

Guessing Students Gender by their Course Plan

Magnin Jonathan, Nonaca Darja, Shmeis Zeinab, Wang Shu

EPFL

A Network Tour of Data Science

January 22, 2020

Project Overview

- Project Goals:
 - Understand the gender distribution of EPFL students
 - Predict Student's gender given his/her course plan

Project Overview

- Project Goals:

- Understand the gender distribution of EPFL students
- Predict Student's gender given his/her course plan

- Roadmap:

- Extract Students features (courses), gender and section
- Reduce dataset dimension while keeping relevant information
- Observe gender distribution
- Classify and predict students gender based on courses

Dataset

- ① Parse the EPFL public and private access to students lists

The screenshot shows a web-based application for searching student enrollment lists. At the top left is the EPFL logo. To its right is a header bar with the text "Liste des étudiants inscrits par semestre". Below the header is a sub-header "Liste des étudiants inscrits par semestre". A "Format:" label is followed by two radio buttons: one selected for "html" and another for "xls". There are four dropdown menus: "Unité académique", "Période académique", "Période pédagogique", and "Type de semestre". A small "ok" button is located at the bottom left of the form area.

Name	Class token	Study field	Gender	Nationality
------	-------------	-------------	--------	-------------

Dataset

- Parse the EPFL public and private access to students lists

The screenshot shows a web-based search interface for EPFL student records. At the top left is the EPFL logo. To its right is the text "EPFL" and "École Polytechnique Fédérale de Lausanne". Below this is a black header bar with the white text "Liste des étudiants inscrits par semestre". Underneath the header, there is a sub-header "Liste des étudiants inscrits par semestre". A "Format:" label is followed by two radio buttons: one selected (blue outline) labeled "html" and another unselected (grey outline) labeled "xls". Below these are four dropdown menus: "Unité académique", "Période académique", "Période pédagogique", and "Type de semestre". At the bottom left of the form area is a grey "ok" button.

Name	Class token	Study field	Gender	Nationality
------	-------------	-------------	--------	-------------

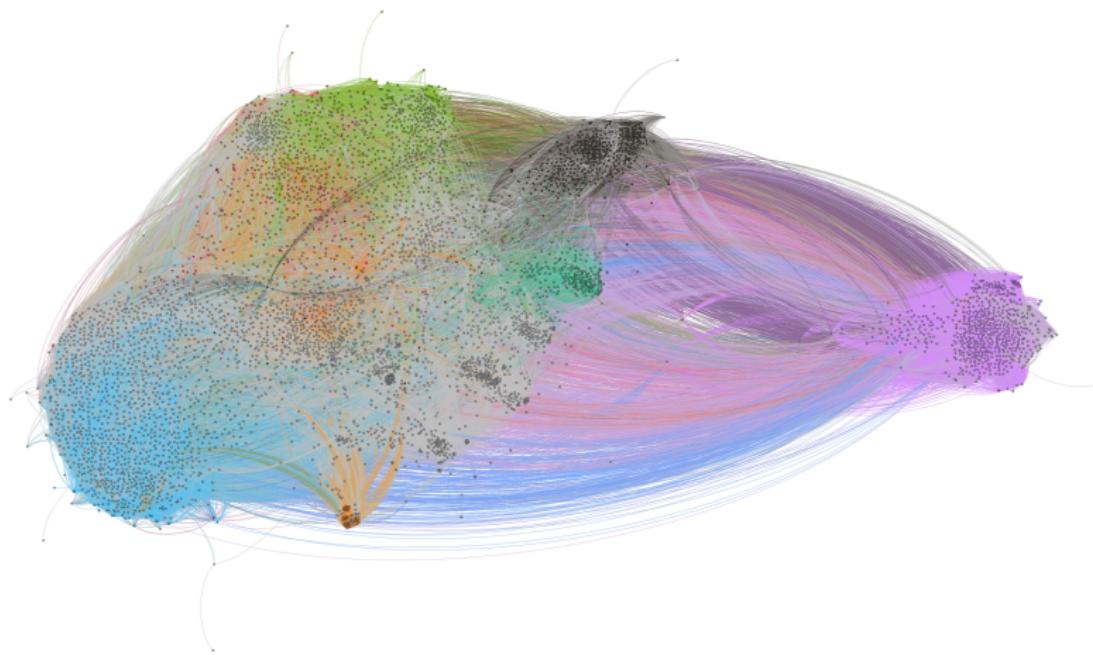
- Clean and filter data:

