CMSC 678 - HOMEWORK 3

Akshay Peshave (peshave1@umbc.edu)

PROBLEM 1

Support vector machines seek to find a linearly separating hyperplane for the training data in the input space (linearly separable case) or in a higher dimension feature space (linearly inseparable case) which maximizes the separator margin while minimizing error.

The hyperplane lies at the midpoint of the pair of training samples, one belonging to either class of the data-set, which are separated by minimum distance. If there are multiple pairs separated by this minimum distance then they all contribute to the orientation of the hyperplane without affecting the location of the hyperplane between the two convex hulls enveloping training vectors from either class.

The margin is the distance on either side of the hyperplane before a training vector from the corresponding class is encountered. We define this margin by two planes, the positive and negative plane, on either side of the hyperplane. All points beyond the positive plane traveling away from the hyperplane are those belonging to the positive class. And similar is the case for the negative plane. The input vectors which lie on the positive and negative planes are termed support vectors. It is clear from the above explanation that the hyperplane is indeed defined by these support vectors, which are the vectors from their corresponding classes separated by minimum distance.

Technically the orientation of the hyperplane is defined by the weight vector w which is normal to the hyperplane. For SVMs the weight vector is given as

$$w = \sum_{k=1..R} \acute{\alpha}_k y_k x_k \ , \quad for\ all\ 0 \leq \acute{\alpha}_k \leq C$$

The latter condition ensures that only support vectors contribute to the weight vector definition. Further the margin is defined as

$$Margin\ (M) = 2\ .\ ||w||^{-1}$$

When we remove a support vector :

1. the hyperplane orientation may change
2. the margin may remain the same or increase, but never decrease

The behavior depends on the following cases which may arise:

Case 1: The only support vector existing in one of the classes is removed.
In this case the margin is bound to increase since a the plane on which the removed SV exists need to be moved further away from the hyperplane till the next support vectors are identified.

Case 2: One of several support vectors existing in one of the classes is removed.
In this case the the margin will remain the same, but the orientation and/or position of the hyperplane may change subject to factors such as:

- Number of support vectors remaining in the class and number of them present in the other class. This influences the position of the hyperplane. This can be imagined as the hyperplane being forced away from or more restricted by the class having more support vectors.
- The shortest distance line between any pair of support vectors now possible. This may affect the orientation and/or position of the hyperplane between the convex hulls now formed for each class' training vectors.

These behaviors are depicted in the following diagrams using a fictional data set which is linearly separable in the input space itself.

PROBLEM 2

The polynomial kernel to be used is

$$K(x_i,x_j) = (x_i \,.\, x_j + 1)^2$$
i.e. c=1 and d=2

The kernel function representing this mapping into the feature space is

$$\varphi(x) = [\,1\ \sqrt{2}.x_1\ \sqrt{2}.x_2\ x_1^2\ x_2^2\ \sqrt{2}.x_1\,.\,X_2\,]$$

Below are the values for each of the factors in the mapping function for the corresponding input vectors.

x_1	x_2	$\sqrt{2}.x_1$	$\sqrt{2}.x_2$	x_1^2	x_2^2	$\sqrt{2}.x_1\,.\,x_2$	y
1	1	√2	√2	1	1	√2	+
1	-1	√2	-√2	1	1	-√2	-
-1	1	-√2	√2	1	1	-√2	-
-1	-1	-√2	-√2	1	1	√2	+

The normal *w,* to the hyperplane is given as :

$$w = \Sigma_{k=1..R}\ \alpha_k y_k\ \varphi(x_k)$$

All the instances are going to be support vectors. Thus all instances can be assumed to have the same α values. Thus the weight vector can be rewritten as:

$$w = \alpha\ \Sigma_{k=1..4}\ y_k\ \varphi(x_k)$$

$$w = \alpha\ \{\ 1\ .\ [1\ \sqrt{2}\ \sqrt{2}\ 1\ 1\ \sqrt{2}] - 1\ .[-1\ -\sqrt{2}\ \sqrt{2}\ -1\ -1\ \sqrt{2}] - 1\ [-1\ \sqrt{2}\ -\sqrt{2}\ -1\ -1\ \sqrt{2}] + 1.[1\ -\sqrt{2}\ -\sqrt{2}\ 1\ 1\ \sqrt{2}]\}$$

$$\textit{This reduces to :}\ \ w = \alpha\ .\ 4.\ \sqrt{2}$$

Therefore the hyperplane is defined by the weight vector along the $\sqrt{2}.x_1\,.\,X_2$ axis and is perpendicular to it. This is depicted in the diagram.

