Project 5: SVMs

Objectives

To train the SVM algorithm using the provided data set. For this project you can use any statistical package, such as MatLab, R, SAS, or use Python with all the available statistical functions. For a detailed explanation as to how to train the SVM procedure, consult the paper "SVMguide.pdf" posted in Moodle under "Relevant papers".

Data Set

You will use the labeled data set provided to you in Moodle. The data set consists of 200 - 400 data points. The data points are 3-tuples is in the form: (x_1, x_2, ℓ), where x_1 and x_2 are two features and ℓ is the tuple's label, $\ell = 1,2$. The two features are continuous real numbers.

Tasks

1. Color your data set and do a scatter diagram to get an idea of how the data is clustered.
2. Scale the values of each feature in the range [0,1].
3. Apply the SVM method with the penalty cost ξ, i.e.

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \\ \text{subject to} \quad & y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0. \end{aligned}$$

   using the radial base function (RBF) kernel:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2),\ \gamma > 0.$$

4. Obtain the best values for *C* and γ using a grid search in conjunction with cross validation as follows.
5. *Cross validation*: Divide your data set into 5 groups of observations. Round up or down the number of data points in each fold if necessary, so that each fold has an integer number of data points. (It is okay if the fold sizes are not exactly the same.) Since the data points in your data set have been shuffled, it is okay to simply divide the data set into 5 groups of successive data points. Alternatively, you can use a stratified sampling function to generate the 5 folds.
6. *Grid search*: Start with a rough search by varying the two parameters as follows: $C = 2^{-5}, 2^{-3}, ..., 2^{15}$, and $\gamma = 2^{-15}, 2^{-13}, ..., 2^{3}$ (see SVMguide.pdf). Identify a small region of values for *C* and γ for which you get the highest accuracy, and then do a finer search within that range.
7. *The basics of the algorithm are as follows*:
   a. Select values for *C* and γ
   b. Run the SVM algorithm using folds 1 to 4 as the training data set. Use fold 5 as a testing data set to check how many data points were accurately classified.
   c. Repeat (b) with folds 1,2,3,5 as a training data set and fold 4 as the testing data set, then with folds 1,2,4,5 as a training set and fold 3 as the testing data set, and so on until you have used all 5 combinations of folds.
   d. Store the total percent accuracy calculated along with the values for *C* and γ and go back to (a).
8. Plot in 3D all your results, and use colors to indicate regions with the same accuracy.

What to report

1. Give the scatter diagram obtained in task 1.
2. Give the best values of *C* and γ and show how you obtained them. Specifically, a) discuss the range of values you used, b) give the 3D plot of percent accuracy vs *C* and γ, and c) identify the best value(s).
3. Discuss your results.

Grading
80 points: 3D plot and identification of best values for C and γ.
20 points: Discussion of you results