- select only master students
- filter classes with low participants

student_id	course ₁	...	course _n	Study field	Gender
------------	---------------------	-----	---------------------	-------------	--------

Student Social Graph

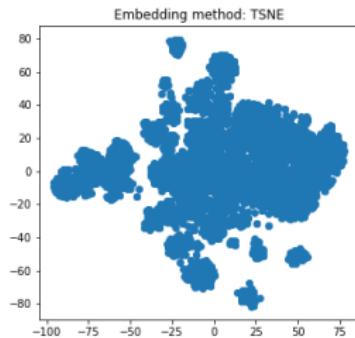
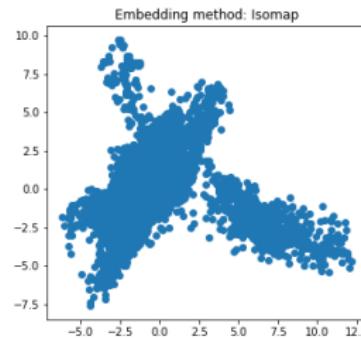
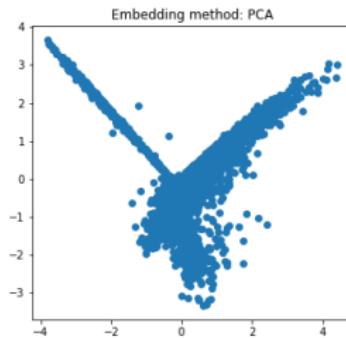
- Each node represents a student
- Students are connected based on the number of common courses



Dimensionality Reduction

Why ?

