Guessing Students Gender by their Course Plan

Magnin Jonathan, Nonaca Darja, Shmeis Zeinab, Wang Shu
Network Tour of Data Science
Project Report

1 Motivation

Choosing a major field of study can be a difficult decision. Today's college students are encouraged to weigh several factors before choosing an area of focus. Important major considerations include overall program cost, salary expectations, and other social aspects (e.g., gender and nationality). Unfortunately, there has been evidence that the student's gender directly influence his/her selection of courses [5] creating a gender gap in majors. Given that, two central questions are discussed in this report. First, is there a gender difference in the course selection plans of EPFL student? And second, can we predict student's gender given his/her courses?

2 Data acquisition and cleaning

2.1 Data acquisition and data structure

Our data consist of EPFL's students lists available on [1]. We could download the `HTML` pages of students subscribed to each class in EPFL since autumn 2012 until autumn 2019 and then to scrap them using the Python module `html.parser`. To obtain the features like gender, nationality and the study field of each student we accessed the complete student's list with the `Gaspar login` while downloading. Finally we merged the student-class list with the student-features list to obtain the complete data needed for our study. Finally our data frame consists of a collection of EPFL's students since autumn 2012 with the following features:

Name	Class token	Study field	Gender	Nationality

2.2 Data cleaning

To preserve the privacy, the student's names has been encrypted by associating each name to an unique ID. Furthermore, the original data frame contained 2086 courses and had 14872 students! Thus while trying to get the first graphs the computer was very slow and the results were very noisy. We understood that we needed to do some data cleaning. First, we decided to select only the Master students for our study, as in Bachelor the class

choice is very limited. Second, to reduce the number of classes, we computed the number of subscribed students to each class and we computed the box diagram. To filter out classes, we select the threshold corresponding the the 3d quantile, which gave us 241 classes. This step was crucial for the clustering part: as we are grouping students by number of classes that they have in common, the number of classes we are considering is proportional to the number of possible clusters in the graph. For this reason we decided to keep the number of classes relatively small, specially much smaller that the number of students remained in the data frame.

3 Exploration

3.1 Dimensionality Reduction

Our data lies in a high-dimensional manifold: the space of the master classes. In order to be able to observe our graphs, we want to project them in 2D. For this purpose we use three different methods: PCA, Isomap and t-SNE. We expect to find several "potato" shapes accumulating the courses that are in the same study field. As we can see from the Figure 1 t-SNE embedding method is the one that make more sense for us.

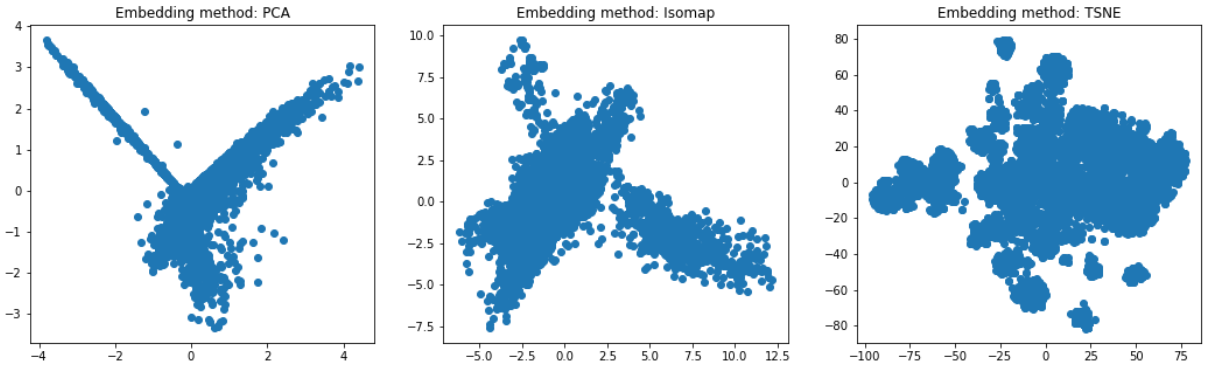

Figure 1: Dimensionality reduction using different methods

3.2 Graph Creation

For our graph analysis, we decided to adapt two approaches. Hence we build two different graphs: Student Social Graph and Epsilon-Similarity Graph.

Student Social Graph: Our first graph is created based on the social connection between students (i.e., based on the number common courses students have). This is obtained by the scalar product between the students course vectors. Initially, the graph consisted of 6212 nodes, 2466115 edges, and a power-law degree distribution with an average of 793.98 signifying a scale-free network. To sparsify the graph we removed all edges with weight less than 1. That is, we consider no relation between two students with only one common course.

This number of edges we reduced to 852869 and the average degree 274.58. However, the network maintained as a scale-free network. We use Gephi to visualize our graph (figure 2). The figure represents the students with their study fields as labels. The clear separation between the majors shows the consistency in our data. More analysis of this graph is shown in the appendix.

