Using Machine Learning for Exploratory Data Analysis

Joshua M. Lewis
josh@cogsci.ucsd.edu
Department of Cognitive Science
University of California, San Diego

Virginia R. de Sa
desa@cogsci.ucsd.edu
Department of Cognitive Science
University of California, San Diego

Abstract

This tutorial will introduce attendees to fundamental concepts in the clustering and dimensionality reduction fields of unsupervised machine learning. Attendees will learn about the assumptions algorithms make and how those assumptions can cause the algorithms to be more or less suited to particular datasets. Hands-on interaction with machine learning algorithms on real and synthetic data are a central component of this tutorial. Students will use the software platform Divvy (freely available from the Mac App Store or divvy.ucsd.edu) to visualize and analyze data in real time while testing the concepts learned during formal instruction. We encourage attendees to bring their Mac laptops and their own datasets for the hands-on portion of the tutorial, and if possible to email their datasets ahead of time to josh@cogsci.ucsd.edu.

Attendees will leave the tutorial with a much better understanding of basic concepts in unsupervised machine learning. Pragmatically they will understand when to apply, e.g., k-means to a dataset versus single linkage clustering. Attendees will also learn how to integrate Divvy into their existing research workflow so that they can quickly test and compare machine learning algorithms on their data.

Objectives and Scope

This tutorial will introduce attendees to fundamental concepts in the clustering and dimensionality reduction fields of unsupervised machine learning. Attendees will learn about the assumptions algorithms make and how those assumptions can cause the algorithms to be more or less suited to particular datasets. Hands-on interaction with machine learning algorithms on real and synthetic data are a central component of this tutorial. Students will use the software platform Divvy to visualize and analyze data in real time while testing the concepts learned during formal instruction. We will encourage attendees to bring their own datasets for analysis in the hands-on portion of the tutorial.

Attendees will leave the tutorial with a much better understanding of basic concepts in unsupervised machine learning. Pragmatically they will understand when to apply, e.g., k-means to a dataset versus single linkage clustering. Attendees will also learn how to integrate Divvy into their existing research workflow so that they can quickly test and compare machine learning algorithms on their data.

Topics

We will split the tutorial into two sections, a morning section focused on clustering and an afternoon section focused on dimensionality reduction. Both sections will start with a brief (approximately 1.5 hours) formal introduction to mathematical and conceptual underpinnings of the topic, followed by a hands-on lab session applying the concepts learned directly before. The lab sessions will start with synthetic datasets designed to reinforce conceptual lessons, and then move to real datasets provided by ourselves and the attendees.

The clustering section will cover centroid-based methods (such as k-means), hierarchical methods (such as single linkage), spectral clustering, and probabilistic modeling (such as Gaussian mixture models). The dimensionality reduction section will cover linear methods (such as PCA and projection pursuit) and that nonlinear methods (such as Isomap, tSNE, and Kernel PCA).

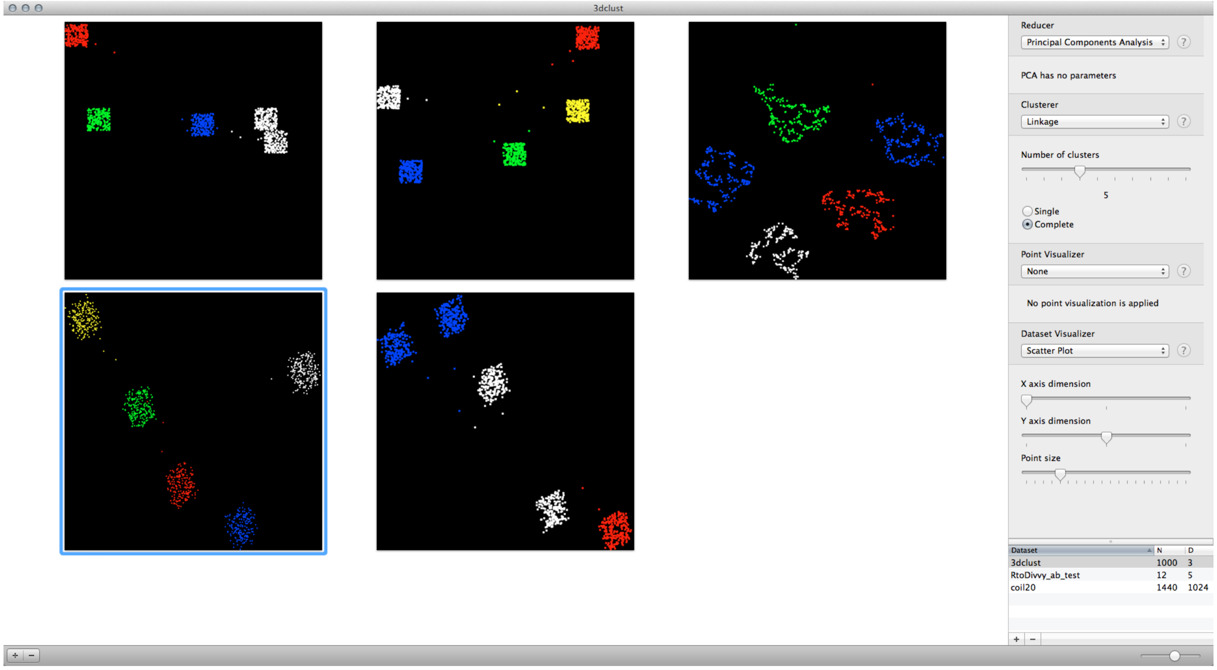
Though we will provide formal mathematical characterizations, our focus will be on conceptual differences between techniques, specifically related to choosing the correct technique based on known structure in a dataset. Additionally, we will emphasize that there is no single best clustering or embedding for any given dataset (in other words, there is no universally agreed upon objective function for clustering and dimensionality reduction). One's own analysis goals can play a significant role in, e.g., determining the number of clusters to search for. Finally, on the topic of evaluation we will cover the visualization and interpretation of algorithmic output as well as formal quality measures such as Silhouette for clusterings and Trustworthiness for embeddings.