- to visualize data and for more efficient processing

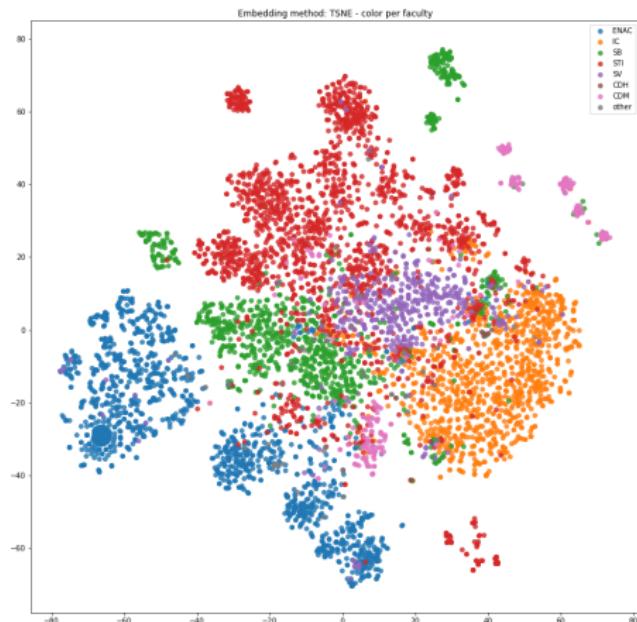


Best method: **t-SNE**

- dataset contains many clusters and is crowded

Dimensionality Reduction

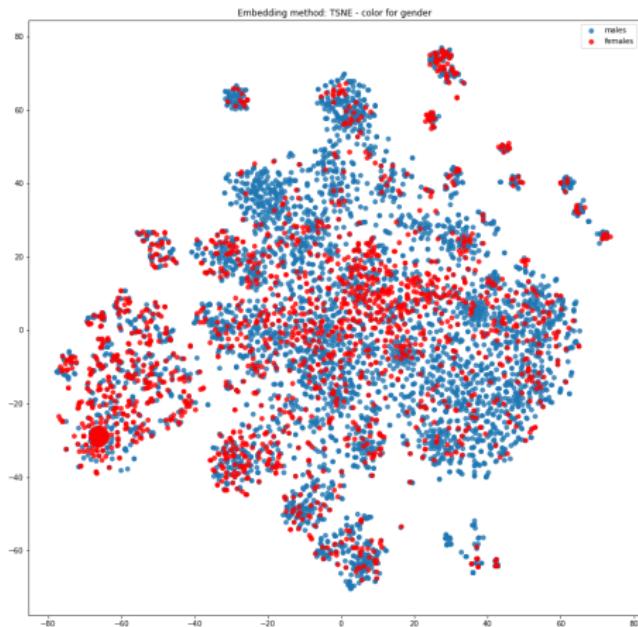
t-SNE embedding with faculty labels



clear relevant clusters

Dimensionality Reduction

t-SNE embedding with gender labels



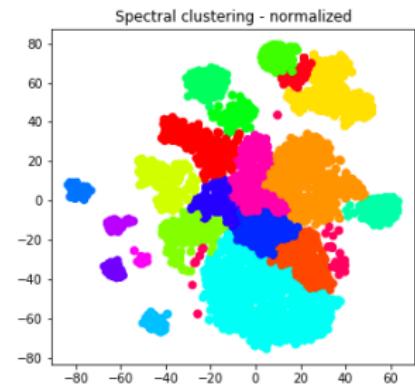
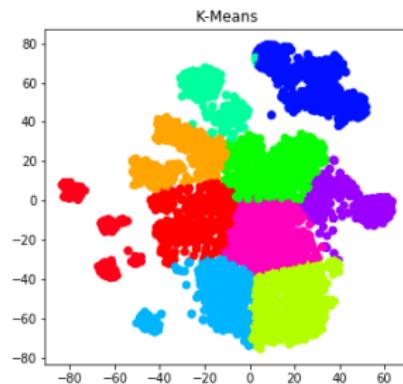
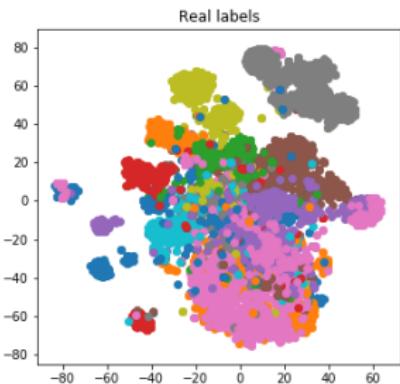
there are males and females everywhere (but no indication on quantity)

Epsilon-Similarity Graph

- Built using the dataset of reduced dimension
- Graph Construction:
 - ① compute euclidean distances
 - ② apply RBF Kernel
 - ③ sparsify using the 3rd quantile

Clustering

Done on Epsilon-Similarity Graph



Gender Classification and Prediction

Classification with Logistic Regression :

Why ?

- Most common method
- Easy to implement

Result : 67.29 %

Gender Classification and Prediction

Classification with SVM :

Why ?

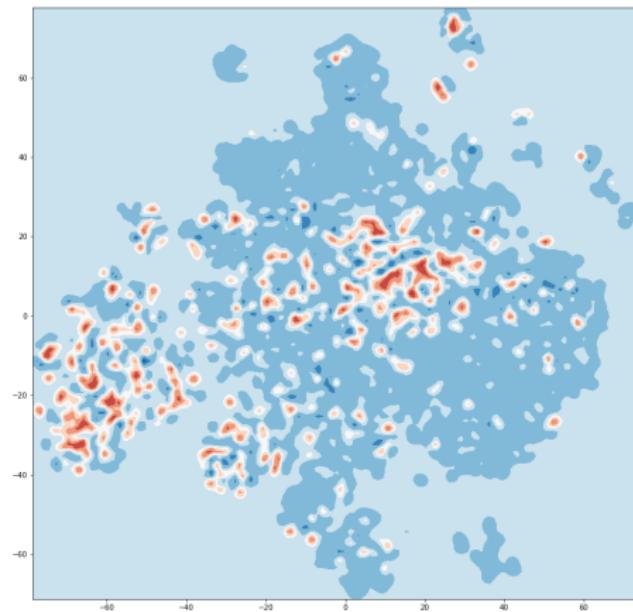
- Convex problem \Rightarrow solved efficiently
- Works well on high dimension data

How ?

