MATH 829 - Spring 2016

Introduction to data mining and analysis

Presentations

Gore 114

May 23, 2016

Time	Speaker(s)	Title
10:00 to 10:25	Kevin Aiton Zach Bailey Kevin Cotter	Portfolio choice with orthogonal multi-arm bandit
10:30 to 10:55	Angela Cuadros	Low Rank Matrix Completion
11:00 to 11:25	Moumita Bhattacharya	Understanding Comorbidity by applying Topic Models to Electronic Medical Records
11:30 to 11:55	Emily Bergman Jake Dynes Amanda Russo	Sentiment Analysis and the Political Twitterverse
12:00 to 1:25	**Lunch break**	
1:30 to 1:55	Mingchang Ding Emma Pollard Lan Zhong Yingxiang Zhou	Scale Invariant Feature Transform (SIFT)
2:00 to 2:25	Benjamin Civiletti Madelyn Houser Jacob Jacavage Amy Janett	It's Trivial: Detecting Condescending Language on Math Stack Exchange
2:30 to 2:55	Ke Jin Shuying Sun Peng Xu	The adversarial multi-armed bandit problem
3:00 to 3:25	Xin Guo	A Probabilistic Language Models based on Neural Networks
3:30 to 3:55	Sergio Matiz Romero	Applications of PCA to Dictionary Learning for Image Classification

Abstracts

10:00–10:25 Kevin Aiton, Zach Bailey, Kevin Cotter, *Portfolio choice with orthogonal multi-arm bandit.*

The classical Markowitz portfolio theory suggests that for a given return, we should choose a portfolio with minimum variance; this naturally leads to a quadratic optimization problem. While this algorithm works well, it is often too conservative. For our presentation we will implement the orthogonal bandit algorithm. This algorithm analyzes the principle components for a group of assets in order to form a portfolio. We will compare this method with the classical Markowitz approach.

10:30–10:55 Angela Cuadros, *Low Rank Matrix Completion.*

In many applications such as system identification in control theory, covariance matrix estimation, machine learning and computer vision one often deals with the problem of recovering missing entries in a data matrix. In general, the problem of filling in these missing values in a matrix is known as: Matrix Completion (Laurent 2001).

The matrix completion problem is ill-conditioned unless the data matrix that wants to be recovered is known to be structured in the sense that it is low-rank or approximately low-rank. In this talk, the matrix completion optimization problem for low rank matrices and algorithms to solve it are introduced. Additionally, the results of solving the Netflix movie challenge problem and image inpainting problems using matrix completion algorithms are presented.

11:00–11:25 Moumita Bhattacharya, *Understanding Comorbidity by applying Topic Models to Electronic Medical Records.*

Close to 80% of Medicare spending in the US is devoted to patients with 4 or more chronic conditions, with costs increasing exponentially as the number of chronic conditions increases. As a result, understanding the presence of multiple diseases in an individual, known as comorbidity, is of growing interest among medical practitioners and researchers. We propose an application of Topic modeling using Latent Dirichlet Allocations (LDA) to identify unique groups of diseases that tend to co-occur. We utilize the information available in Electronic Medical Records (EMRs) containing diagnosed conditions for several thousand patients. Our results establish that groups of diagnosed conditions obtained using topic modeling are statistically significant and can reveal medically relevant commonalities among diseases.

11:30–11:55 Emily Bergman, Jake Dynes, and Amanda Russo, *Sentiment Analysis and the Political Twitterverse.*

In this research we follow the sentiment of Bernie Sanders and Hillary Clinton on Twitter for three primaries during May 2016. We aim to answer a question about the sentiment towards Clinton as she gets closer to taking the democratic nomination. Based on research done for the 2012 GOP primaries, we use the popular Natural Language Toolkit (NLTK) library as well as a very simple analysis of the results to see if, in the last four years, political discourse on Twitter has made any progress.

12:00–1:25 Lunch break.

Talks resume at 1:30 PM.

1:30–1:55 Mingchang Ding, Emma Pollard, Lan Zhong, Yingxiang Zhou, *Scale Invariant Feature Transform (SIFT).*

We realize object recognition in 2D images by using the Scale-invariant feature transform (SIFT) algorithm. The SIFT algorithm can be applied to many tasks such as 3D reconstruction, panorama creation, object recognition, and motion tracking. Implementation of this algorithm consists of four main stages: first, scale-space extrema detection; second, key-point localization; third, orientation assignment; fourth, keypoint description. The keypoint descriptors can then be matched against other images. We can detect many highly distinctive features of images through this algorithm and these features are somewhat robust to image scaling and rotation, especially to change of noise.

2:00–2:25 Benjamin Civiletti, Madelyn Houser, Jacob Jacavage, Amy Janett, *It's Trivial: Detecting Condescending Language on Math Stack Exchange.*

The emergence of the Internet has irrevocably changed the way in which we communicate. Now we can instantaneously "talk" with friends and strangers across the globe. One unintended but harmful consequence of Internet communication is our ability to hide anonymously behind screens with little accountability for what we say. As a result, chat rooms, public forums, and other sites teem with condescending, rude, and hurtful interactions. We investigate Math Stack Exchange, an open forum of mathematical discussion, for condescending language in the answers to hundreds of posted math questions. Using sentiment analysis and support vector machines, we explore models able to distinguish between condescending and non-condescending speech.

2:30–2:55 Ke Jin, Shuying Sun, Peng Xu, *The Adversarial Multi-Armed Bandit Problem.*

In this talk, we will study the model of adversarial multi-armed bandit problem: given K arms, the adversary selects gains for each arm, and the player simultaneously chooses one arm and receives the corresponding gain. For this model, there is a randomized strategy which is called Exp3, for the player such that the expected cumulative gain will be close to the best arm . We will use this model to analyze the simulated slot machines and also apply this model to stock picking among those stocks in the Dow-Jones index . As a nice application, for the repeated unknown game, the Exp3 algorithm will guarantee that the expected gain per round converges to the Nash equilibrium very quickly.

3:00–3:25 Xin Guo, *A Probabilistic Language Models based on Neural Networks.*

In this report, the basics of natural language models such as mathematical representation, n-gram models, perplexity are introduced. Then the very first feed-forward neural network language model proposed by Y. Bengio et.al.in 2003 is demonstrated and implemented. The experimental results show that the feed-forward neural network language model is suitable of processing natural language applications.

3:30–3:55 Sergio Matiz Romero, *Applications of PCA to Dictionary Learning for Image Classification.*

Recently proposed methods for dictionary learning can also be used for classification tasks. This project focuses on techniques based on sparse coding, where classification is performed on the sparse representations of testing vectors. Since vectors are often high-dimensional, PCA offers an effective way to reduce the dimensionality of the optimization problems required for dictionary learning. In this project the use of PCA for facial recognition tasks is explored and reported. Experiments conducted on two face recognition databases, the Extended YaleB database and the AR database, demonstrate the computation time savings obtained through PCA.