IE506 Machine Learning: Principles and Techniques **Jan-May 2023**

Assignment 2: Due On 30^{th} March 2023 (11:59 PM IST)

1 Instructions

Answer all questions. Write your answers clearly. You can use `scikit-learn` package wherever required. You can score a maximum of 70 marks in this assignment.

Use different python notebook (.ipynb) files for each question. Use the same random seed value for all questions so that the splits are same in all questions.

Name the .ipynb files as "IE506_rollno_assignment2_q1.ipynb", "IE506_rollno_assignment2_q2.ipynb" and "IE506_rollno_assignment2_q3.ipynb".

Make sure that your answers and plots are clearly visible in .ipynb files.

Create a folder "IE506_rollno_assignment2" and place all your solution .ipynb files in the folder.

If required, answer to Question 2 part (a) can be in a `.pdf` file which should be added to folder "IE506_rollno_assignment2".

Zip the folder "IE506_rollno_assignment2" to create "IE506_rollno_assignment2.zip". Upload the single zip file "IE506_rollno_assignment2.zip" in moodle.

There will be no extensions to the submission deadline.

Note: Submissions not following the instructions will not be evaluated.

2 Questions

1. Consider the data set `Data_Q1.txt` posted in moodle. Each line in the file `Data_Q1.txt` contains details about a particular sample and has the following format:

   `Label FeatureID:val FeatureID:val FeatureID:val ... FeatureID:val`

   For example if the i-th line in the file is of the form `-1 10:1 13:0.5 19:2 135:5` then it means that the i-th sample has a label `-1` and the 10-th feature of i-th sample has a value 1, 13-th feature of i-th sample has a value 0.5, 19-th feature of i-th sample has a value 2 and 135-th feature of i-th sample has a value 5. All other features of i-th sample have value 0.

   Thus the format is a way to sparsely represent the non-zero feature values of certain attributes.

   (a) **[3 marks]** Write code to read the dataset in `Data_Q1.txt` into two `numpy` arrays: X and y containing the features and labels respectively.

   (b) **[2 marks]** Print the number of classes in the data set and the number of samples belonging to each class. Indicate if there is class imbalance issue.

   (c) **[1 marks]** Split the data into two sets such that 80% of the data is considered as set $\mathcal{T}_1$ and 20% of the data is considered as set $\mathcal{T}_2$. Justify if set $\mathcal{T}_1$ and set $\mathcal{T}_2$ have similar class label proportions.

(d) **[16 marks]** Using $\mathcal{T}_1$ as training data, train each of the following algorithms by tuning only the hyperparameters specified below (keep all other hyperparameters fixed to the default values in `scikit-learn`):

i. Logistic regression with L2 regularizer (Hyperparameter: regularization constant C)
ii. Logistic regression with L1 regularizer (Hyperparameter: regularization constant C)
iii. Soft-margin SVM with L2 regularizer (Hyperparameter: regularization constant C)
iv. Soft-margin SVM with L1 regularizer (Hyperparameter: regularization constant C)
v. Kernel SVM with RBF kernel (Hyperparameter: kernel parameter γ)
vi. KNN (Hyperparameter: number of neighbors)
vii. Decision tree (Hyperparameter: min_weight_fraction_leaf)
viii. Random forest (Hyperparameter: number of estimators)

Choose appropriate ranges for the hyperparameters to be tuned. Clearly indicate the range you choose and justify your choice for the range chosen. Tune the hyperparameters using 5-fold cross-validation procedure. If there is class imbalance, you should take care of it during cross-validation and training.

(e) **[5 marks]** Tabulate the accuracy, precision, recall, specificity and senstivity values for training set $\mathcal{T}_1$ and test set $\mathcal{T}_2$ for each model trained in part (d) for the best hyperparameter choices. Discuss your observations.

(f) **[3 marks]** Discuss if the $L1$ regularizers used in logistic regression and soft-margin SVM resulted in sparse models when compared to $L2$ regularizers. Also compare and contrast the performance of the models obtained using $L1$ and $L2$ regularizers. Using these observations, what would you suggest to the practitioner regarding the use of $L1$ regularizer?

