Machine Learning

Data Set Generation

This assignment consists of theoretical and implementation questions. Question 1 here uses new data. To obtain the new data you need to do the following:

- Go to https://ranger.uta.edu/∼huber/cse6363/Hwk2/Hwk2gen.php
- Enter your student ID number (the 1000... number on your student ID) and hit submit
- Save the generated web page and submit it with your assignment
- Copy the generated data to files on your computer and use them with the corresponding questions

Make sure that you enter your own student ID. Results on data for other student ID numbers will not be considered correct solutions.

Questions 2 and 3 will again consider the problem and data sets from Question 2 in the first assignment where we want to predict the type of material (among 3 material types) of a mug based on four measurements, namely the height, diameter, weight, and hue (color). Assume the same datasets you generated for the first assignment. Make sure that you use the data you got using your own student ID. Results on data for other student ID numbers will not be considered correct solutions.

Support Vector Machines

1. Considering the linearly separable training data generated above.

   a) Formulate the optimization function as well as the constraints for the corresponding linear maximum margin optimization problem without a regularization term. Also show the corresponding Lagrangian as well as the Lagrangian Dual for this problem.

   b) Manually perform 3 iterations of the SMO algorithm on this data. In each iteration you need to pick two α parameters, compute the unconstrained (unclipped) α values that maximize the modified performance function, clip them to make sure that both α values are ≥ 0, and then compute the corresponding decision boundary values w and b. At the end of each iteration, also provide a plot of the data points with the decision boundary described by the values of w and b that you obtained. You do not have to use any specific heuristic to pick the two α parameters in each iteration but picking them according to how likely they are to be on the margin will yield more meaningful initial decision boundaries and allow the algorithm to converge in fewer iterations.

2. Consider again the problem from the previous assignments (Question 2c in Homework 1) where we want to predict the type of material of a mug based on four measurements, namely the height, diameter, weight, and hue (color). To make this into a 2 class problem, we will here consider a "one-against-all" classification scenario where we want to predict whether the material is "Plastic" or not (you do not have to address the other cases of the "one-against-all" classification scenario, i.e. you do not have to learn additional binary "Metal" or not and "Ceramic" or not classifiers). To evaluate the learned classifier you should split the larger dataset from Homework 1, Question 2c) into a test set containing the first 6 examples of each of the materials and a training set that contains the remaining data points.

   a) Use a SVM solver (e.g. MatLab's *fitcsvm* function or for Python Scikit-Learn's *svm* class) to learn the linear SVM parameters for this problem (linear here means that we are not using a kernel function). Since this data is not linearly separable you need to use a non-zero value for the regularization weight C (you can use the default value or experiment with different values to see the differences). Show the classification accuracy you achieved on both the test and the training set and indicate if you think the system is overfitting. Plot the data points and the resulting decision boundary's projection in the 3-dimensional height/diameter/weight space (ignoring "hue"). You can do this by plotting the 2-D projections of this 3D space onto the 2D subspaces height/diameter, height/weight, and diameter/weight. Also identify the support vectors in this problem.

   b) Repeat the classification experiment of part a) but using the SVM with Gaussian Kernels to allow a non-linear decision boundary. In this case you have two parameters, namely the regularization weight C and the standard deviation for the Gaussian Kernels, σ. Again you can use the default values or experiment with different values to see the difference in classification accuracy and overfitting. Indicate the accuracy you achieved on the test and training set and whether you observe overfitting. Compare the results with the ones for the linear SVM and discuss your observations. Also show the classification results by plotting the data points, colored by whether they fall into the positive ("Plastic") or negative class in the 3-dimensional height/diameter/weight space (ignoring "hue"). You can do this again by plotting the 2-D projections of this 3D space onto the 2D subspaces height/diameter, height/weight, and "diameter/weight" (note that since the decision boundary is highly non-linear, you do not have to plot the actual decision boundary which will be between the two colors of the points. Finally, identify the support vectors in this problem.

Decision Trees

3. Consider again the 3-class problem from the previous assignments where we want to predict the type of material (among 3 material types) of a mug based on four measurements, namely the height, diameter, weight, and hue (color). Here we will use Decision Trees to make this prediction. Note that as the data attributes are continuous numbers yo have to use the $\leq$ attribute and determine a threshold for each node in the tree. As a result you need to solve the information gain for each threshold that is half way between two data points and thus the complexity of the computations increases with the number of data items.

   a) Show the construction steps in the construction of a 2 level decision tree using a single step lookahead search and maximum information gain as the construction criterion. You should include the entropy calculations and the construction decisions for each node you include in the 2-level tree. Since the size of the depth-limited search used in the construction of the tree depends on the number of features and the training set size, you should limit the construction to the first 3 features (ignore "hue") and the data to only the first 2 data items for each material type in the data set you generated for Question 2 a) (the smaller data set for manual work) in Homework 1.

b) Implement a decision tree learner for the full classification problem (using all 4 features and the full data set) that can derive decision trees with an arbitrary, pre-determined depth (up to the maximum depth where all data sets at the leaves are pure) using the information gain criterion.

c) Divide the data set from Question 2c) in Homework 1 (the large training data set) into a test set containing the first 6 data items of each material type and a training set containing the remaining data points. Use the resulting training set to derive trees of depths 1 until 8 and evaluate the accuracy of each of the 8 resulting trees for the training samples and for the test set. Compare the classification accuracy on the test set with the one on the training set for each tree depth. For which depths does the result indicate overfitting ?