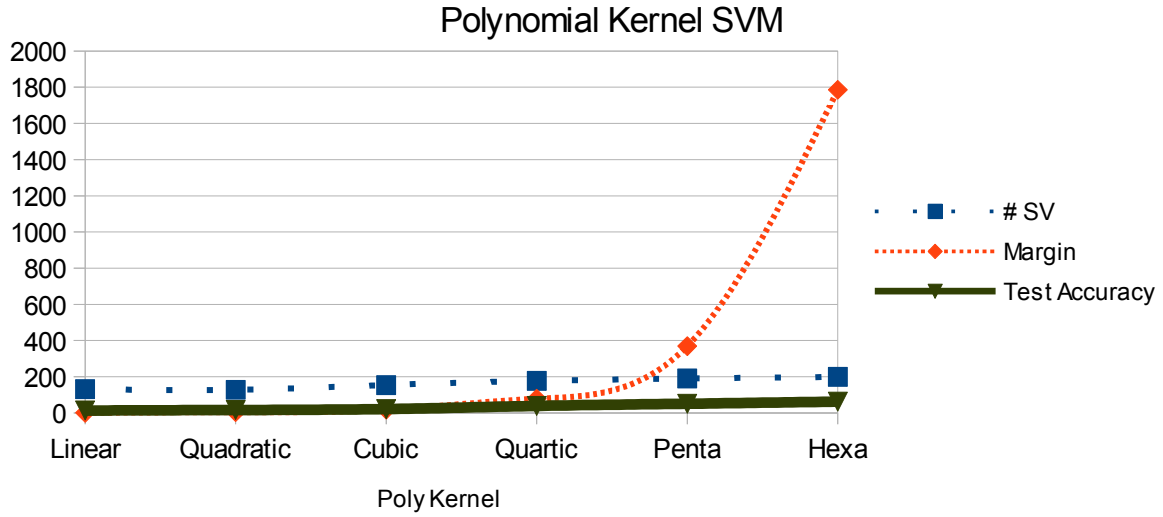
PROBLEM 3

Analysis of the effect of parameters for the various kernels used in conjunction with SVM yield 3 significant observations/trends.

Trend 1: Accuracy and margin width increases as degree of the polynomial kernel increases.
An experiment conducted to measure various parameters of the SVM when degree of the polynomial kernel is increased shows:

1. The accuracy increases.
2. The margin width increases steadily until the degree-5 kernel. For the degree-6 kernel the margin width spikes up.
3. The number of Support Vectors increases steadily.
4. The VC-dim parameter shows a steep rise.
5. The number of training mis-classifications show no appreciable variance.

Poly Kernel	Tr. Misclassifications	# SV	|w|	Margin	VC dim	Test Accuracy
Linear	31	132	1.73107	1.1553547806	99.8879	13.21
Quadratic	23	128	0.47382	4.2210121987	245.26225	16.04
Cubic	22	154	0.11109	18.00342065	444.48063	21.7
Quartic	25	177	0.02546	78.554595444	769.94633	38.69
Penta	36	190	0.00542	369.00369004	1336.87553	50.94
Hexa	44	199	0.00112	1785.7142857	1928.22961	62.26



Obs. 1 & 2 indicate that the training set is linearly separable with the most margin of separation between the two classes in the 6^{th} dimension feature space. It may continue to increase in succeeding degrees too.
Obs. 3 & 4 provide evidence that the increase in accuracy is achieved at the expense of successively increasing complexity of the classifier model, one whose margin whose separator in the input space is increasingly more convoluted.
Obs. 5 suggests that the increased accuracy is not achieved at the expense of neglecting training set mis-classifications but only through increased complexity of the model which helps generalize the

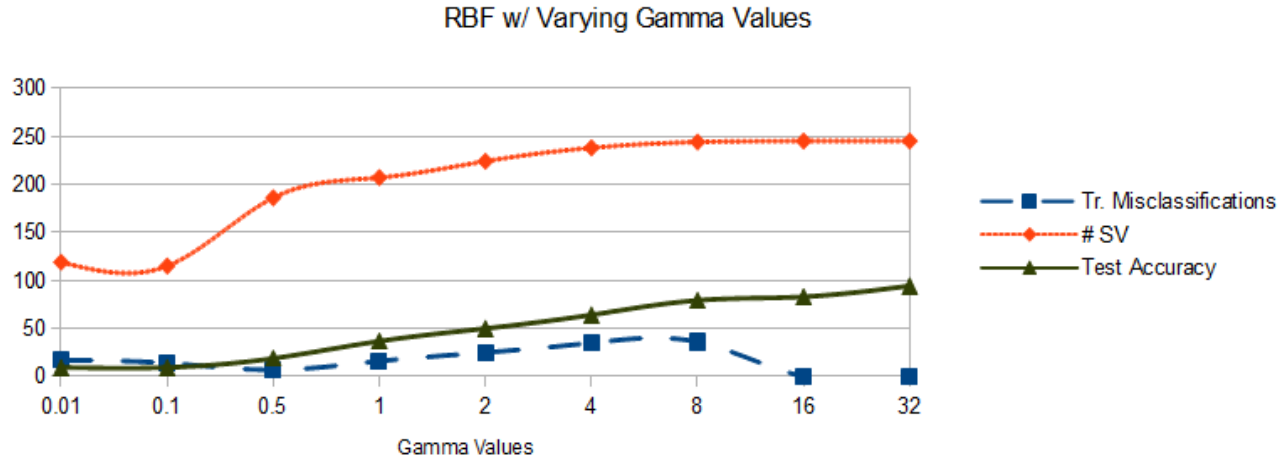
classifier.

