Machine Learning	**Homework 4**
CSCI 5622 Fall 2017	**Due Time Nov 10, 2017**

Submission Guidelines

- In order to download files required for the homework, clone `https://github.com/BoulderDS/csci_5622_hws`.
- For programming questions, submit python source files in a zip file.
- For other questions, submit a PDF file of no more than 4 pages.

All homework submissions are done through Moodle.

1 Support Vector Machines (50 pts)

In this homework you'll explore the primal and dual representations of support vector machines, as well as explore the performance of various kernels while classifying handwritten digits.

1.1 Programming questions (20 pts)

Finish `svm.py`.

1. Given a weight vector, implement the *find_support* function that returns the indices of the support vectors.
2. Given a weight vector, implement the *find_slack* function that returns the indices of the vectors with nonzero slack.
3. Given the alpha dual vector, implement the *weight_vector* function that returns the corresponding weight vector.

1.2 Analysis (30 pts)

Use *svm_fours_nines.py* to help answer the analysis questions.

Please do NOT submit *svm_fours_nines.py* to Moodle.
This file is to just help read in data and run the *GridSearch*.

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 homework).
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.

   Hint: http://scikit-learn.org/stable/modules/grid_search.html
3. Comment on classification performance for each model for optimal parameters by either testing on a hold-out set or performing cross-validation.

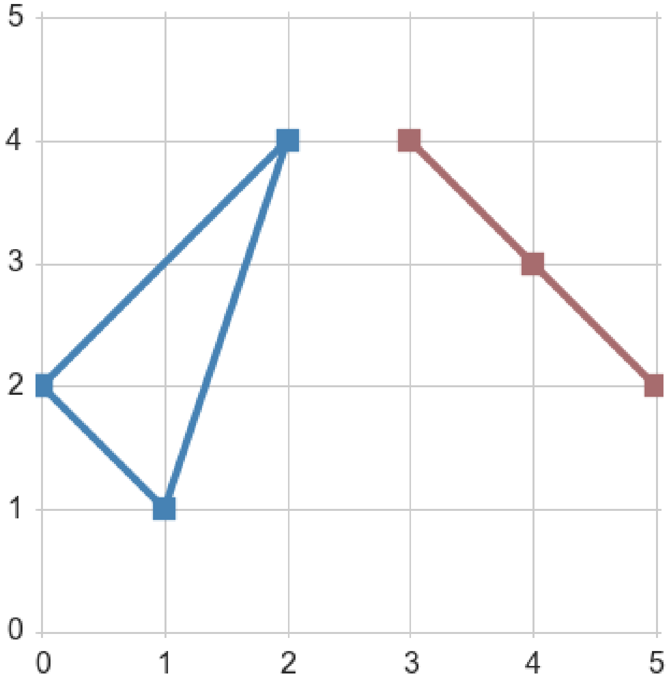
4. Give examples (in picture form) of support vectors from each class when using a polynomial kernel.

2 Learnability (25 pts)

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.

Note: To make life easier, we'll allow degenerate triangles in C. That is, triangles where the vertices are collinear. The following image depicts an example of a degenerate and non-degenerate triangle.



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 family of classifiers.

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 lie on the boundary or interior of the circle, and classify points as negative if they lie on the exterior of the circle. State and prove (rigorously) the VC dimension of this family of classifiers.