2. In this problem we shall apply kernel ridge regression for the data set `Data_Q2.csv` posted in moodle.

(a) **[10 marks]** Recall that in ridge regression, we solve the problem $\min_\beta \frac{\lambda}{2}\|\beta\|_2^2 + \|y - X\beta\|_2^2$. Using the first order optimality conditions, show that $X^\top X\beta + \lambda\beta = X^\top y$ holds, and show that this can be equivalently written as $\beta = X^\top\alpha$ where $\alpha = \frac{1}{\lambda}(y - X\beta)$. Further show that $\alpha = (X^\top X + \lambda I)^{-1}y$. Note that in this relation $X^\top X$ can be effectively replaced by a kernel matrix using the kernel function idea discussed in class as $\alpha = (K + \lambda I)^{-1}y$. Now, represent the inference function $\langle\beta, x\rangle$ using α and hence using kernels. This extension to ridge regression is called kernel ridge regression.

(b) **[1 mark]** Read the data set in `Data_Q2.csv` into a `pandas` dataframe.

(c) **[1 mark]** Perform standardization of each column in the data frame and create a new data frame.

(d) **[1 marks]** Split the data into two sets such that 80% of the data is considered as set $\mathcal{T}_1$ and 20% of the data is considered as set $\mathcal{T}_2$. Justify if the splits $\mathcal{T}_1$ and $\mathcal{T}_2$ have similar spread in `Consumption` column.

(e) **[3 marks]** Using $\mathcal{T}_1$ as training data, train kernel ridge regression model. Use RBF kernel and tune the *gamma* parameter using 5-fold cross-validation.

(f) **[2 marks]** Compute and display the RMSE and R^2 values on the training set $\mathcal{T}_1$ and test set $\mathcal{T}_2$.

(g) **[2 marks]** Consider the original data in `Data_Q2.csv` and load it into a different `pandas` dataframe called `frame2`. Add another column with name `Class` to the data frame `frame2` such that the following hold:

- samples having `Consumption` values ≤ 6500 are labeled as class 1

- samples having `Consumption` values > 6500 and ≤ 7000 are labeled as class 2
- samples having `Consumption` values > 7000 and ≤ 7500 are labeled as class 3
- samples having `Consumption` values > 7500 and ≤ 8000 are labeled as class 4
- samples having `Consumption` values > 8000 and ≤ 8500 are labeled as class 5
- samples having `Consumption` values > 8500 and ≤ 9000 are labeled as class 6

(h) [**2 marks**] Perform standardization of samples in `frame2` belonging to each class separately. Ignore `Class` column during standardization procedure.

(i) [**4 marks**] Split `frame2` into train and test splits $\mathcal{T}_3$ and $\mathcal{T}_4$, such that the samples in $\mathcal{T}_3$ are the same as in $\mathcal{T}_1$. Consider $\mathcal{T}_3$ as training set, ignore the `Consumption` column and considering `Class` as labels, train a kernel SVM model with RBF kernel. Tune *gamma* parameter using 5 fold cross-validation. Take care of class imbalance issues if they exist.

(j) [**5 marks**] Now consider samples belonging to a particular class i in $\mathcal{T}_3$: build a kernel ridge regression model with RBF kernel (ignore the `Class` column for this task). Tune *gamma* parameter using 5 fold cross-validation restricted to samples belonging to only class i. Repeat this for each class. Thus, at the end, for each class i, you would now have a kernel ridge regression model M_i.

(k) [**5 marks**] For testing (or) inference, implement the following procedure: for any sample, first predict the class label as j and then based on the class label j, use model M_j to predict the `Consumption` value. Using this procedure, find the RMSE values for $\mathcal{T}_3$ and $\mathcal{T}_4$.

(l) [**4 marks**] Compare and contrast the RMSE values obtained in part (f) and part (k). Using your observations, suggest when the two-stage approach of classification-followed-by-regression would be useful when compared to the simple regression approach on the full data set.
