CH 5650: Molecular Data Science

Project Report

Shashwat Patel (MM19B053)

13th May 2022

1 Note

There are 2 ipynb files in the zip file, one is for data processing and storing, other is for model building.

2 Objective

The objective of this project is to use the structural features of the polymers and using those features, predict the different properties of polymers like glass transition temperature and melting temperature. The Graph convolutional network has been used for featurization purposes and then a artificial neural network has been trained for the predictions.

3 The Data Set

The data has been obtained from PolyInfo, it's an open source database for polymers. The glass transition temperature, melting temperature data has been collected manually for around 200 polymers. The feature matrix and adjacency matrix of each polymer was extracted from the mol file present in the database. The **rdkit** package available in python has been used to process the mol files. Separate csv files for feature matrix and adjacency matrix is created for each polymer. The density was aslo collected but due to insufficient data it is not used. All the data collected belongs to polyamide polymer class.

4 Introduction to GCN

GCN takes as input: An input feature matrix N $\times$ F feature matrix, X, where N is the number of nodes and F is the number of input features for each node and an N $\times$ N matrix representation of the graph structure such as the adjacency matrix A.

At each layer, these features are aggregated to form the next layer's features using the propagation function.

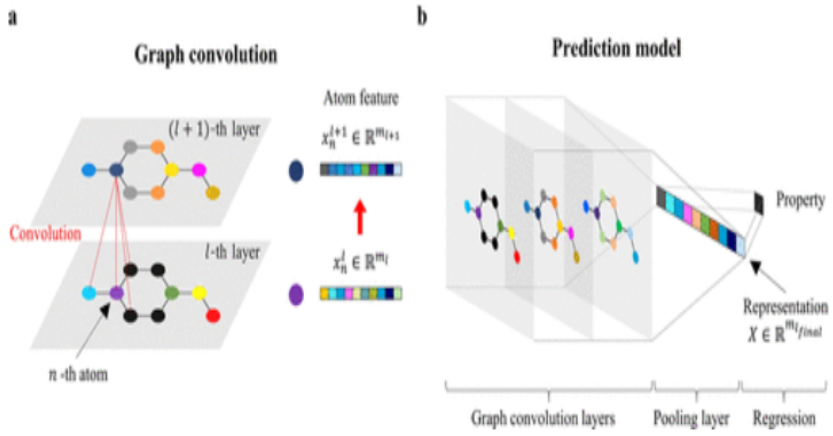
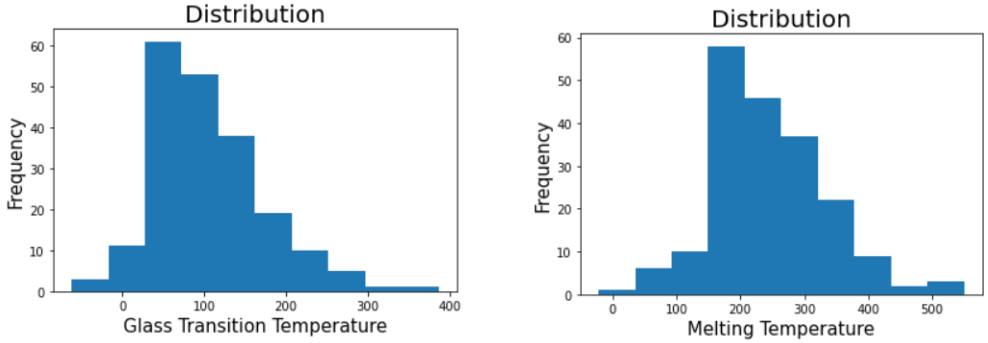


Figure 1: Description of GCN & NN

5 Distributions



(a) Glass Transition Temperature

(b) Melting Temperature

Figure 2: Distributions

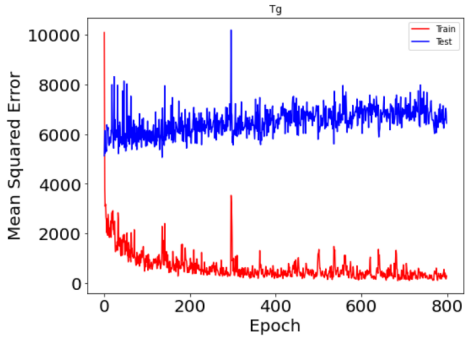
6 Model

For structural featurization, single layer GCN is used and MaxPooling layer is used for dimensionality reduction purposes. The neural network has 4 hidden layers and

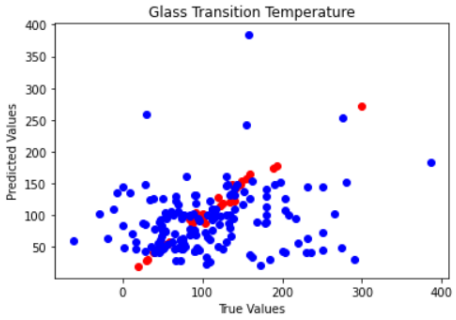
3 dropout layers as well with a dropout rate of 0.2. The initial layer has 500 nodes, the hidden layers have 400, 300, 200, 100 nodes respectively with 'relu' activation function. 800 epochas with batch size of 3 is used for training the model. The dataset has been splitted in the ratio of 0.8 . For both glass temperature and melting temperature prediction, the above-mentioned model has been used.

7 Prediciton Performance

7.1 Glass Transition temperature



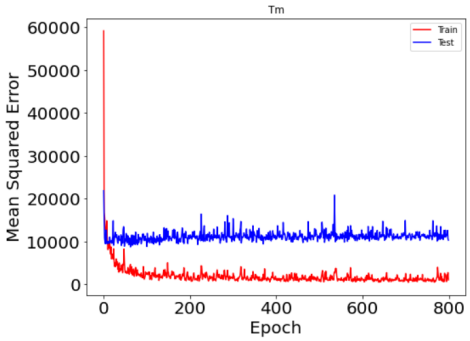
(a) Mean Squared Error Vs No. of epochs



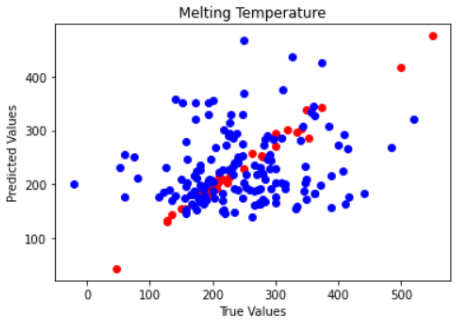
(b) Actual Vs Predicted values

Figure 3: Glass Transition Temperature

7.2 Melting Temperature



(a) Mean Squared Error Vs No. of epochs



(b) Actual Vs Predicted values

Figure 4: Melting Temperature

For both the cases, the model did not perform that well. The model performed comparatively well for prediction of glass transition temperature as compared to

prediction of melting temperature. The test error did not reduce much with increasing epoch in both the cases. The training error did reduce quite a bit. There are many possible reasons of high test error, first being insufficient data. Very less amount of data was used for training purposes, that would have surely an effect. The model's accuracy might also be affected due to presence of outliers and leverage points in the data.

8 Conclusion

The model with a single layer GCN didn't perform well, with high mean squared errors on test set. The prediction performance of the model was better for predicting glass transition temperature as compared with the melting temperature.

For future steps, we should conduct the study once more with larger number of data points. We should be extra careful regarding featurization and try to implement multiple layers of GCN instead of a single layer. We can also try for hyperparameter tuning to optimize the parameters used for the model.