Data Science Decal Fall 2017 Homework 2

Dataset Credits: I made it

1 Not Quite Plug-and-Chug

There are two datasets included in this week's homework which correspond to the training and test datasets respectively. For each dataset, the first three columns correspond to the dimensions of a data point, and the final column corresponds to the label for that data point. Your task is as follows:

1.1 Tools

- You will be using a specific sklearn model for this assignment. This is the SVC class from the module sklearn.svm. It is a worthwhile exercise to figure out how to import this model and use its associated methods. All of this information will be found on the same webpage, since that webpage contains useful examples too.
- You will need to use matplotlib to visually represent error for a two dimensional grid of values. We recommend googling "imshow matplotlib" and figuring out how to represent a numpy array graphically (grayscale works fine).

1.2 Specifications

In class we covered the notion of hyperparameters and the importance of tuning them for the dataset for the particular problem at hand. The SVM is a good model with which we can see this part of the process at work. Recall that the soft-margin SVM has a hyperparameter called C which denotes the level with which we penalize misclassified points. Also recall that the SVM model can make non-linear decision boundaries by making them in a higher polynomial dimension space and that the degree, d of this polynomial is another hyperparameter to tune.

Your assignment is to train the sklearn model on all combinations of (C, d)
with $C \in \{0.0001, 0.001, 0.01, 0.1, 1\}$ and $d \in \{1, 2, 3, 5\}$.

Additionally, graphically represent the training and test error in grid form so that the error is easily interpretable for a given value of C, d or (C, d) (Two 4 x 5 grids).

1.3 Analysis and Conclusion

The results of the exercise has many insights hidden in plain view. Please answer the following questions.

- On what basis would you decide that a hyperparameter setting is optimal? Which setting of (C, d) gave the optimal results?
- You'll notice that between C and d, one factor mattered far more than the other. What can you conclude about the structure of the datasets and how they were generated?
- With as much granularity as possible, which hyperparameter settings are underfitting and which are overfitting? What allows you to make this claim?

2 K-Fold Cross Validation

This exercise is meant for you to take a well-discussed concept and try implementing it in code. You will only be using the training set of the provided data for this question.

2.1 Tools

You will use the same SVC class as in Question 1 and will likely benefit from having Numpy documentation on hand.

2.2 Specifications

Please perform 7-fold cross-validation on the training set using the SVC configured to use the RBF kernel with the following values of γ (gamma): $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$
In your submission, please provide a table that gives the validation error for each setting of γ used.