Machine learning

18/06/2019

Machine learning for loneliness score prediction

Question: Is the loneliness of students linked to the distance between their domicile and place of study?

the Support Vector Algorithm (SVM) is used in an attempt to classify student data into three classes. Lonely, neutral or not lonely. The result of the classifier would therefore be interpreted as, the input data correspond to a student who is likely to be lonely/ not likely to be lonely. Knowing the domicile and the distance travelled between the domicile and the place of study.

By mapping the loneliness score of the region in England to the student higher education provider location training sets are created. They contain the longitude and latitude of the student, the distance(km) between the domicile and the he provider and a binary representation of the loneliness zscore.

The classes are defined as follow: - class 0: region loneliness zscore <= 0 - class 1: 0 < region loneliness zscore < 1 - class 2: region loneliness zscore >= 1

The Linear Discriminant Analysis (LDA) technique allows to reduce the 3 dimentions(distance travelled, domicile latitude, domicile longitude) data set in a 2D space which can be visualized.

Results: Testing and visualization of the data reveals that the features used in the input, namely, the distance travelled, the domicile latitude and longitude do not describe how lonely a student may be. However the input data lack of resolution in the sense that it only takes into consideration the loneliness zscore of the region in England which induce very large amount of bias. Indeed the visualization tool developed shows that the zscore of the counties in England is highly variable. More importantly the distance travelled is not accurate as the exact location of study is not known, only the region of study is known in the data provided by HESA(Higher Education Statistics Agency) https://www.hesa.ac.uk/data-and-analysis/students/table-11. Further work would include the repeat of the experiment with data containing the exact migration of the students. The impact of features such as the level of study and mode of study should be explored too.

The source code for the trainning set generation and the classifier can be found at https://github.com/mcguinlu/JGI-Comp/tree/master/ml

distance_travelled	domicile_lat	domicile_long	class
9572.332449	-6.86997	-75.045900	1
1536.741614	64.57320	11.528000	1
9618.499896	14.89720	100.833000	2
6500.568201	-2.98143	23.822300	2
90.182577	54.66670	-1.750000	1
135.335597	52.82470	-2.007450	1
119.412442	52.63610	-1.130600	2
167.481670	51.08330	-1.166670	0

Figure 1: *subset of trainning dataset.*

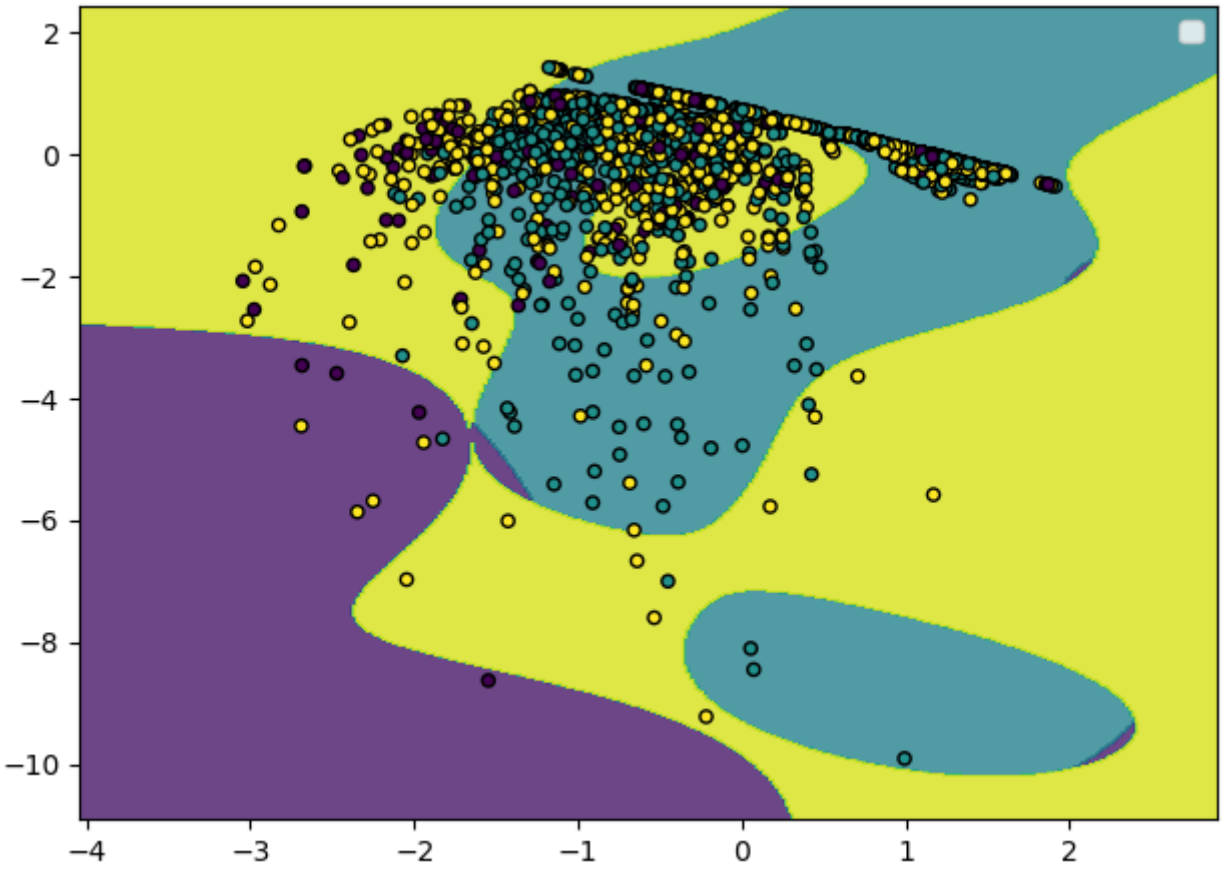

Figure 2: *2D LDA projection of the trainning sets.*