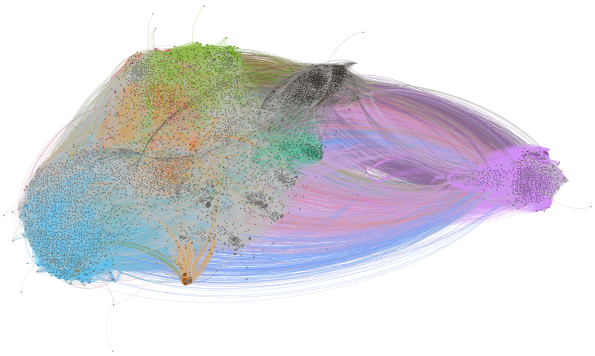

Figure 2: Visualization of the Student Social Graph with Study Field labels

Epsilon-Similarity Graph: This graph is build based on the data obtained from the dimensionality reduction using t-SNE. To obtain the adjacency matrix we first calculate the euclidean distances between the reduced vectors of the students. Then we compute the RBF Kernel using the following formula: $w_{ij} = \exp(-\text{distance}_{ij}^2 / \, 2\sigma^2)$. And finally sparsify it by removing the weights that are smaller than epsilon. This generates a graph with 6212 nodes, 2891065 edges, and a binomial degree distribution with an average of 930.8.

In the Figure 3 we plot the Epsilon-Similarity Graph with the true label of the student's sectin.

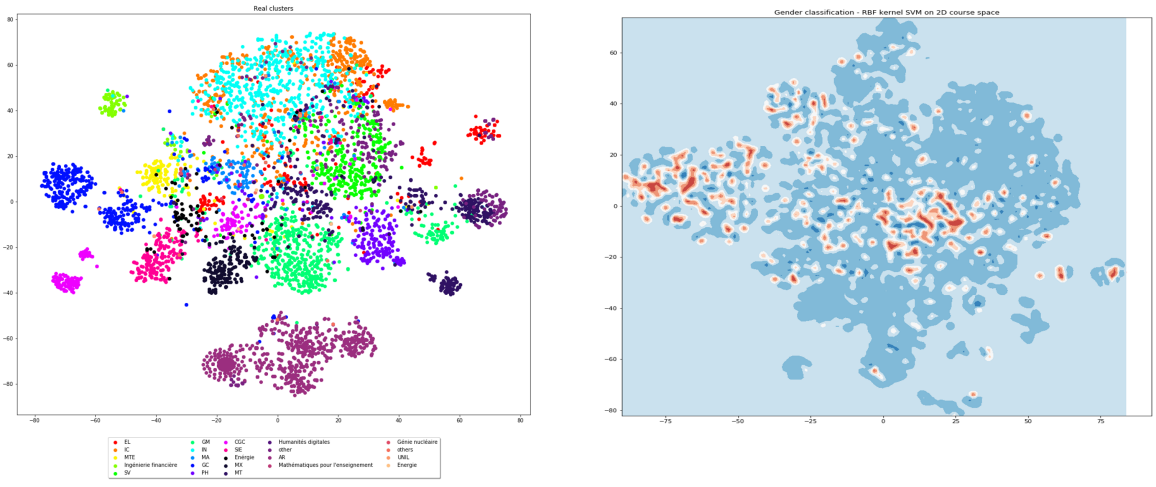

Figure 3: To the left: Reference graph. Master students labelled with real faculty names. To the right: RBF kernel SVM gender classification in 2D "course space" (blue = males, red = females)

We clearly observe what we were expecting: the data form clusters of study fields. The graph is also able to give us further insights by looking at the distance between different clusters: more two clusters are closer, more is similar the education of the students forming them. Also we can see how many clusters exists for each study field (same color on the graph).

4 Exploitation

4.1 Clustering

Without knowing the student's section, we want to see if the clustering algorithms are able to recover the student section by knowing the classes that a student takes. We try two different algorithms: K-means and Spectral Clustering. In Figure 4 we see the comparison of this two methods with the real section labels.

