CS304 - Homework 2

The answers of each question are given at the end of each part so that you can check your solutions.

Part 1 – Gradient Descent

Consider the nonlinear error surface $E(u,v) = (ue^v - 2ve^{-u})^2$. We start at the point $(u,v) = (1,1)$ and minimize this error using gradient descent in the uv space. Use $\eta = 0.1$ (learning rate, not step size).

1. What is the partial derivative of $E(u,v)$ with respect to u, i.e., $\frac{\partial E}{\partial u}$?
   (a) $(ue^v - 2ve^{-u})^2$
   (b) $2(ue^v - 2ve^{-u})$
   (c) $2(e^v + 2ve^{-u})$
   (d) $2(e^v - 2ve^{-u})(ue^v - 2ve^{-u})$
   (e) $2(e^v + 2ve^{-u})(ue^v - 2ve^{-u})$

2. How many iterations (among given choices) does it take for the error $E(u,v)$ to fall below 10^{-14} for the first time? In your programs, make sure to use double precision to get the needed accuracy.
   (a) 1
   (b) 3
   (c) 5
   (d) 10
   (e) 17

3. After running enough iterations such that the error has just dropped below 10^{-14}, what are the closest values (in Euclidean distance) among the following choices to the final (u,v) you got in problem 2?
   (a) $(1.000, 1.000)$
   (b) $(0.713, 0.045)$
   (c) $(0.016, 0.112)$
   (d) $(-0.083, 0.029)$
   (e) $(0.045, 0.024)$

4. Now, we will compare the performance of "coordinate descent." In each iteration, we have two steps along the 2 coordinates. Step 1 is to move only along the u coordinate to reduce the error (assume first-order approximation holds like in gradient descent), and step 2 is to reevaluate and move only along the v coordinate to reduce the error (again, assume first-order approximation holds). Use the same learning rate of $\eta = 0.1$ as we did in gradient descent. What will the error $E(u,v)$ be closest to after 15 full iterations (30 steps)?
   (a) 10^{-1}
   (b) 10^{-7}
   (c) 10^{-14}
   (d) 10^{-17}
   (e) 10^{-20}

Solutions: 1-(e), 2-(d), 3-(e), 4-(a)

Part 2 – Support Vector Machines

We will apply soft-margin SVM to handwritten digits from the processed US Postal Service Zip Code data set given in the files "**features_train.txt**" and "**features_test.txt**". Each row consists of three numbers representing: **digit intensity symmetry.**

We will train two types of binary classifiers: i) one-versus-one (one digit class $+1$ and the other digit class -1), with the rest of the digits disregarded), and ii) one-versus-all (one digit class $+1$ and the rest of the digits are class -1).

When evaluating the training error (E_{train}) and test error (E_{test}) of the resulting classifier, use binary classification error.

Pracitical remarks:

(i) You may use the `np.loadtxt()` method to read the .txt files as numpy arrays
(ii) For the purpose of this homework, do not scale the data, otherwise you may get different results.
(iii) Use the **SVC** class of scikit-learn library:
https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
https://scikit-learn.org/stable/modules/svm.html#svm-classification
The Kernel Functions are defined as: https://scikit-learn.org/stable/modules/svm.html#svm-kernels
(iv) In 10-fold cross validation, if the data size is not a multiple of 10, the sizes of the 10 subsets may be off by 1 data point.

Polynomial Kernels

Consider the polynomial kernel $K(\boldsymbol{x}_n, \boldsymbol{x}_m) = (1 + \boldsymbol{x}_n^T \boldsymbol{x}_m)^Q$, where Q is the degree of the polynomial.