Trend 2: Accuracy increases as gamma value of the RBF kernel increases without much increase in complexity of the classifier.

An experiment conducted to measure various parameters of the SVM when gamma value of the RBF kernel is increased shows:

1. The accuracy increases with a decreasing rate of growth showing a trend of converging near (100-ε) percent. Assume ε to be a threshold on the maximum achievable accuracy for the data set, influenced by the presence of noise, probability of not having seen some configuration or input vectors in the training set etc.
2. The margin width show no appreciable variance.
3. The number of Support Vectors increases steadily.
4. The VC-dim parameter shows a steady but not so steep a rise.
5. The curve for training mis-classifications shows no clear trend except that it converges on 0.

gamma	Tr. Misclassifications	# SV	|w|	Margin	VC dim	Test Accuracy
0.01	17	119	14.14857	0.1413570417	110.64015	9.43
0.1	14	115	6.16704	0.3243046907	73.96413	9.43
0.5	7	186	6.80665	0.2938302983	93.66101	18.87
1	16	207	7.44736	0.2685515404	111.92628	36.79
2	25	224	8.27391	0.2417236832	137.91522	50
4	35	238	8.43252	0.2371770242	143.21464	64.15
8	36	244	8.26611	0.2419517766	137.65726	79.25
16	0	245	8.06764	0.2479039719	131.17376	83.02
32	0	245	7.88598	0.2536146427	125.37734	94.34



Obs. 1 & 2 indicate that the training set is linearly separable with the increasing accuracy as the gamma value increases.

Obs. 3 & 4 provide evidence that the increase in accuracy is achieved at the expense of a substantially lesser increase in complexity of the classifier model as opposed to higher degree polynomial kernels.

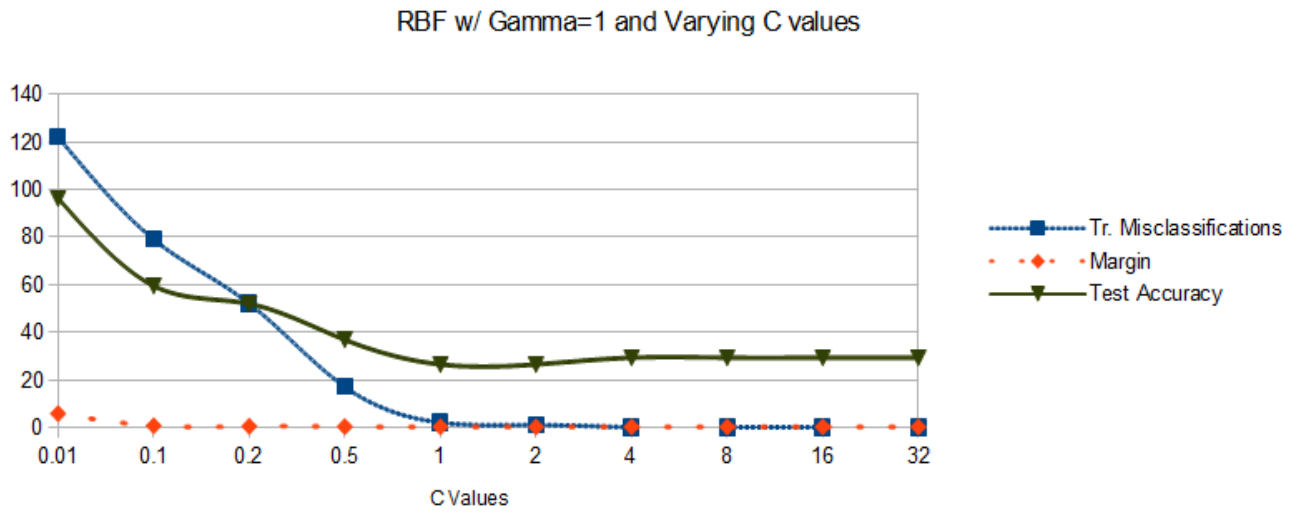
Obs. 5 coupled with Obs. 1 suggests that the model does not exhibit over nor under fitting. The classifier manages to fit the training set really well and also generalize to achieve maximum accuracy on the test set.

