CECS 551 Programming Assignment 4

Dr. Todd Ebert

Due Date: 3:30pm on April 12th

Presenting Your Work.

Submit a paper containing the solutons to each of the exercises found below. The paper should be divided into two sections: Solutions, and Appendix. Each solution should be clearly presented and accompanied by narrative when necessary. Source code should *not* be included in the Solutions section unless explicitly requested. However, all supporting source code should be included in the appendix. Both Solutions and Appendix should be divided into sections, titled EXERCISE 1, EXERCISE 2, etc.. **Important:** credit for an exercise will not be awarded if there is insufficient supporting source code in the appendix.

Exercises

1. Review and download the abalone data set at

   `http://archive.ics.uci.edu/ml/datasets/Abalone?pagewanted=all`

   Use the `e1071` library's `svm` function on the data set with at least 20 different combinations of polynomial degree and C cost. For each combination perform the following: i) 10-fold cross validation, and ii) training accuracy from training over the entire data set. Make a table showing the results. The table should order the combinations by increasing complexity. Note: assume (d_1, C_1) induces a more complex model than (d_2, C_2) iff either $d_1 > d_2$, or $C_1 < C_2$. Highlight the combination that resulted in highest average CV accuracy. Note: the 20 different combinations for d and C should provide a good variation of svm model possibilities. For the best classifier in the table, provide the average distance of the predicted class from the true class. Provide a histogram that shows the frequency of how often a predication is m rings away from the true number of rings, where $m = 0, 1, 2, \ldots, 29$.

2. Consider the following alternative method for classifying the ring count of an abalone. This method uses the following nine binary classifiers: $f_{\leq 9 \text{ vs } \geq 10}$, $f_{\leq 7 \text{ vs } 8-9}$, $f_{\leq 5 \text{ vs } 6-7}$, $f_{8 \text{ vs } 9}$, $f_{6 \text{ vs } 7}$, $f_{10-11 \text{ vs } \geq 12}$, $f_{12-13 \text{ vs } \geq 14}$, $f_{10 \text{ vs } 11}$, $f_{12 \text{ vs } 13}$. For example $f_{\leq 7 \text{ vs } 8-9}$ classifies an abalone data point as either having 7 or fewer rings, or having either 8 or 9 rings. Thus, the training set for this classifier consists of all training points with 9 or fewer rings. As another example $f_{8 \text{ vs } 9}$ classifies an abalone data point as either having 8 rings or 9 rings. Thus, the training set for this classifier consists of all training points with 8 or 9 or fewer rings. These binary classifiers are used to form a classification algorithm that behaves in a manner similar to binary search. For example, on input $\overline{x}$, we first evaluate $f_{\leq 9 \text{ vs } \geq 10}(\overline{x})$. Suppose the output is $+1$ (i.e. $\overline{x}$ is classified as having at least 10 rings). Next we evaluate $f_{10-11 \text{ vs } \geq 12}(\overline{x})$. Suppose the output is -1 (i.e. . $\overline{x}$ is classified as having 10 or 11 rings). Finally, the algorithm evaluates $f_{10 \text{ vs } 11}(\overline{x})$ and returns either 10 or 11 as the final ring classification. In the case where $\overline{x}$ is classified as having 5 or fewer rings, then the algorithm outputs 5. Similarly, in the case where $\overline{x}$ is classified as having 14 or more rings, then the algorithm outputs 14.

   For each of the nine classifiers, use a method similar to that used in Exercise 1 for finding a best svm model for the classifier. Provide a table having nine rows, where each row reports on the best classifier found for each of the nine different classifiers. Each row should include i) a description of the two classes (e.g. "10 vs 11"), ii) size of the training set, iii) degree value of best learning-parameter (BLP) combination, iv) C value of BLP combination v) the average CV accuracy for the BLP combination, and vi) the training accuracy of the final model constructed with the best learning parameters.

3. Implement the binary-search learning algorithm described in the previous exercise, and apply it to the entire abalone data set. Report on the training accuracy and the average distance of the predicted class from the true class. Provide a histogram that shows the frequency of how often a predication is m rings away from the true number of rings, where $m = 0, 1, 2, \ldots, 29$.

4. Use the two-dimensional data in file `Exercise-4.csv` to build a data frame `df`. Use R's `plot` function to visualize the data. Verify that the relationship between x and y appears to be quadratic. Use the `e1071` library's `svm` function (with the following options

   ```
   kernel = ''polynomial'', degree = 2, type = ''eps-regression''
   ```

   held constant) on the data set with at least 20 different combinations of ϵ and C cost. For each combination perform the following: i) 10-fold cross validation of mean squared error (mse), and ii) mse over the entire data set. Make a table showing the results. The table should order the combinations by increasing complexity. Note: assume (ϵ_1, C_1) induces a more complex model than (ϵ_2, C_2) iff either $\epsilon_1 < \epsilon_2$, or $C_1 < C_2$. Highlight the combination that resulted in highest average CV mse. Again, the 20 different combinations for ϵ and C should provide a good variation of svm model possibilities.

5. Provide a graph that shows the plotted data points against the curve provided by the best svm from the previous exercise. Plot the svm model using 1,000 data points equally spaced between 0 to 10. Make sure the plotted data points and plotted model points are clearly distinguishable.

6. Try different combinations of d, C, and ϵ to find a good support-vector regression machine for the abalone data set. Report on the average distance of the predicted class from the true class. Provide a histogram that shows the frequency of how often a predication is m rings away from the true number of rings, where $m = 0, 1, 2, \ldots, 29$.