E9 205 – Machine Learning for Signal Processing

Homework # 2

February 28, 2025

Due date: March 12, 2025

Analytical part, prepared in writing, can be scanned. This should be attached to the report on the coding part. Finally, a single pdf file containing the response (no in-person handouts) is to be submitted.
Source code also needs to be included.
Name of file should be "Assignment2_FullName.pdf" submitted to teams channel.
Assignment should be solved individually without collaboration with other human or online (GPT-like) resources..

1. **One-class SVM** Let $\boldsymbol{X} = \{\boldsymbol{x}_1,\ \boldsymbol{x}_2, .., \ \boldsymbol{x}_l\}$ be dataset defined in $\mathbb{R}^n$. An unsupervised outlier detection method consist of finding a center $\boldsymbol{a}$ and radius R of the smallest sphere enclosing the dataset in the high dimensional non-linear feature space $\boldsymbol{\phi}(\boldsymbol{x})$. In a soft margin setting, non-negative slack variables ζ_j (for $j = 1, .., l$) can be introduced such that, $||\boldsymbol{\phi}(\boldsymbol{x}_j) - \boldsymbol{a}||^2 \leq R^2 + \zeta_j$

   The objective function in this case is to minimize radius of the sphere with a weighted penalty for slack variables, i.e., $R^2 + C\sum_{j=1}^{l} \zeta_j$ where C is a penalty term for allowing a trade-off between training errors (distance of points outside the sphere) and the radius of the smallest sphere.

   (a) Give the primal form Lagrangian and the primal constraints for the one-class SVM. (**Points** 5)

   (b) Find the dual form in terms of kernel function and the KKT constraints for the one-class SVM. What are the support vectors ? Will support vectors change when $C > 1$ is chosen ? Give a numerically stable estimate of R (**Points** 10)

   (c) For a new data point $\boldsymbol{x}$, how will we identify whether it is an outlier or not (using kernel functions) ? (**Points** 5)

2. Use the following data source for the remaining two questions
   `leap.ee.iisc.ac.in/sriram/teaching/MLSP25/assignments/data/Data.tar.gz`
   Implementing Linear SVMs - 15 subject faces with happy/sad emotion are provided in the data. Each image is of 100×100 matrix. Perform PCA to reduce the dimension from 10000 to K. Implement a classifier on the training images with linear kernel based support vector machine (You are permitted to use tools like *scikit – learn* to perform SVM learning).

(a) Use the SVM to classify the test images. How does the performance change for various choice of kernels, parameter C and ϵ. How does the performance change as a function of K. Make necessary plots and provide your inference.

(b) Compare the SVM classifier with LDA classifier and comment on the similarity and differences in terms of the problem formulation as well as the performance.

(**Points** 15)

3. **Supervised Sentiment Analysis** - Download the movie review data (each line is a individual review)
`leap.ee.iisc.ac.in/sriram/teaching/MLSP25/assignments/data/movieReviews1000.txt`

    a Split the data into two subsets randomly. One for training (first 3000 reviews) and the other for testing (last 1000 reviews).

    b Use TF-IDF features (*scikit-learn* has tools for feature extraction) and train PCA (using the training data) to reduce the data to 30 dimensions. (Use your own code for PCA from the first assignment).

    c Split the training data randomly into set of 2500 for model training and 500 for validation. Train a logistic regression model. Implement the stochastic gradient descent algorithm by hand without using any tools. Show the loss value for each epoch on the training and validation dataset (for 20 epochs). Test on the test data and report the performance in terms of review classification accuracy. Compare the performance for different choices - batch size [32,64,128], learning rate [1e-3,1e-2, 1e-1]. Show the loss curves for each case.

    d Implement the logistic regression with L2 weight regularization. Show the loss curves for regularization coefficient value of $1e-2$, $1e-1$ and 1. Do you see any overfitting/under-fitting for any of these choices. Test on the test data and report the performance on the test set in terms of review classification accuracy.

(**Points** 20)

4. **Neural Networks - Cost Function** - Let us define a NN with softmax output layer and $\{\mathbf{o}_i\}_{i=1}^M$ and $\{\mathbf{y}_i\}_{i=1}^M$ denote the input and targets to the NN. The task is classification with hard targets $\mathbf{y} \in \mathcal{B}^{C\times 1}$, where $\mathcal{B}$ denotes boolean variable (0 *or* 1), and C is the number of classes. Note that every data point $\mathbf{o}_i$ is associated with only one class label c_i where $c_i \in \{1, 2, .., C\}$ classes. The output of the network is denoted as $\mathbf{v}_i$ where $\mathbf{v}_i = \{v_i^1, v_i^2, .., v_i^C\} \in \mathcal{R}^{C\times 1}$ and $0 < \mathbf{v}_i < 1$. The NN cost function can defined using mean square error (MSE).

$$J_{MSE} = \sum_{i=1}^{M} ||\mathbf{v}_i - \mathbf{y}_i||^2$$

Show that the MSE is bounded in the following manner

$$\sum_{i=1}^{M} \left[1 - v_i^{c_i}\right]^2 \leq J_{MSE} \leq 2\sum_{i=1}^{M} \left[1 - v_i^{c_i}\right]^2$$

(**Points** 10)

5. **Learning of NN weights** - Consider a quadratic error of the form

$$J = J_0 + \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}^*)^T \boldsymbol{H}(\boldsymbol{w} - \boldsymbol{w}^*)$$

where $\boldsymbol{w}^*$ represents the minimum of the function and $\boldsymbol{H}$ represents the Hessian matrix which is positive definite. Let the eigenvalues and eigenvectors of $\boldsymbol{H}$ be denoted by λ_j and $\boldsymbol{u}_j$ respectively for $j = 1, .., n$. The gradient descent based update of $\boldsymbol{w}$ is given by

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \eta \frac{\partial J}{\partial \boldsymbol{w}}$$

where t is the iteration index. If the weights $\boldsymbol{w}$ were initialized to $\boldsymbol{w}^0 = \mathbf{0}$, then show that after Q steps of gradient descent, the weights are given by

$$w_j^Q = ([1 - (1 - \eta\lambda_j)]^Q) w_j^*$$

where $w_j = \boldsymbol{w}^T \boldsymbol{u}_j$. Show that as $Q \rightarrow \infty$, $\boldsymbol{w}^Q \rightarrow \boldsymbol{w}^*$ if $|1 - \eta\lambda_j| < 1$.

(**Points** 15)

6. **Implementing BackPropagation**
`leap.ee.iisc.ac.in/sriram/teaching/MLSP25/assignments/EmoData.tar.gz`
The above dataset has training/test subject faces with happy/sad emotion are provided in the data. Each image is of 100×100 matrix. Perform PCA to reduce the dimension from 10000 to $K = 12$. Implement a 1 hidden layer deep neural network (layer containing 10 neurons) to classify the happy/sad classes. Use the cross entropy error function with softmax output activations and ReLU hidden layer activations. Perform 20 iterations of back propagation and plot the error on training data as a function of the iteration. What is the test accuracy for this case and how does it change if the number of hidden neurons is increased to $15, 20, 25, 30$. [Implement backpropagation by hand without using any tool]. (**Points** 20)