

Машинное обучение

Данный курс является, в некотором смысле, продолжением эконометрики. Курс будет состоять из двух независимых частей, программа каждой из них представлена далее. В первой части будут рассматриваться методы работы с пространственной и географической информацией, а во второй – некоторые методы машинного обучения, решающие задачи классификации, регрессии и кластеризации.

Общая оценка за курс будет на 50% состоять из оценки за первую его половину и на 50% из оценки за вторую. Первая половина будет преподаваться на английском языке, а вторая - на русском.

* Данная программа является предварительной и может быть изменена к моменту проведения курса.

Machine Learning, First Half: Geographic Information Systems for Economists

Prof. Andrea Matranga

A geographic information system (GIS) is a system designed to capture, store, manipulate, analyze, manage, and present all types of spatial or geographical data. The objective of this course is to teach students how to integrate GIS in the standard applied econometrics workflow, so as to be able to use spatial within their own research projects, as research assistants for faculty, or within industry jobs.

By the end of the course, students will be able to acquire data from a wide variety of electronic and paper resources, organize it in a consistent and workable format, design the geographic component of an empirical strategy, apply the most commonly used transformation and analysis techniques, and output the data in a format which will allow them to continue their analysis in STATA or any other statistical package.

GIS-literate Research Assistants are generally in high demand in academic departments, so the course is particularly recommended for students that plan to pursue a PhD. GIS programs are also used in a wide variety of industry analytical applications.

Here are some examples of tasks you will be able to perform after completing the course.

1. Download a panel dataset of rainfall. Scan and digitize a map of US railways from 1870, and correctly project on top of the rainfall data. Find all areas and times in the US that were within 20km of a railway, and suffered a drought between 1870 and 1880. Map the results for both black and white and color publication.
2. Import four or five datasets which use different formats and projections. Harmonize them in order to use them for analysis. Upload them in a way that an entire workgroup can use them.
3. Construct a plausible instrument for which European areas should see more invasions through history, given the position of major rivers and mountain ranges.
4. Replicate the Analysis from Alesina Giuliano and Nunn fertility and the plow paper.

Evaluation

50% of the grade will come from weekly problem sets, and 50% will come from a end-of-term individual project.

Topics

1. Introduction to GIS data and formats
2. Creating simple maps
3. Using the model builder and ArcPy
4. Importing and converting data
5. Georeferencing and geocoding paper data
6. Analyzing and transforming data
 - a. Distances
 - b. Buffers
 - c. Slope and Elevation
 - d. Cost Paths
 - e. Network Analysis
7. Datasets commonly used in economics
8. Exporting data

Машинное обучение, вторая половина

Екатерина Лобачева

Во второй части курса будут рассмотрены основные алгоритмы машинного обучения, которые могут быть полезны для обработки данных, в том числе экономических. Будут рассматриваться более сложные методы классификации, регрессии и анализа взаимосвязей в данных.

В настоящее время машинное обучение является очень популярным направлением, так как во многих областях собираются большие объемы данных, которые нужно как-то анализировать. Методы машинного обучения позволяют находить взаимосвязи и закономерности в этих данных. Несколько примеров практических задач, в решении которых может помочь машинное обучение (задачи, на мой взгляд, близкие к интересам экономистов):

- кредитный скоринг,
- маркетинговые исследования рынка,
- прогноз цен на недвижимость.

Прerequisites

Для изучения данного курса студенты должны обладать базовыми знаниями линейной алгебры, математического анализа и теории вероятностей. Также крайне желательно уметь программировать на Python: все домашние задания нужно будет выполнять на Python, однако на необходимом уровне им можно овладеть и в ходе курса. Знание эконометрики упростит понимание материала, но оно абсолютно не обязательно.

Основные темы:

1. Основные задачи машинного обучения: классификация, регрессия, кластеризация.
2. Метод ближайшего соседа.
3. Линейные методы для классификации и регрессии.
4. Метод опорных векторов: линейный и ядровой.
5. Нейронные сети.
6. Деревья решений.
7. Композиции алгоритмов: бэггинг и бустинг.
8. Методы кластеризации.

В ходе курса кроме теоретических основ будут разбираться примеры применения перечисленных алгоритмов на практике, в том числе в рамках практических домашних заданий.

Оценивание

Курс предполагает сдачу 3-4 практических домашних заданий и итоговой письменной контрольной работы. Итоговая оценка на 70% состоит из оценок за домашние задания и на 30% из оценки за контрольную работу. Ни одна из оценок не является блокирующей.