CS436/536: Introduction to Machine Learning
Homework 5

(1) [600 points] Support Vector Machines.

For this question, you will use the data the data you generated in HW2 from the MNIST Digits Dataset for the training and test datasets in `ZipDigits.train` and `ZipDigits.test` respectively with the two features you computed for the problem of classifying 1s vs. 5s.

(a) Use this method for training support vector machines and this method for model selection with cross validation from the `scikit-learn` python library to find the value C for the regularization parameter with the smallest cross validation error using 5-fold cross validation and the training dataset with two features you formed from the data in `ZipDigits.train`. Report E_{CV} for the best value for C.

(b) For the chosen value of C, learn a support vector machine using all of the training data. Compute and report its E_{in}.

(c) Use the test dataset from `ZipDigits.test` to compute E_{test} for the classifier you just learned and report it. Compare it with the results from HW2 using the linear model.

Solution:

a:

```
The best calculated value for C is : 0.1
Cross-validation error for selected best value of C is :  0.9987199967231917
```

b:

```
The In-Sample Error (Ein) for the selected best value of C=0.1 is 0.45863272816214906
```

c:

```
The Test error (Etest) for the selected best value of C=0.1 is 0.4722321110715557
```

(2) [600 points] Support Vector Machines with the Polynomial Kernel.

For this question, you will use the data you generated in HW3 from the MNIST Digits Dataset for classifying 1s vs. Not 1s, where you created $\mathcal{D}$ with 300 randomly selected data points and $\mathcal{D}_{\text{test}}$ consisting of the remaining data points.

(a) Use this method (not the same as the one for the previous question) for training support vector machines using the kernel for the 10-th order polynomial feature transform and this method for model selection with cross validation from the `scikit-learn` python library to find the value C for the regularization parameter with the smallest cross validation error using 5-fold cross validation and the training dataset with two features you formed from the data in `ZipDigits.train`. Report E_{CV} for the best value for C.

(b) For the chosen value of C, learn a support vector machine using all of the training data. Compute and report its E_{in}.

(c) Use the test dataset from `ZipDigits.test` to compute E_{test} for the classifier you just learned and report it. Compare it with the results from HW3 using the linear model with the 10th order polynomial feature transform.

Solution:

a.

```
The best calculated value for C is : 1
Cross-validation error for selected best value of C is :  0.9733333333333334
```

b.

```
The In-Sample Error (Ein) for the selected best value of C=1 is 0.23606666666666667
```

c.

```
The Test error (Etest) for the selected best value of C=1 is 0.21605218536375745
```

Access Link

HarsimranSinghDhillon_HW_5.ipynb