We will transact our lab sections in Divvy, a free and open-source software platform for performing unsupervised machine learning (see `http://divvy.ucsd.edu` where there is a video of Divvy in action). Divvy will allow attendees to rapidly cluster, reduce and visualize a wide variety of datasets without having to write any code. Divvy can concurrently visualize several perspectives on a dataset and can switch between datasets with one click, even when algorithms are computing in the background. Divvy integrates well with existing research workflows—it can import data from Matlab and R and it exports data and visualizations in standard formats for further analysis.

Qualifications

Joshua Lewis recently completed his PhD thesis, Anthropocentric Data Analysis, on the topic of reintegrating humans into the data analysis process. He is the lead software architect behind Divvy, and has done several studies on the relationship between human reasoning and machine learning. He is a postdoc in UCSD's Natural Computation Lab under the supervision of Virginia de Sa. He has attended CogSci and presented papers every year starting in 2009. Joshua will lead the tutorial.

Virginia de Sa is an associate professor at UCSD in the Cognitive Science department. She has done extensive re-



Reducer

Principal Components Analysis

PCA has no parameters

Clusterer

Linkage

Number of clusters

5

○ Single

● Complete

Point Visualizer

None

No point visualization is applied

Dataset Visualizer

Scatter Plot

X axis dimension

Y axis dimension

Point size

Dataset	N	D
3dclust	1000	3
RtoDivvy_ab_test	12	5
coil20	1440	1024

Figure 1: The full Divvy UI. Each visualization represents a different view of the same dataset (generated by combining a dimensionality reduction technique, a clustering technique and a dataset visualizer) and users can set the properties of each view using the tools to the right. A list of datasets resides in the bottom right, allowing the user to switch between them at any time, even while results are computing in the background.

search in the fields of machine learning and human perception and has ten years of experience in teaching undergraduate and graduate courses in data analysis and machine learning to people with both weak and strong mathematical backgrounds. She is the PI on NSF Grant #SES-0963071, which funds Divvy's development. Virginia will assist in developing the tutorial curriculum.

For detailed CVs, please see our websites (listed in the Contact Us section below).

Relevance to CogSci

Data analysis is a fundamental part of most scientific endeavors, and the judicious application of machine learning techniques to the analysis process is often quite profitable. Further, in the field of Cognitive Science in particular, a basic understanding of machine learning techniques is valuable for interpreting the work done in computer science-focused subfields such as artificial intelligence and computational neuroscience. Clustering and dimensionality reduction are established methodologies for performing data analysis, and this tutorial presents them from the unique perspective of enabling the human researcher to use them wisely.

Audience

This tutorial will introduce attendees to the area of unsupervised machine learning for data analysis. It does not presuppose any machine learning background and thus will be appropriate for any graduate student or faculty member interested in integrating machine learning techniques into their research process. On the other hand, it will be less valuable to those researchers who already have extensive experience applying machine learning techniques.

Attendee Requirements

Divvy a Mac OS X 10.6/10.7 application, and the hands-on portions of the lab will require a Mac laptop with a 64-bit Intel processor (basically any Mac made in the last three years). Attendees will be able to easily download and install the software at the conference or ahead of time. Those who do not have or do not wish to bring a laptop will be grouped with those who do. Hopefully enough attendees can bring laptops to have groups of about 2 to 4 people per analysis team. Last year's conference was suffused with glowing white Apple logos, so we don't think this will be an onerous requirement. Additionally, attendees can submit datasets to us ahead of time that we can integrate into the instruction as examples of real-life data analysis problems.

Contact Us

Joshua M. Lewis - josh@cogsci.ucsd.edu - http://cogsci.ucsd.edu/~josh - UCSD Cognitive Science - 115 Dufour St, Santa Cruz, CA 95060 - (831)-246-1578

Virginia de Sa - desa@cogsci.ucsd.edu - http://cogsci.ucsd.edu/~desa - UCSD Cognitive Science - 9500 Gilman Dr, La Jolla, CA 92093 - (858)-822-5095