Note: Call the `SVC` class using parameters `gamma = 1.0` and `coef0= 1.0` since the polynomial kernel in sci-kit learn is defined as: $K(\boldsymbol{x}_n, \boldsymbol{x}_m) = (\gamma \boldsymbol{x}_n^T \boldsymbol{x}_m + r)^Q$ where r is defined by `coef0`.

```
svm.SVC(C=C_param, kernel="poly", degree=Q_param, gamma=1.0, coef0 = 1.0)
```

1. With $C = 0.01$ and $Q = 2$, which of the following classifiers has the **highest** E_{train} ?
   (a) 0 versus all
   (b) 2 versus all
   (c) 4 versus all
   (d) 6 versus all
   (e) 8 versus all
2. With $C = 0.01$ and $Q = 2$, which of the following classifiers has the **lowest** E_{train} ?
   (a) 1 versus all
   (b) 3 versus all
   (c) 5 versus all
   (d) 7 versus all
   (e) 9 versus all
3. Comparing the two selected classifiers from Problems 1 and 2, which of the following values is the closest to the difference between the number of support vectors of these two classifiers?
   (a) 600
   (b) 1200
   (c) 1800
   (d) 2400
   (e) 3000

4. Consider the 1 versus 5 classifier with $Q = 2$ and $C \in \{0.001, 0.01, 0.1, 1\}$. Which of the following statements is correct? Going up or down means strictly so.
   (a) The number of support vectors goes down when C goes up.
   (b) The number of support vectors goes up when C goes up.
   (c) E_{test} goes down when C goes up.
   (d) Maximum C achieves the lowest E_{train}.
   (e) None of the above.

5. In the 1 versus 5 classifier, comparing $Q = 2$ with $Q = 5$, which of the following statements is correct?
   (a) When $C = 0.0001$, E_{train} is higher at $Q = 5$.
   (b) When $C = 0.001$, the number of support vectors is lower at $Q = 5$.
   (c) When $C = 0.01$, E_{train} is higher at $Q = 5$.
   (d) When $C = 1$, E_{test} is lower at $Q = 5$.

Cross Validation

In the next two problems, we will experiment with 10-fold cross validation for the polynomial kernel. Because E_{cv} is a random variable that depends on the random partition of the data, we will try 100 runs with different partitions and base our answer on how many runs lead to a particular choice.

Note:

You can use the `sklearn.model_selection.KFold` class for this cross validation:

https://scikit-learn.org/stable/modules/cross_validation.html#k-fold

https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html#sklearn.model_selection.KFold

6. Consider the 1 versus 5 classifier with $Q = 2$. We use E_{cv} to select $C \in \{0.0001, 0.001, 0.01, 0.1, 1\}$. If there is a tie in E_{cv}, select the smaller C. Within the 100 random runs, which of the following statements is correct?
   (a) $C = 0.0001$ is selected most often.
   (b) $C = 0.001$ is selected most often.
   (c) $C = 0.01$ is selected most often.
   (d) $C = 0.1$ is selected most often.
   (e) $C = 1$ is selected most often.

7. Again, consider the 1 versus 5 classifier with $Q = 2$. For the winning selection in the previous problem, the average value of E_{cv} over the 100 runs is closest to
   (a) 0.001
   (b) 0.003
   (c) 0.005
   (d) 0.007
   (e) 0.009

RBF Kernel

Consider the radial basis function (RBF) kernel $K(\boldsymbol{x}_n, \boldsymbol{x}_m) = exp\left(-||\boldsymbol{x}_n - \boldsymbol{x}_m||^2\right)$ in the soft-margin SVM approach. Focus on the 1 versus 5 classifier.

8. Which of the following values of C results in the lowest E_{train}?
   (a) $C = 0.01$
   (b) $C = 1$
   (c) $C = 100$
   (d) $C = 10^4$
   (e) $C = 10^6$

9. Which of the following values of C results in the lowest E_{test}?
   (a) $C = 0.01$
   (b) $C = 1$
   (c) $C = 100$
   (d) $C = 10^4$
   (e) $C = 10^6$

Solutions: 1-(a), 2-(a), 3-(c), 4-(d), 5-(b), 6-(b), 7-(c), 8-(e), 9- (c)

References

[1] Ski-kit learn library: https://scikit-learn.org/stable/user_guide.html

[2] Learning from Data course: https://work.caltech.edu/telecourse.html