COMPUTATIONAL SOCIAL SCIENCE METHODS IN R

Machine learning in Social Science Studies

Taehee Kim

Summer Semester 2022

WHAT IS MACHINE LEARNING?

Reading

Raschka, Sebastian, and Vahid Mirjalili. Python machine learning. Packt Publishing Ltd, 2017.

WHAT IS MACHINE LEARNING?

- Field of study that gives computers the ability to learn without being explicitly programmed (Arthur Samuel, 1959)
- it learns a function that maps an input to an output based on a training dataset (set of examples of input and output)
- Machine learning is not a single approach but rather a diverse array of techniques: classification, regression, clustering, Bayesian networks, etc…
- spam filter, image detection, self-driving car, AlphaGo….etc

Definition of machine learning algorithm

A machine learning algorithm is an algorithm that is able to learn from data.

> *“ A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at task in T, as measured by P, improves with experience E.” (Mitchell, T. 1997)*

Three elements

- Experience E : train the algorithm (model) by maximizing the performance P on the training set E.
- Task T : task is solved by the model trained by E.
- Performance P : it should increase with E.

Goal

- learning from data
- execute task T based on experience E with optimal performance P

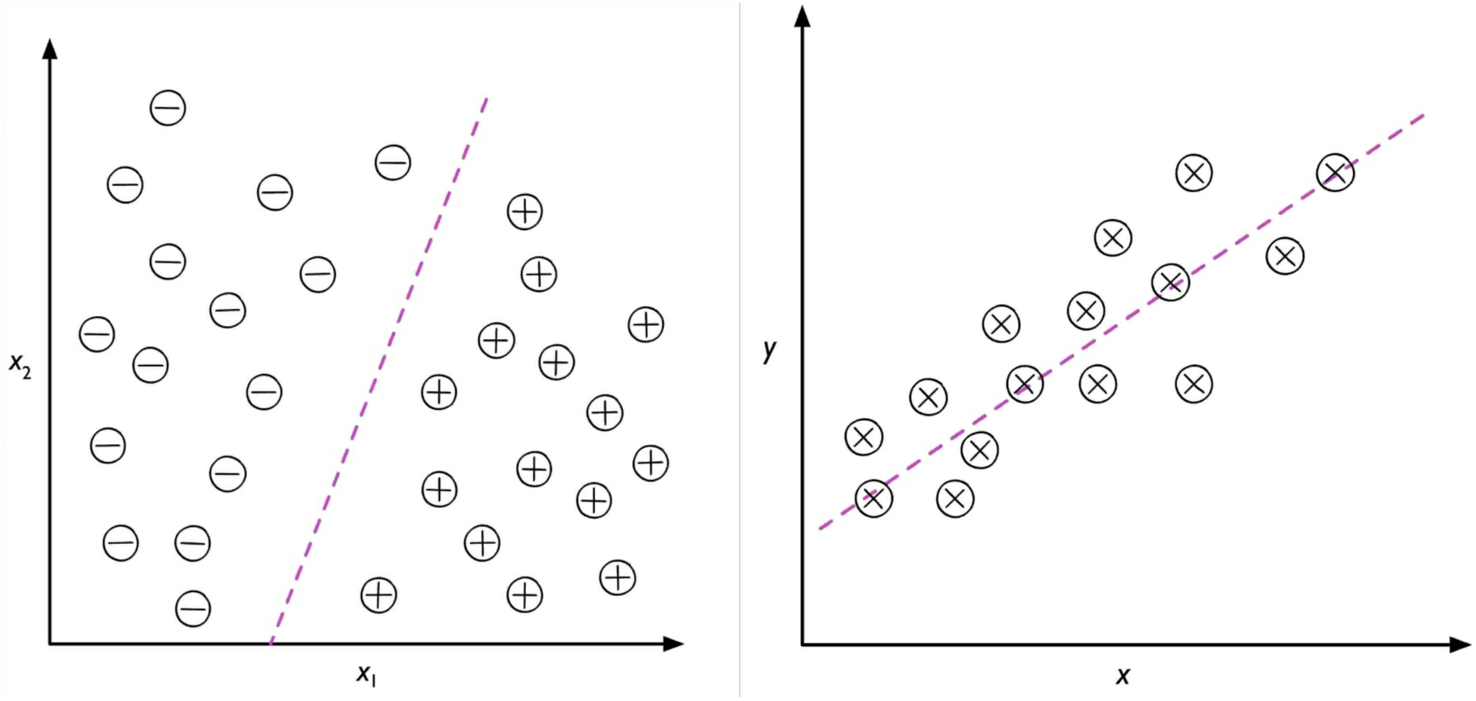
Three types of machine learning

- Supervised learning
- Unsupervised learning
- Reinforcement learning

SUPERVISED LEARNING

- Right answers are given: labeled data
- Direct feedback
- Predict outcome/future
- Two kinds in terms of outputs:
  - Regression: predict continuous valued output
  - Classification: Predict discrete valued output
- Example: linear/logistic regression, SVM, Neural Network etc

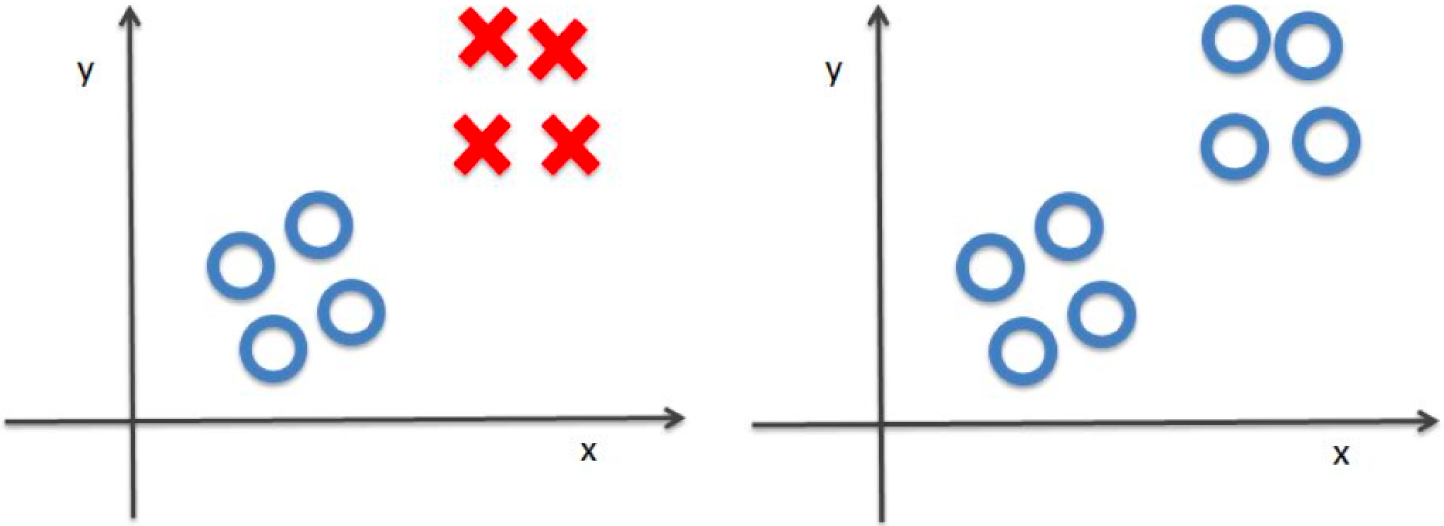
CLASSIFICATION VS. REGRESSION



UNSUPERVISED LEARNING

- We can’t give right answer to all data as the data increase exponentially with the development of technology
- No labeled data
- No feedback
- Find hidden structure in data
- Example: K-means, Gaussian mixture models, PCA etc

