CS 596 Machine Learning

**Homework Assignment 6**

Due: Apr 24, 11:59pm, 2018

The purpose of this homework assignment is to practice the classification model: Support Vector Machine (SVM).

**Overview**

In this project, we try to build a SVM classifier that can identify the gender of a crab from its physical measurements. For every crab, there are six physical features: species, front-allip, rear- width, length, width and depth. You will need to train a binary SVM model from a set of training samples, and apply this model over the testing samples to evaluate its performance.

The dataset is provided in the file ‘crab.csv’, which 200 samples with gender labels (1: male, -1: female). The starter codes are provided in the file ‘main\_svm.py’. Please this file following the instructions inside.

The starter codes include five major steps. To implement the training & testing, you are recommended to use the provided SVM demo codes available in blackboard.

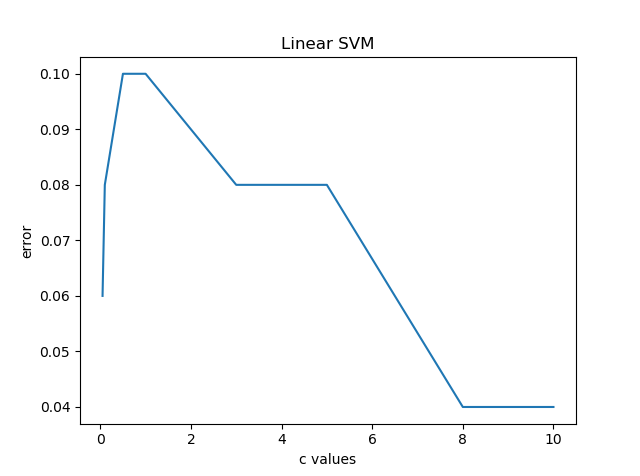
**Step-1**: Load data from ‘crab.csv’ to get feature matrix X and label vector Y. X is of 6 by 200 dimension, where each column represents a crab sample. The matrix Y is of 1 by 200 dimension, including the related gender labels (1 or -1). The starter codes will randomly split these 200 samples into two even subsets: use one for training & validation and another for testing. You don’t need to change the codes in this step.

**Step-2**: Randomly divide the training set into two EVEN subsets: use one for training your model, and another for validation. You will need to implement your codes to do such random splitting. This step includes the “placeholder 1: training/validation”.

Note that, you will need to use the same training/validation/testing splitting (steps 1 and 2) while evaluating different models in the later steps – you can either store and load the splitting data, or make sure all model selection processes are performed with the exactly same training/validation.

**Step-3**: select the optimal model parameters using validation samples. You will need to consider the parameter C (the weighting parameters), kernel types (linear or others) and kernel parameters (if applicable). Please try as many parameters as you can to get the best result. In particular, you will need to generate two figures (in placeholders 2 and 3)

*Figure 1:* **S**election of Cs. Please use different C (e.g., 2, 4, 6, 10) to train the SVM classifier (with other hyper-parameter fixed). For each classifier, calculate its validation error rate (number of misclassified samples over the total number of validation samples). With these results, please generate the following figure:



where the horizontal direction represents different values of C, and the vertical direction represents validation errors. Use at least 3 different values for C.

*Figure 2: selection of kernels* Plot the validation errors while using linear, RBF kernel, or Polynomial kernel (with other hyper-parameters fixed);

**Step-4**: Select the best hyper-parameters (C, kernel types, etc.) and apply them over the testing subset. You may write a script to do the selection, or manually pick up the hyper-parameters based on your results. To do the latter, you might run steps 1 to 3 while temporarily commenting step-4 and step-5.

This step includes the “placeholder 4: testing”.

**Step-5:** evaluate your results using confusion matrix and other metrics, including accuracy, precision and recall rates. Analyze your results through visualizing both success and failure examples. Include 5 success examples and 5 failure examples in your report.

This step includes the “placeholder 5:　metrics” and “placeholder 6:　success and failure examples”.

**How to submit**

* Submit your source codes and report through the SDSU Blackboard. The codes should be self-contained, and run without any error. Otherwise, severe penalty will be applied.
* Two re-submissions are allowed.
* ***No hard copy required.***