Lecture 9: Unsupervised Learning

Intro to Data Science for Public Policy, Spring 2016

by Jeff Chen & Dan Hammer, Georgetown University McCourt School of Public Policy

Contents

No, this is not an independent study session. Unsupervised learning techniques such as clustering and principal components analysis helps to identify recognizable patterns when no labels are provided. In sales and recruitment offices, customer segmentation may use current customer data, then use clustering techniques to identify k-number of distinct customer profiles. In resourceful law firms, data scientists may develop topic modeling algorithms to automatically tag and cluster hundreds of thousands of documents for improved search. This session will focus on clustering methodologies that are commonly employed in applied research.

Overview

Three common unsupervised learning problems in public policy:

Example #1

[Example goes here]

Example #2

[Example goes here]

Example #3

[Example goes here]

Unsupervised Learning

No, it's not independent study

[text goes here]

Structure of an unsupervised learning problem

- Target variables
- Input variables
- Objective function and evaluation measures
- model experiment design
- Cross validation versus train/validate/test

Framing Dataset

[text goes here]

Types of unsupervised learning

[text goes here]

Models: statistical assumptions and mechanics, risks/strengths, implementation, sanity checks, non-technical explanation, K-means clustering (K-means), Principal Components Analysis (PCA)/Dimensionality Reduction, Hierarchical clustering (if time permits)

Applications of unsupervised learning

Appropriate uses

Univariate clustering application: k-means

#

Multivariate clustering application: Customer segmentation using Census American Community Survey

#