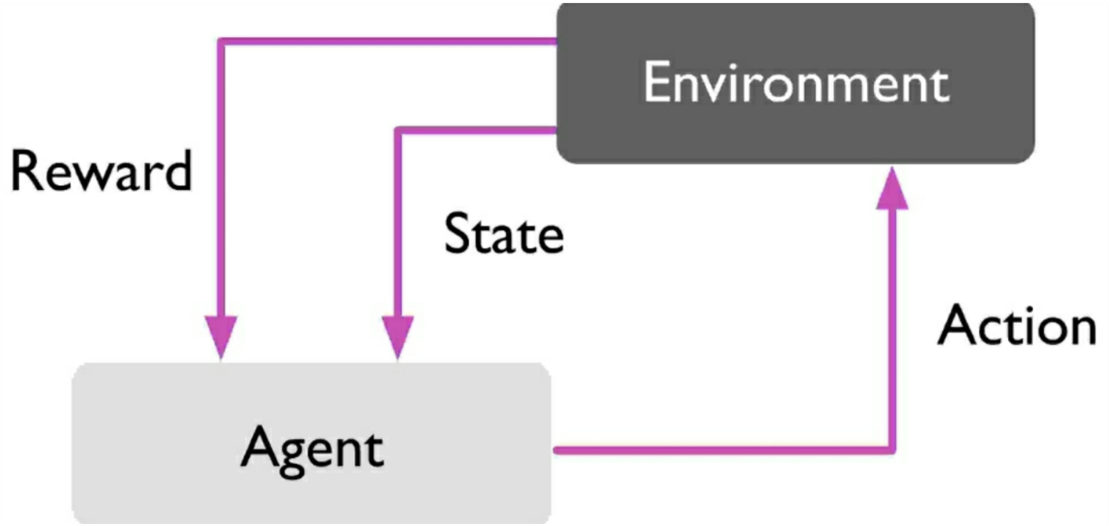
SUPERVISED VS. UNSUPERVISED



classification vs. clustering

REINFORCEMENT LEARNING

- Decision process
- Reward system
- Learn series of actions
- Example: Chess game



MACHINE LEARNING IN SOCIAL SCIENCE STUDIES

Reading and Acknowledgement

Grimmer et al. 2021. Machine Learming for Social Science: An Agnostic Approach. Annual Reviews of Political Science. 24:395-419 link

WHY MACHINE LEARNING?

- Data abundance (large, new form) + computational power
- Machine learning tools are used to extract meaning from data sets both massive and small
- Provide more diverse approach to answer social science questions

How it is used (can be used) in Social Science

- Discovery
- Measurement
- Inference
  - Causal
  - Prediction

> *The abundance of data and resources facilitates the move away from a deductive social science to a more sequential, interactive and ultimately inductive approach to inference. (Grimmer et al. 2021)*

Machine learning focuses **Prediction**

- approxiated by something like R^2
- focuses on $\hat{y}$ (the prediction of the outcome)
- introduce more complex models than traditional linear models common in the social science
- provide better predictive performance

(Traditional) Social science focuses on the relationships of the variables

- Social Scientists focuses on $\hat{\beta}$ (the parameter of the model)
  - the predictions can be a more opaque function of the inputs
  - the models may not easily provide meaningful estimates of uncertainty
  - the estimation routines are more computationally intensive

Deductive model of social science?

- the most common process in the social science
- quantitative tests and variables must be defined a priori (e.g., King et al. 1995)
    - help to reduce false discoveries that can result from researcher discretion.
    - most efficient way to conduct research when data were scarce

Inductive approach?

- the importance of more inductive forms of analysis in qualitative research has been suggested by qualitative scholars
- researchers often discover new directions, questions, and measures within quantitative data
- machine learning can facilitate those discoveries

Measurement

- Machine learning methods are useful for a wide range of measurement strategies
- it can improve the classification of observations into categories
    - it is used to extrapolate the hand coding decisions of coders to other data
    - it greately reduce the cost of using otherwise difficult-to-use data
    - **text, image**, video

Example: Text classification

- Kim and Ogawa. 2022. The Impact of Politicians' Behaviors on Hate Speech Diffusion: Analysis Focusing on Adoption Threshold of Hate Speech on Twitter in Japan. SocArXiv link
- Hate speech text classification
- Classified 200,000 tweet posts into two categories, hate speech or not.

Example: Image classification

- Won et al. 2017. “Protest activity detection and perceived violence estimation from social media images.” link
- collected geotagged tweets’ images associated to major protest events.
- A multi-task convolutional neural network is used to classify images
- predict its visual attributes, perceived violence and exhibited emotions