CSCI 390 Machine Learning – HW3

DUE: Sunday, November 10 - 11:59PM

In this assignment you will be exploring the application of linear regression and Support Vector Machines.

Part 1 (50 points)

For part 1, you will be exploring linear regression as applied to the California Housing dataset. You can stream this dataset using sklearn functions (`datasets.fetch_california_housing`) just like you did in the previous homework for the Iris dataset. Explore this dataset. There are 8 features describing houses sold in California on a per-block basis, with a target variable of the median house value in that block. Make sure to follow all standard machine learning practices with this dataset, including normalization, train-test split, etc.

Create a linear regressor using `linear_model.LinearRegression` in sklearn. Provide the final accuracy using the score function on the test set. Also print out the coefficient using **`model.coef_`** method. What do these coefficients say about each of the features? Are there features that matter more for the housing price than others? Are there any with inverse relationships?

Part 2 (100 points)

In this part, use the dataset "weatherAUS.csv". You will be developing several binary classifiers, each to predict the binary variable "RainToday". For predictors, you should use all the scalar variables in the dataset: 'MinTemp', 'MaxTemp', 'WindSpeed9am', 'WindSpeed3pm', 'Humidity9am', 'Humidity3pm', 'Pressure9am', 'Pressure3pm', 'Temp9am', and 'Temp3pm'. Perform normalization, imputation and train/test split in the usual manner.

Test several different SVM classifiers. Try the linear, poly (with degree=3), rbf and sigmoid kernels. For each, report the classification rate. Which is the most accurate?

The documentation says that the performance of the RBF kernel depends on the gamma and C parameter values. Use GridSearchCV to find the values of C and gamma that give the highest accuracy (look for "RBF SVM parameters" in the sklearn API documentation for more information). What are the best parameters, and what is the accuracy of the resulting model? Note: using the grid search can take a long time; you will need to subsample the dataset to do this step (but compute the accuracy on the entire test set).

Submission

Your submission will consist of a Word or pdf file, containing the items below, and your source code for all problems. Your Word file should contain:

- Your responses to problems 1-2
- Your complete Python code for parts 1-2