Machine Learning	**Homework 1**
CSCI 5622 Fall 2017	**Due Time Nov 10, 2017**
Name: Payoj Jain	**CU identitykey: paja3104**

1 Support Vector Machine (50pts)

1.1 Programming questions (20pts)

1.2 Analysis (30pts)

1.2.1 Use the Sklearn implementation of support vector machines to train a classifier to distinguish 4?s from 9?s (using the MNIST data from the KNN home- work).

1.2.2 Experiment with linear, polynomial, and RBF kernels. In each case, perform a GridSearch to help determine optimal hyperparameters for the given model (e.g. C for linear kernel, C and p for polynomial kernel, and C and γ for RBF). Comment on the experiments you ran and optimal hyperparameters you found.

Solution. Below are the optimal hyperparameters I have found using GridSearch for different kinds of kernels.

Linear Kernel : Best C: 0.1, Best Gridscore: 0.9697

Table 1: *Experiments with hyperparameters for Linear Kernel*

Cs	**0.1**	1	10	100

Polynomial Kernel : Best C : 1000, Degree (p) : 2, Best Gridscore :0.9911

Table 2: *Experiments with hyperparameters for Polynomial Kernel*

Cs	1	10	100	**1000**
Degrees (p)	**2**	3	4	5

RBF Kernel : Best C : 100, Gamma (γ) : 0.01, Best Gridscore : 0.9917

Table 3: *Experiments with hyperparameters for RBF Kernel*

Cs	1	**10**	100	1000
Gammas (γ)	0.1	**0.01**	0.001	0.0001

1.2.3 Comment on classification performance for each model for optimal parameters by either testing on a hold-out set or performing cross-validation.

Solution. Parameter C helps in avoiding misclassification, i.e. it adds a slack variable.
For linear kernel, as C increases Gridscore decreases hence, the accuracy decreases. For Polynomial kernel, as we increase C, we are avoiding misclassification even more, hence, as C increases Gridscore should also increase. For RBF, as we increase gamma (γ), variance decreases. A small gamma means large variance and low bias. Thus, higher accuracy for a low bias, high variance model. Thus, for RBF model, low gamma increases accuracy. Table below shows accuracy for different models for best hyperparameters.

Table 4: *Classification performance for each model for optimal parameters on test set*

Model	Linear Kernel	Polynomial	RBF Kernel
Accuracy	97.08%	98.94%	99.01%

1.2.4 Give examples (in picture form) of support vectors from each class when using a polynomial kernel.

Solution. We have binary classes (4 and 9). Support vectors for each class lie on the margin of the classes. Below are the pictorial representation of support vectors from each class.

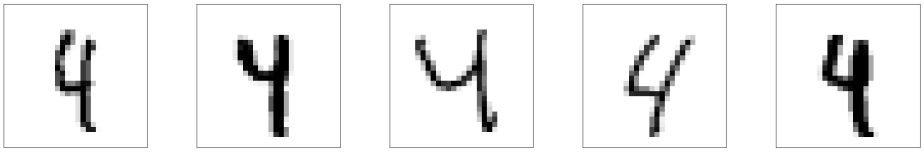

Figure 1: *Support vectors for class 4*

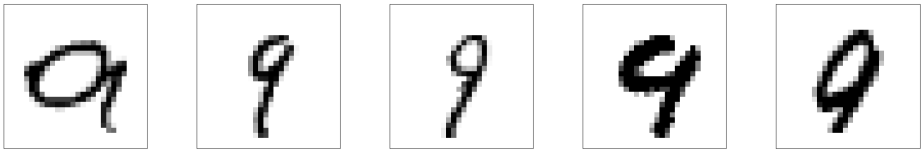

Figure 2: *Support vectors for class 9*

2 Learnability (25pts)

Consider the class C of concepts defined by triangles with distinct vertices of the form (i, j) where i and j are integers in the interval [0, 99]. A concept c labels points on the interior and boundary of a triangle as positive and points on the exterior of the triangle as negative. Give a bound on the number of randomly drawn training examples sufficient to assure that for any target class c in C, any consistent learner will, with probability 95%, output a hypothesis with error at most 0.15.

Solution. Total number of integral points or vertices available in the interval of i,j ([0,99],[0,99]) is 100*100 = 10000. So, total number of triangles we can form out of these 10000 integral points will be choosing 3 different points out of 10000 points, i.e. $\binom{10000}{3}$. Hence, $|H|$ i.e. the hypothesis class size is $\binom{10000}{3}$. Also, the hypothesis class is finite and consistent. Therefore,

$$m \geq \frac{1}{\epsilon}(\ln |H| + \ln \frac{1}{\delta})$$

where m is total number training examples, ϵ is the error rate, δ is the confidence.
We need to find bound on number of training examples sufficient to assure an error of at most 15% with the confidence of 95%. Therefore, ϵ is 0.15 and δ is 0.05.

$$m \geq \frac{1}{0.15}(\ln \binom{10000}{3} + \ln \frac{1}{0.05})$$

$$m \geq \frac{1}{0.15}(\ln \binom{10000}{3} \times \frac{1}{0.05})$$

$$m \geq \frac{1}{0.15} \times 28.834$$

$$m \geq 192.231292635$$

So, a minimum of **193** randomly drawn training examples should be sufficient.

