Indian Institute of Technology Dharwad

Department of EE/CSE/ME

EE 405: Pattern Recognition and Machine Learning

Total Marks: 20 **Mid-semester Exam** **Date: 23/10/2020**

Instructions:

- All the solutions must be numbered. Solution with missing question number will not be graded. Combine your solutions in appropriate order in a single pdf.
- The final pdf should include link to colab notebook, appropriate block diagram (wherever required), explanations (handwritten), assumptions, equations, and figures. Do not include code in compiled pdf.
- Use the assignment codes and functions wherever required.
- Any genuine queries regarding any question must be posted in moodle. Queries to any of the TAs through WhatsApp or email will not be entertained.
- During the exam period TA's will not reply to any queries regarding the exam (If you feel any data or information is missing make appropriate assumptions and mention the same in your answer script).
- The deadline is strict and submission must be via moodle. Late submissions moodle will be penalized even if delayed by seconds. Submissions via email will not be accepted.

---

1. **(10 marks)**

   a **(5 marks)** Let

$$f_1(x_1, x_2) = x_1^2 + x_2^2 + 2x_1 + 2x_2, -10 \leq x_1 \leq 10, -10 \leq x_2 \leq 10$$

$$f_2(x_1, x_2) = x_1 sin(x_1) + x_2 sin(x_2), -10 \leq x_1 \leq 10, -10 \leq x_2 \leq 10$$

For both the functions:

- Draw the surface using appropriate library and comment on the type of surface (convex/ non-convex).
- Find the minimum value of x_1 and x_2 analytically and verify the same from the plot.
- Use gradient descent and normal equation method, and obtain the minimum value of x_1 and x_2. From the observation clearly illustrate the pros. and cons. of both the techniques.
- Using various learning rates and initialization tabulate the results and justify your observation using appropriate theoretical illustration.

b **(5 marks)** What is the difference between regression and k-means clustering task. Based on the cost function of both the techniques, comment on the convergence (i.e will they reach the global minima). Justify your solution analytically (using appropriate mathematical equations or intuition).

- Take an appropriate example and illustrate the same (coding is expected).
- If you think that any of the methods will not reach the global minima, propose a technique that may help (to some extend) to reach the global minima. Justify your proposed method by extending your earlier illustration.

2. **(10 marks)** Consider credit card fraud detection task (downloaded the data-set from: kaggle-link).

- Follow these steps to pre-process the data and obtain its subset:
  1. Export the data in .csv file using any library of your choice.
  2. As the dataset is imbalanced (i.e less no. of fraud examples compared to genuine), choose randomly the same number of genuine examples as fraud examples and obtain a balanced dataset .
  3. Remove the outliers from the selected examples (use threshold of 2.5 IQR).
  4. Normalize the data (features should have zero mean and unit standard deviation).
  5. Use this data for further processing.
- **(2 Marks)** Use t-SNE plot (refer: t-SNE link) to visualize the subset of data in 2D. Intuitionally comment which clustering algorithm (k-means, GMM, agglomerative hierarchical clustering, DBSCAN) will work better and provide justification for the same using your theoretical knowledge.
- **(2 Marks)** Propose and clearly explain an evaluation measure to evaluate the clustering algorithms. Using a piece of code, prove your intuition of clustering algorithm selection using the proposed evaluation measure (use the same subset of data obtained in the first part).
- **(3 Marks)** Take the whole (imbalanced) data, do the data pre-processing task (outlier removal, data normalization), perform clustering using all the above mentioned techniques. If degradation in performance is observed, comment on the possible causes for it. Propose a possible solution to overcome this issue (without reducing the size of the dataset). Justify the same through a code.
- **(3 Marks)** Split the whole data in training and testing set (90%, 10%, perform class specific division as the data is imbalanced). Use any one of the above mentioned clustering technique and perform classification task (using a piece of python code). In the report, clearly mention the steps involved in training and testing, and the performance measure with appropriate mathematical equations.