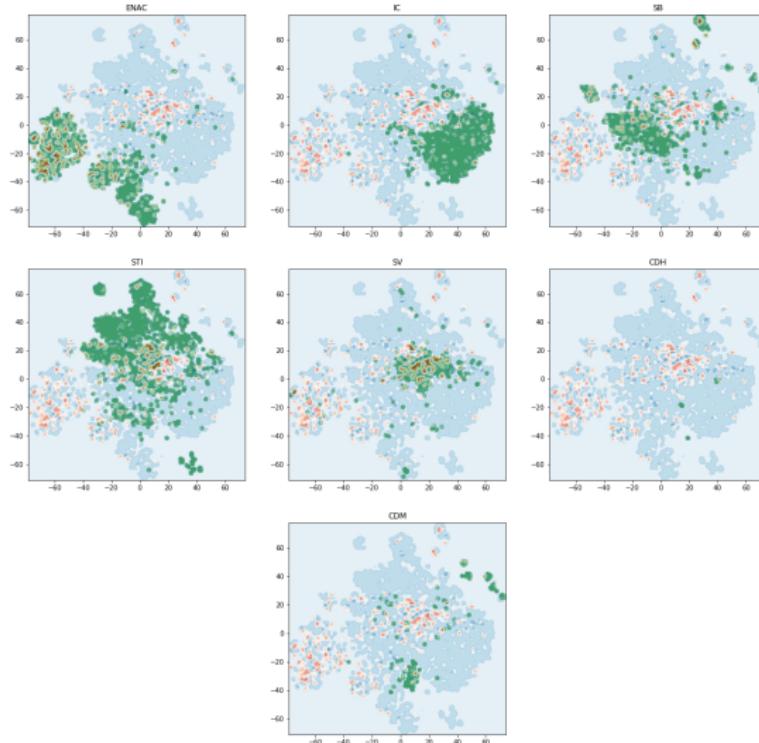
- Linear SVM on high dimension data (no visual)
- RBF SVM on 2D data, assumed as equivalent to previous SVM (visualizable)

Gender Classification and Prediction

Classification with SVM, visualization :



Gender Classification and Prediction



Gender Classification and Prediction

Gender Prediction with SVM, results :

Method	Score
Linear SVM in high dimension	67.29 %
RBF SVM in two dimensions	69.51 %

Gender Classification and Prediction

Classification with Graph Convolutional Network :

Why ?

- capture the structural relations among data
- gives more insight than analyzing data in isolation

How ?

- Approximate GFT with a polynomial of degree 3
- Learn on 1000 epochs

Results :

Graph	score
social students graph	68.27 %
epsilon similarity graph	73.13 %

Conclusion on Gender Distribution

- Females minority among master students
 - less critical in SV and AR sections¹
- The distribution is explained by the psychological differences between males and females regarding STEM [2]:
 - females tends towards organic STEM and arts
 - males favor mathematics, mechanics and engineering
 - differences are amplified in societies where individuals are free to choose their field of career.
- Conclusion: there is no apparent bias on the relation between gender and choice of major in the scope of EPFL.

¹confirmed by EPFL statistics [1]

Conclusion on Gender Prediction

- SVM prediction is not bad but cannot be used as a reliable result
- Graph Convolutional Networks is almost equal
- Conclusion: our gender prediction accuracy is limited
 - predicting a female in a male dominated section is difficult
 - wrong prediction of females as males
- Improvements: for better minority detection:
 - use different weights for features based on the locally dominant gender of a course
 - the "cluster membership" information is useful for this purpose

References

- [1] EPFL internal statistics (GASPAR required),
<https://tableau.epfl.ch/#/home>
- [2] David Geary (2018), "Sex Differences in Science, Tech, Engineering, and Math",
www.psychologytoday.com/us/blog/male-female/201911/sex-differences-in-science-tech-engineering-and-math

Thank You

True labels by class

