree	Mean ROC AUC	Mean Precision
≤ 100	≤ 5	0.651 ± 0.009	0.301 ± 0.026
1000	1	0.623 ± 0.010	0.281 ± 0.021
1000	2	0.821 ± 0.007	0.640 ± 0.020
1000	3	0.797 ± 0.007	0.602 ± 0.015
1000	4	0.730 ± 0.016	0.490 ± 0.029
1000	5	0.653 ± 0.012	0.359 ± 0.026
10000	1	0.626 ± 0.009	0.286 ± 0.021
10000	2	0.959 ± 0.006	0.897 ± 0.015
10000	3	0.961 ± 0.004	0.908 ± 0.014
10000	4	0.941 ± 0.009	0.872 ± 0.016
10000	5	0.879 ± 0.007	0.755 ± 0.011
100000	1	0.626 ± 0.010	0.285 ± 0.021
≥ 100000	≥ 2	≥ 0.991	≥ 0.969

Figure 5: Polynomial SVM Performance with Vowel Data

This dataset seems to have the worst performance when attempted to separate the classes linearly (degree = 1). Furthermore, the SVM (with polynomial kernel) only performs well for extremely large C values, which is bad generalization. These same performance measurements are shown, again, for a radial basis kernel SVM, below, in Figure 6.

Hyper-Parameter C	Hyper-Parameter γ	Mean ROC AUC	Mean Precision
≤ 1	≤ 1	≤ 0.636	≤ 0.334
1	10	0.992 ± 0.004	0.984 ± 0.007
1	≥ 100	1.0 ± 0.0	1.0 ± 0.0
10	≤ 0.1	≤ 0.597	≤ 0.253
10	1	0.921 ± 0.009	0.831 ± 0.012
10	≥ 10	1.0 ± 1.0	1.0 ± 1.0
100	≤ 0.1	≤ 0.718	≤ 0.458
100	1	0.997 ± 0.001	0.989 ± 0.005
100	≥ 10	1.0 ± 0.0	1.0 ± 0.0
1000	≤ 0.01	≤ 0.630	≤ 0.298
1000	0.1	0.920 ± 0.010	0.821 ± 0.018
1000	≥ 1	1.0 ± 0.0	1.0 ± 0.0
10000	≤ 0.01	≤ 0.722	≤ 0.461
10000	0.1	0.989 ± 0.007	0.965 ± 0.012
10000	≥ 1	1.0 ± 0.0	1.0 ± 0.0
10000	≤ 0.001	≤ 0.637	≤ 0.310
10000	0.01	0.916 ± 0.011	0.811 ± 0.018
10000	≥ 0.1	1.0 ± 0.0	1.0 ± 0.0

Figure 6: Radial Basis Function SVM Performance with Vowel Data

Likewise from Section 4.1, a low C is chosen for the fine grid search. Specifically, the fine grid search is for $C \in \{1, 2, 3, 4, 5, 6, 7, 8\}$, and $\gamma \in \{1, 2, 3, 4, 5, 6, 7, 8\}$. Most of these results are omitted here, since there are exactly 64 instances to look through (See file "output/vowels/vowels_train_rbf_fine.txt"). Because of this, only the relevant results are shown, below in Figure 7, of which, the optimal hyper-parameters are chosen.

Hyper-Parameter C	Hyper-Parameter γ	Mean ROC AUC	Mean Precision
1	4	0.915 ± 0.008	0.840 ± 0.017
1	5	0.945 ± 0.010	0.898 ± 0.019
1	6	0.964 ± 0.007	0.934 ± 0.013
1	7	0.975 ± 0.006	0.954 ± 0.012
1	8	0.984 ± 0.004	0.970 ± 0.009
2	3	0.947 ± 0.009	0.894 ± 0.019
2	4	0.973 ± 0.003	0.948 ± 0.009
2	5	0.989 ± 0.008	0.977 ± 0.016
2	6	0.995 ± 0.005	0.989 ± 0.008
2	7	0.997 ± 0.003	0.993 ± 0.005
2	8	0.998 ± 0.002	0.995 ± 0.004
3	2	0.924 ± 0.008	0.848 ± 0.013
3	3	0.971 ± 0.006	0.940 ± 0.012
3	4	0.990 ± 0.006	0.978 ± 0.011
3	≥ 5	≥ 0.996	≥ 0.991
4	2	0.947 ± 0.006	0.891 ± 0.012
4	3	0.984 ± 0.006	0.964 ± 0.012
4	≥ 4	≥ 0.995	≥ 0.988
5	2	0.960 ± 0.006	0.916 ± 0.010
5	3	0.991 ± 0.005	0.979 ± 0.010
5	≥ 4	≥ 0.997	≥ 0.992
6	2	0.970 ± 0.007	0.936 ± 0.011
6	≥ 3	≥ 0.993	≥ 0.984
7	2	0.977 ± 0.008	0.948 ± 0.015
7	≥ 3	≥ 0.995	≥ 0.988
8	1	0.901 ± 0.003	0.795 ± 0.009
8	2	0.982 ± 0.006	0.958 ± 0.010
8	≥ 3	≥ 0.996	≥ 0.990

Figure 7: Radial Basis Function SVM Performance with Vowel Data

C and γ are chosen low, but not too low as to lower precision. A good elbow point is $C = 2$ and $\gamma = 5$, where their performance scores tend to level off after increasing any more. The precision for execution of the SVM (with Radius Basis kernel) with the chosen optimal hyper-parameters is 0.942, and the ROC AUC is 0.976, which is comparable to the training dataset. This means that this SVM is likely a good generalization for this dataset.

