Logo, company name

Description automatically generated

**CSE523 Machine Learning**

**Prof. Mehul Raval**

**Weekly report**

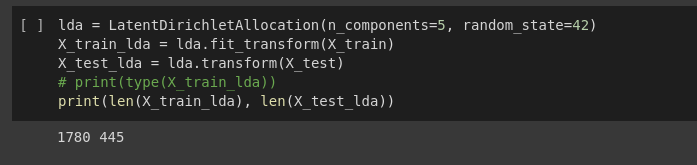
**Group number: 17**

**Group name: The Mandelbrot set**

| **Name** | **Enrolment Number** |
| --- | --- |
| **Aastha Gaudani** | **AU2040032** |
| **Khushi Patel** | **AU2040068** |
| **Devyash Shah** | **AU2040152** |
| **Simran Khoja** | **AU1910606** |

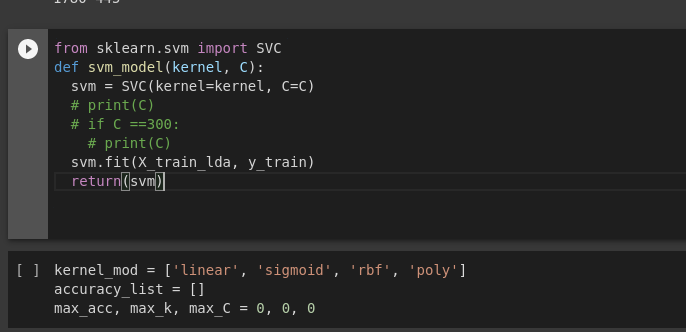
**Weekly Report**

After completing with the **LDA** part, hereby showing the results we move forward to the SVM.



In the Latent Dirchlet Allocation we have given the n\_components to be 5 as there exist 5 categories in our dataset namely entertainment, news, sports, education and business. The random\_state is taken to be 42 as the documentation showed 0 to 42 as the most commonly used values and it proved to give indeed satisfactory results when we kept it to be 42 which acts as the seed value which determines that each category gets involved in the process and the process remains unbiased towards the categories. LDA provides probability values to the words such that the chance of a particular word appearing in a document of a particular category which starts by LDA assigning words to any of the one category at the start and changing the prob values as we move ahead. In short it creates chunks pf the categories. Hence LDA finishes the topic modelling.

**SVM**

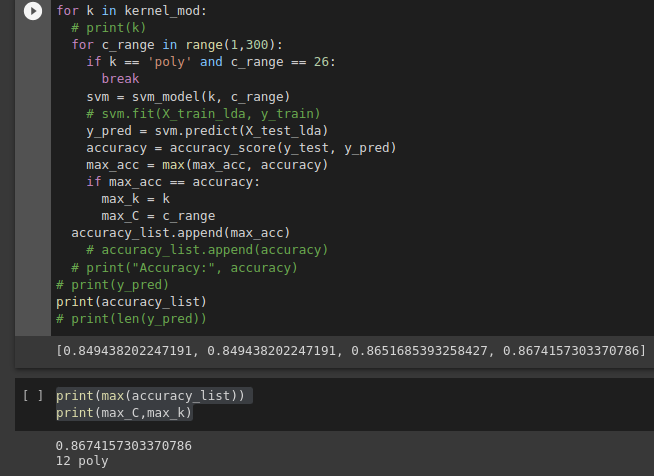


Support vector machine (SVM) classifiers can be trained using the Python function svm model with a specified kernel and regularisation parameter c by using the SVC class from the sklearn.svm package.

For classification problems, the SVC class implements the SVM algorithm. In order to translate input data to a high-dimensional feature space where the SVM can more readily separate the classes, the kernel parameter determines the kind of kernel function to utilise. The C parameter regulates the trade-off between margin maximisation and classification error minimization. A higher C number indicates a smaller margin and more emphasis on reducing the classification error, whereas a lower C value indicates a greater margin and less emphasis on this goal. A linear kernel is specified by linear, a sigmoid kernel by sigmoid, a rbf kernel by radial basis function, and a polynomial kernel by poly. These are some of the kernel operations for SVMs that are most frequently utilised.

The classification accuracy of the SVM models trained with different kernel functions will be stored in accuracy list, an empty list. To ensure that they have some initial value even if there is no SVM, these variables are initialised to 0 at the beginning of the script.

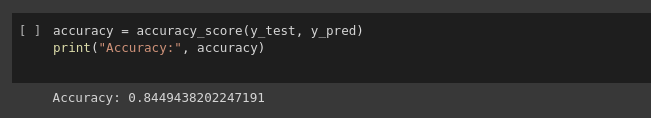
SVM draws out the hyper planes that would separate the data into categories . We have taken the kernal to be poly. Polynomial kernel is a good choice when the decision boundary between classes is curved. It maps the original data to a higher-dimensional feature space, where the classes can be separated by a linear boundary.



The variable max acc stores the maximum classification accuracy attained by any of the SVM models, and the variables max k and max C store the equivalent values for the kernel function and regularisation parameter. After each inner loop for the current kernel function k is finished, the greatest accuracy obtained is added to the accuracy list. This enables us to monitor the highest accuracy attained for each kernel function and to graph accuracy as a function of kernel.

The accuracy list, which contains the highest accuracy across all models (max(accuracy list)), the maximum classification accuracies for each kernel function, and the matching kernel function and regularisation parameter values (max k and max C), is then printed out by the code.

**Accuracy**



Hence we get the accuracy to be 0.84 after implementation of LDA+SVM.

Following are the test dataset implementations.