Trend 3: Accuracy increases as C value(margin-error tradeoff) of the SVM using RBF (gamma=1) kernel decreases decreasing complexity bound and increasing margin.

An experiment conducted to measure various parameters of the SVM when gamma value of the RBF kernel is increased shows:

1. The accuracy increases with decreasing C value converging near (100-ε) percent.
2. The margin increases steadily.
3. The number of Support Vectors increases steadily.
4. The VC-dim parameter shows a steep decline.
5. The curve for training mis-classifications shows a sharp increase as C-value decreases.

C	Tr. Misclassifications	# SV	|w|	Margin	VC dim	Test Accuracy
0.01	122	244	0.34627	5.7758396627	1.23981	96.23
0.1	79	238	3.05789	0.6540457636	19.70132	59.43
0.2	52	223	4.64863	0.4302342841	44.2196	51.89
0.5	17	207	7.38288	0.2708969941	110.01394	36.79
1	2	192	10.04356	0.1991325785	202.74629	26.42
2	1	190	11.32104	0.1766622148	257.33201	26.42
4	0	190	11.80547	0.1694129925	279.73811	29.25
8	0	189	12.02477	0.1663233476	290.19024	29.25
16	0	189	12.02477	0.1663233476	290.19024	29.25
32	0	189	12.02477	0.1663233476	290.19024	29.25



Obs. 1, 2 and 4 suggest that this model is the best case classifier for this data, if considered alone. Obs. 5 though, serves evidence to refute this claim. It shows that the training mis-classifications increase sharply as the C-value falls below 1. Thus it the increase in accuracy on the test set is based on generalization of the classifier at the expense of reduced accuracy on the training set itself. This may be due to a bad choice of training versus test set or other factors such as noise. This is undesirable behavior. Thus decreased C value results in a model that under-fits the training set. Increased C-value on the other hand results in a model that over-fits, as is evident from the fact that the models for C-value greater than 1 eliminates training set validation errors with no change in the test-set accuracy.

Conclusion:

1. The data set is not linearly separable considering the low accuracy (13.21%) of the linear SVM model and slow growth in accuracy for increasing degree of the polynomial kernels and for the RBF kernel.
2. Choice of the kernel is a trade-off decision involving complexity of the model, margin-size and overall accuracy over training and test sets. The best case choice of classification model based on the above experiments is the SVM utilizing:
    1. RBF kernel
    2. Default C (margin-error tradeoff) value of 0.5
    3. Gamma > 16

    This decision is subject to acceptable VC complexity bounds.
3. The last experiment provides evidence that, in addition to empirical analysis of various kernels and SVM parameters, the choice of training and test set substantially affects the choice and future performance of the classifier model to be chosen.

PROBLEM 4

J-48 decision tree classifier is a high variance, low bias classifier in general. We also know that the J48 decision tree classifier may over-fit the data. By pruning the classifier we eliminate the over-fit to a certain extent. Also, over-fitting causes variance to increase while under-fitting reduces variance and simultaneously increases bias.
Thus un-pruned J48 classifiers will have higher variance and low bias, whereas pruned J48 classifiers will tend to have lower variance and higher bias. Also deeper trees tend to have higher variance and lower bias and vice-versa.

10-FOLD CROSS VALIDATION

Un-pruned	
Method	**Accuracy**
Vanilla J48	91.453
Bagging	92.8775
Boosting	100

Pruned	
Method	**Accuracy**
Vanilla J48	88.0342
Bagging	91.7379
Boosting	97.151

Size of un-pruned vanilla J48 decision tree : 15 → height ≈ 8
Size of reduced-error pruned vanilla J48 decision tree : 9 → height ≈ 3

In brief we state the following about the models:

1. Unpruned J48 classifier is deeper and tends to over-fit. Thus the variance is higher than in the pruned J48 classifier. It has higher Vu.
2. Considering the pruned J48 classifier has lesser depth and exhibits reduced over-fitting, the bias will be higher in case of the pruned classifier. Also Vb will be higher.

Our observations are :

1. The bias is less in case of unrpuned trees causing Vu to be the major factor in reduce accuracy. Bagging increases accuracy by eliminating Vu. Boosting further increases accuracy by reducing bias.
2. In the pruned classifier bias is higher. Thus Vb will be comparatively higher. Thus even after bagging and boosting the accuracy is comparatively lower due to the presence of variance due to bias.

References:

1. *VC dimension tutorial slides by Andrew Moore [Online] : http://www.autonlab.org/tutorials/vcdim08.pdf*
2. *SVM dimension tutorial slides by Andrew Moore [Online] : http://www.csee.umbc.edu/courses/graduate/678/spring13/svm.pdf*
3. *Support Vector Machines, Bennett et. al. [Online] : http://www.csee.umbc.edu/courses/graduate/678/spring13/bennett-svm.pdf*