MALAVIYA NATIONAL INSTITUTE OF TECHNOLOGY JAIPUR

Department of Computer Science and Engineering

CSP610 Lab Elective

Note: Lab- I Time: 2 hrs Write the program in Python.

1. Consider the following three tables of observations:

Id	Exercise	Family	Risk
1	daily	yes	low
2	weekly	yes	high
2	weekly	yes	high
5	rarely	no	high
5	rarely	no	high
Id	Smoker	Obese	Risk
1	false	false	low
2	true	false	high
2	true	false	high
4	true	true	high
5	true	true	high
Id	Obese	Family	Risk
1	false	yes	low
1	false	yes	low
2	false	yes	high
4	true	yes	high
5	true	no	high

The three tables above (separated by line) list three bootstrap samples, using these bootstrap samples create the decision trees that will be in the random forest model using entropy based information gain as the feature selection criteria. Assuming the random forest uses majority voting, what prediction will it return for the query: EXERCISE = rarely, SMOKER = false, OBESE = true, FAMILY = yes.

2. (15 points, Labs - 2) Handwritten Digits Data: You should download the two data files with handwritten digits data: training data (ZipDigits.train) and test data (ZipDigits.test). Each row is a data example. The first entry is the digit, and the next 256 are grayscale values between -1 and 1. The 256 pixels correspond to a 16×16 image. For this problem, we will only use the 1 and 4 digits, so remove the other digits from your training and test examples. Please submit your Python code implementing the logistic regression for classification using gradient descent.

   - (5 points) Familiarize yourself with the data by giving a plot of two of the digit images.
   - (5 points) Develop two features to measure properties of the image that would be useful in distinguishing between 1 and 4. You may use symmetry and average intensity (as discussed in class).
   - (5 points) As in the text, give a 2-D scatter plot of your features: for each data example, plot the two features with a red redx if it is a 4 and a blue blueo if it is a 1.

3. (20 points, Labs - 3) In this problem you will implement forward and backward propagation

methods for a multi-layer neural network with K hidden layers. Assume that K is a user input less than 10. Implement the networks separately with the following activation functions:

- Sigmoid: Derive the gradient of the activation function. Confirm with numerical differentiation.
- Tanh: Derive the gradient of the activation function. Confirm with numerical differentiation.

Assume that the last layer has a linear activation function and the loss function is $l(y, \hat{y}) = ||y - \hat{y}||_2^2$. Submit your code (along with any instructions necessary to run it), the forward pass outputs at each layer and the gradients of the parameters $\left(W_{ij}^k, b_i^k\right)$. The input, output and the parameters of the network can be found in the MAT file associated with this problem.

4. (20 points, Labs - 2) In this problem you will train a multi-layer neural network to recognize handwritten digits. Use the multi-layer neural network (with ReLU activation) that you implemented in the previous homework. Use 32 nodes in each layer and initialize the weights randomly. The data is also provided to you in a MAT file.

5. Report the training and validation accuracy as a function of iterations (with 5 hidden layers). Report the convergence speed of the training procedure (with 5 hidden layers) for the Stochastic Gradient Descent optimization algorithm.

    - Determine the number of hidden layers required via cross-validation. Report the training and validation accuracy for cross-validation.
    - Finally, report the best test error that you can achieve.

6. (20 points, Labs - 2) Classify the digits data as given for exercise 1 using a Support Vector Machine. Compute the values of W and an offset b, also draw the hyperplane.