4.3 Satellite Dataset

For this multi-class classification problem, the discriminant hyper-plane is determined from the multiple classes using OVA (one versus all). The K-fold cross-validation mean AUC ROC and precision measurements from the coarse grid search using a polynomial kernel (See file "output/sat/sat_train_poly.txt") are calculated with 95% confidence interval. The performance on this dataset for the polynomial kernel is extremely low and omitted here. For execution with the Radial

Basis kernel (See file "output/sat/sat_train_rbf.txt"), see Figure 8 below.

Hyper-Parameter C	Hyper-Parameter γ	Mean ROC AUC	Mean Precision
≤ 0.1	-	≤ 0.850	≤ 0.703
1	≤ 0.1	≤ 0.844	≤ 0.686
1	1	0.917 ± 0.005	0.807 ± 0.009
1	10	0.972 ± 0.003	0.936 ± 0.006
1	≥ 100	1.0 ± 0.0	1.0 ± 0.0
10	≤ 0.1	≤ 0.868	≤ 0.729
10	1	0.952 ± 0.004	0.880 ± 0.007
10	≥ 10	1.0 ± 0.0	1.0 ± 0.0
100	≤ 0.01	≤ 0.852	≤ 0.699
100	0.1	0.923 ± 0.007	0.814 ± 0.009
100	1	0.982 ± 0.002	0.951 ± 0.005
100	≥ 10	1.0 ± 0.0	1.0 ± 0.0
1000	≤ 0.01	0.878 ± 0.012	0.744 ± 0.019
1000	0.1	0.950 ± 0.003	0.869 ± 0.007
1000	1	$.999 \pm 0.001$	0.997 ± 0.002
1000	≥ 10	1.0 ± 0.0	1.0 ± 0.0
10000	0.0001	0.817 ± 0.005	0.646 ± 0.011
10000	0.001	0.853 ± 0.004	0.701 ± 0.009
10000	0.01	0.924 ± 0.009	0.813 ± 0.012
10000	0.1	0.971 ± 0.004	0.922 ± 0.007
10000	≥ 1	1.0 ± 0.0	1.0 ± 0.0
100000	0.0001	0.831 ± 0.006	0.668 ± 0.011
100000	0.001	0.881 ± 0.012	0.747 ± 0.017
100000	0.01	0.949 ± 0.004	0.863 ± 0.005
100000	0.1	0.994 ± 0.002	0.980 ± 0.003
100000	≥ 1	1.0 ± 0.0	1.0 ± 0.0

Figure 8: Radial Basis SVM Performance with Satellite Data

$C = 1$ is chosen for the fine grid search, along with $\gamma \in \{8, 9, 10, 11, 12, 13, 14, 15, 20, 25\}$, and the average performance metrics is shown, below, in Figure 9.

Hyper-Parameter C	Hyper-Parameter γ	Mean ROC AUC	Mean Precision
1	8	0.967 ± 0.004	0.923 ± 0.009
1	9	0.970 ± 0.004	0.930 ± 0.008
1	10	0.972 ± 0.004	0.936 ± 0.007
1	11	0.974 ± 0.003	0.941 ± 0.006
1	12	0.976 ± 0.003	0.945 ± 0.006
1	13	0.977 ± 0.003	0.949 ± 0.006
1	14	0.979 ± 0.003	0.952 ± 0.006
1	15	0.980 ± 0.003	0.956 ± 0.005
1	20	0.986 ± 0.002	0.968 ± 0.004
1	25	0.990 ± 0.001	0.979 ± 0.002

Figure 9: Radial Basis SVM Performance with Satellite Data

There is no extremely obvious elbow point, here. Instead, a 'best guess' is performed here, with $\gamma = 13$, since there seems to be less increase from $\gamma = 12$ to 13 than from $\gamma = 13$ to 14. To this end, $C = 1$ and $\gamma = 13$ are chosen to run on the testing dataset. The precision from the execution on the testing dataset is 0.783, and the ROC AUC is 0.891, which implies a better fit (slight over-fit) for the training dataset.

5 Conclusion

In general, the Radial Basis kernel SVM has higher performance metrics for lower C values than the polynomial kernel SVM. For the multi-valued categorical datasets, the polynomial kernel does not perform well at all. For the first two datasets (ionosphere dataset and vowel dataset), the hyper-parameters had noticeable elbow points from the grid searches that helps determine the optimal values. Consequently, the execution of the SVM on the testing data with the training data's respective optimal hyper-parameters (from the ionosphere and vowel datasets) yields high performance comparable to the performance of the cross-validation. This indicates a good generalization from the training dataset to the testing dataset. However, the satellite dataset does not have an obvious elbow point in the performance metrics of the cross-validation, which may be the reason the chosen optimal hyper-parameters for this set yields lower performance on the testing dataset than on the cross-validation, indicating an over-fit of the training data.