Deep Learning & Machine Intelligence Task 1

Hesam Asadollahzadeh, Nesa Abbasi, Hadi Babalou

Deadline: **18th Mordad, 1402**

1. Introduction

In this task, you will extract information from a CSV file. You have to load this file from the link in the kaggle. This file includes some information about Spotify songs.

2. Steps

All of the steps should be done using **vectorization**. Pandas includes a generous collection of vectorized functions for everything from mathematical operations to aggregation. For an extensive list of available functions, check out Pandas docs.

2.1. Read the file and save it as a DataFrame.

2.2. Explore a bit about what train, test, and validation datasets are. Then split the dataset into train and test datasets.

2.3. Check the structure of the data using describe, tail, head and info methods.

2.4. Some columns may not contain meaningful or useful data for prediction; therefore, they should be dropped. Identify and remove these columns from the dataset.

2.5. Some columns are Numerical and others are Categorical. Show which category each column belongs to.

2.6. There are some missing values in the dataset. Explain different methods to handle missing data. Subsequently, handle the missing values using the best method.

2.7. *msPlayed* shows the duration in milliseconds that the track was played. Sort the songs based on their *msPlayed*.

2.8. Show the mean of the *msPlayed* column. do this once with pandas and once with simple loops. compare the execution time of these two approaches.

2.9. Add a new column named *danceable*. If *danceability* in each row is less than 0.5, *danceable* is 0; otherwise, it is 1.

2.10. Visualize the Top 10 *Genres* by their *Energy* level. Only consider genres with at least 10 songs present in the dataset.

2.11. Plot the histogram of the numerical columns in subplots.

2.12. One way to improve the performance of machine learning models is by normalizing the data. Search and figure out why normalization helps the model. After understanding the benefits, proceed to normalize the numerical columns in the dataset.

2.13. Show the correlation matrix of the dataset.

2.14. For each feature in the dataset, calculate the mean and standard deviation separately for the rows where danceable=1 and where danceable=0. Then, plot the probability density function (PDF) of the normal distribution with the calculated means and standard deviations (you can use *scipy.stats* for this part.) Make sure to use different colored curves on a single plot for each feature to compare the danceable=1 and danceable=0 groups.

2.15. Of all the features, select the one that seems most useful for predicting the danceability of a song. Explain your reasoning for choosing this feature by discussing how its distribution differs between danceable and non-danceable songs based on the analysis performed.

2.16. Using the normal distribution plotted for the selected feature, make predictions on the danceability of songs in the test dataset. Specifically, based on the selected feature values of the test set, calculate the probability they come from the danceable distribution versus non-danceable distribution and classify songs based on these values.

2.17. Save the predictions in a CSV file with columns for the song ID and predicted danceable label. Then compare to the true labels in the test set and calculate the accuracy of the predictions.

3. Theoretical Questions

3.1. Suppose that $X_1, X_2, ..., X_n$ form a random sample from a normal distribution for which the value of the parameter μ is unknown. Determine the **maximum likelihood estimator** and estimate of μ.

$$p(x|\mu) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

3.2. Find the **eigenvalues** and associated **eigenvector** bases for each eigenvalue for the following matrix.

$$\boldsymbol{A} = \begin{bmatrix} 10 & -9 \\ 4 & -2 \end{bmatrix}$$

3.3. Find the **rank** of the following matrix.

$$A = \begin{bmatrix} 0 & -1 & 5 \\ 2 & 4 & -6 \\ 1 & 1 & 5 \end{bmatrix}$$

Notes:

- It's recommended to use Jupyter Notebooks for these exercises, mostly because it displays DataFrames clearly.
- Try to use LaTeX to include mathematical formulas and answer theoretical questions in your Jupyter notebook. (Tutorial)
- Please upload your responses in this Google form.

Good Luck!