3 VC Dimension (25 pts)

This questions concerns feature vectors in two-dimensional space. Consider the class of hypotheses defined by circles centered at the origin. A hypothesis h in this class can either classify points as positive if they lie on the boundary or interior of the circle, or can classify points as positive if they lie on the boundary or exterior of the circle. State and prove (rigorously) the VC dimension of this

Solution. We have a positive hypothesis class (**A**) that will classify positive if a point is in the interior or boundary of the circle and will classify negative otherwise and also a negative hypothesis class (**B**) that will do the reverse (negative inside or on boundary and positive outside).
For lower bound, we can shatter 2 points using these two hypothesis classes as shown in the figures **A** and **B** below:

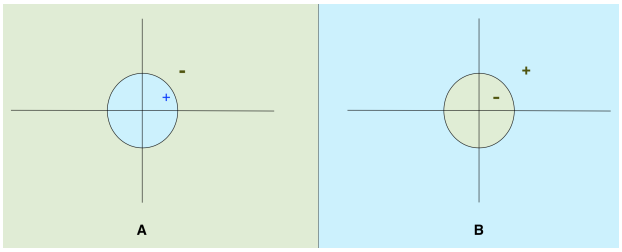

Figure 3: ***A*** *shows that positive hypothesis class can shatter two points when positive point is nearer to the origin. Similarly, negative hypothesis class* ***B*** *can shatter two points when negative point is nearer.*

We take 3 points on the plane such that $||x_1||^2 \leq ||x_2||^2 \leq ||x_3||^2$ and labels of x_1 and x_3 is "+" and of x_2 is "-" or vice-versa. Then, we cannot find a circle centered at origin and include both x_1 and x_3 without including x_2 or include x_2 without including x_1 (in the case of negative hypothesis class) . Hence, we cannot shatter three points. Therefore, VC dimension of this family of classifiers is < 3.

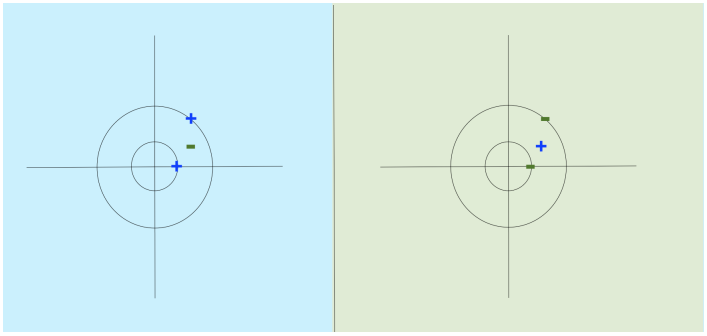

Figure 4: *3 points cannot be shattered using circular classifiers even though we have two sets of hypothesis classes. Hence, VC dimension is less than 3 but greater than or equal to 2.*

EXTRA CREDIT (10 pts): Consider the class of hypotheses defined by circles anywhere in 2D space. A hypothesis h in this class will classify points as positive if they are on the boundary or interior of the circle and classify points as negative if they are on the exterior of the circle.. State and prove (rigorously) the VC dimension of this family of classifiers.

Solution. For lower bound, below figure covers every possible cases for 3 points which can be shattered.

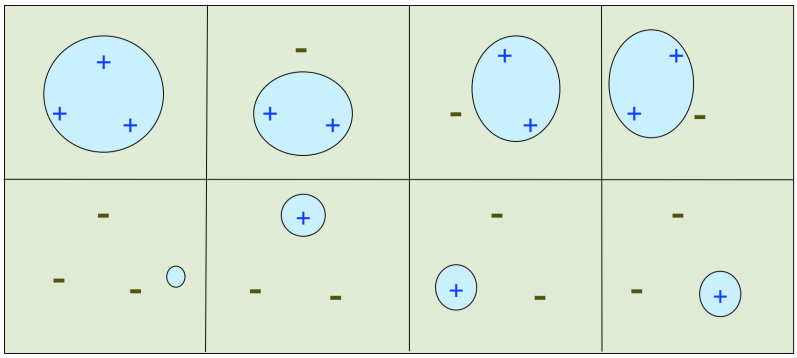

Figure 5: *Every possible labelling can be covered by a circle, so we can shatter 3 points.*

Let's see the case for shattering 4 points.
Radon's Theorem states that, given a d+2 dimensions, the points can always be partitioned into two sets A and B where the convex hulls of A and B have a non-empty intersection, i.e., convex(A) $\cap$ convex(B) ? ϕ
Proof for 4 points in 2 dimension space: Form a triangle with 3 vertices. Then there will be 2 cases for where the final vertex can go - either inside the triangle or in outside of the triangle, since it's a 2D only 1 outer region is possible. Let A be the set of points on the face of intersection and B be the set of remaining points. These CONVEX hull of two sets clearly have a common point.
If either A or B has 3 points, the fourth point will be inside the triangle formed by first three points. If A and B both have two points each then we have a case as shown below. Here, labels of two opposite points are "+" and labels of the other two opposite points are "-".
The smallest circle which passes through both "+" labelled points will be the one which has the line joining these two point as a diameter. Now, the sum of angles of opposite points (labeled as "- ") in the quadrilateral **should be greater than or equal to 180° (Convex polynomial property)**. But if they were outside the circle then their sum would become **less than 180°**. This is a **contradiction**. Thus, there is no circle containing "+" points and not containing "-" points. Hence, VC dimension for a circle anywhere in a 2D space < 4.