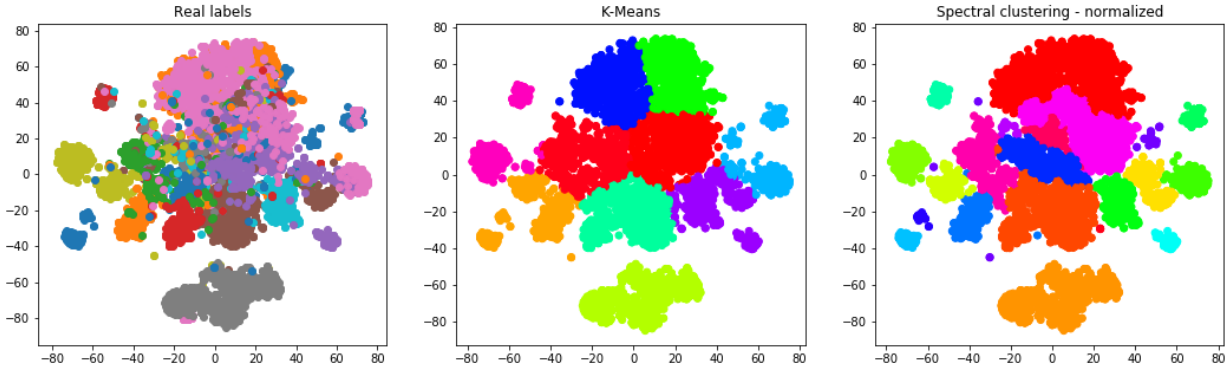

Figure 4: Real labels (on the left), K-Means clustering (in the middle) and Spectral Clustering (to the right).

We conclude that Spectral Clustering is the one that is working better.

4.2 Gender Classification and Prediction

We would like to know if it is possible to predict, with a relative accuracy, the gender of a student given its course choices (i.e. feature vector). To do so, we try multiple methods.

Logistic Regression: Logistic Regression is the most common classification method in the literature. Hence, we started by trying the logistic regression and get a score 0.69%.

Linear and RBF Kernel SVM: A well known way to classify data is the support vector machine (SVM). It has the advantage of being a convex problem and it will therefore be optimally solved efficiently. Moreover, SVM usually needs high dimension data to classify it and uses a kernel to augment this dimension. Since we already have data in high dimension, it is interesting to use a linear SVM directly on the high dimension data and a kernel SVM

(RBF kernel because it works well with irregular shapes) on the reduced data. The 2D SVM classification boundaries are shown in the figure 3.

The figure shows clearly that females are not uniformly distributed over the campus which means that there is a relation between gender and major choice. We want want to know how well this classification can be used to predict the gender of students given their courses. There is one problem with using a SVM on high dimension data. The dimension is defined by the number of available courses and it will change with time since some courses are removed and other are created. Therefore, a model created in year Y may not be applicable to predict on data of year Z. That is why, for the purpose of prediction, we use a SVM classification based on 2D data (with RBF kernel) and test data reduced from high dimension to 2D. Nevertheless, we can compare the performance of linear SVM on high dimension data and RBF kernel SVM on 2D data by testing them with their own training set. This indicates how well the classification fits the data. In the next table we summarize the test accuracy results that we got with these three methods :

ML method	Score
Linear High Dimension SVM	0.6729 %
RBF kernel 2D SVM	0.6951 %
Logistic Regression	0.6729 %

From these results we conclude that all the methods have approximately the same score. The prediction is not good but not bad either: it is high enough to show a correlation between classes token by the students and their gender.

Graph Convolutional Networks: We also tried to see if the structure of the graphs can help in the prediction of student's gender. For this we use the code from [2] and craft it to be used with out built graphs. We get a test accuracy of 68.27% for the social graph and 73.13% for the epsilon-similarity graph. Graph convolutional networks had better scores than the previous methods as they use the graph structure in the learning process. Hence, they captures the structural relations among data which allows it to harvest more insights compared to analyzing data in isolation. Furthermore, comparing the results for both graphs, we see that the the epsilon-similarity graph performs better than the social graph since it is more dense than the latter thus it contains more information.

5 Conclusion and Future Work

In this report we proved that there is a correlation between the student class choices and his/her sex. We used simple ML algorithms, that could be calibrated further to improve the results. For example the dataset could be balanced on as we are dealing with an imbalanced dataset. To do this we could weight the classes, making the model preferential to the minority class. We could also try to predict the minority class using "outlier detection" techniques [3], where the graphs structure play a fundamental role. Furthermore, a potential approach that could be interesting to tackle would be to use and compare the different student graphs constructed for different years in order to study and predict the changes in the gender gap over time.

References

[1] Accès public sur les études bachelor et master. https://www.epfl.ch/campus/services/ressources/is-academia/acces/accesspublic-bachelor-master/. Accessed: 10-01-2020.

[2] Machine learning on graphs. https://github.com/mdeff/ntds_2019/blob/master/assignments/2_learning_with_graphs.ipynb.

[3] Outlier detection techniques. https://towardsdatascience.com/outlier-detection-with-one-class-svms-5403a1a1878c.

[4] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[5] Basit Zafar. College major choice and the gender gap. *Journal of Human Resources*, 48(3):545–595, 2013.

A Student Social Graph Analysis

To check the presence of a gender gap at EPFL students network we plot the social graph with the student's gender labels using Gephi (figure 5a). The green color represent male students and the pink represent females. According to graph there a huge difference in the male and female enrollment.

We did another analysis to validate our social graph were we used the Louvain method [4] from Gephi to detect the different communities in the graph. We observe in figure 5b that the obtained communities properly reflects the students' distribution over EPFL major in figure 2.

(a) with gender labels (b) detected communities

Figure 5: Visualization of the Student Social Graph