Machine Learning for Data Science (CS4786) Lecture 1

Tu-Th 10:10 to 11:25 AM
Phillips Hall 101

Instructor : Karthik Sridharan

The Awesome TA’s

1. Esin Durmus
2. Vlad Niculae
3. Jonathan Simon
4. Ashudeep Singh
5. Yu Sun [TA consultant]
6. Yechuan(Jeff) Tian
7. Felix Wu

Course Information

- Course webpage is the official source of information: `http://www.cs.cornell.edu/Courses/cs4786/2016sp`
- Join Piazza: `https://piazza.com/class/ijxdhmmko1h130`
- TA office hours will start from next week
- While the course is not coding intensive, you will need to do some light coding.

Course Information

- Assignments worth 60% of the grades
- Two competitions (worth 40% of the grade)
- TA office hours will start from next week
- Course is not coding intensive, light coding needed though (language your choice)

Assignments

- Diagnostic assignment 0 is out: for our calibration.
  - 3% of assignment grades allotted only to hand in A0
    (we wont be giving grades for solutions)
  - Students who want to take course for credit need to submit this, only then you will be added to CMS.
  - Hand in your assignments beginning of class on 4th Feb.
  - **Has to be done individually**

Assignments

- Diagnostic assignment 0 is out: for our calibration.
  - 3% of assignment grades allotted only to hand in A0 (we wont be giving grades for solutions)
  - Students who want to take course for credit need to submit this, only then you will be added to CMS.
  - Hand in your assignments beginning of class on 4th Feb.
  - **Has to be done individually**
- Three assignments A1, A2 and A3
  - Can be done in groups of size at most 4.
  - Only one write up/submission per group

Assignments

- Diagnostic assignment 0 is out: for our calibration.
  - 3% of assignment grades allotted only to hand in A0 (we wont be giving grades for solutions)
  - Students who want to take course for credit need to submit this, only then you will be added to CMS.
  - Hand in your assignments beginning of class on 4th Feb.
  - **Has to be done individually**

- Three assignments A1, A2 and A3
  - Can be done in groups of size at most 4.
  - Only one write up/submission per group

- Diagnostic assignment P1 (sometime mid semester)
  - **Has to be done individually**
  - Worth 10% of class grades

Competitions

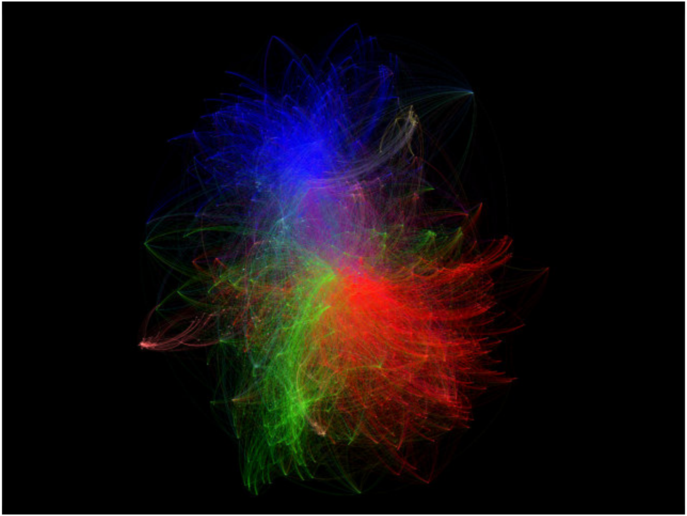
- 2 competition/challenges,
  - Clustering/data visualization challenge
  - Prediction challenge with focus on feature extraction/selection
- Will be hosted on “In class Kaggle”!
- Grades for project focus more on thought process (demonstrated through your reports)
- Kaggle scores only factor in for part of the grade.
- Groups of size at most 4.

Lets get started …

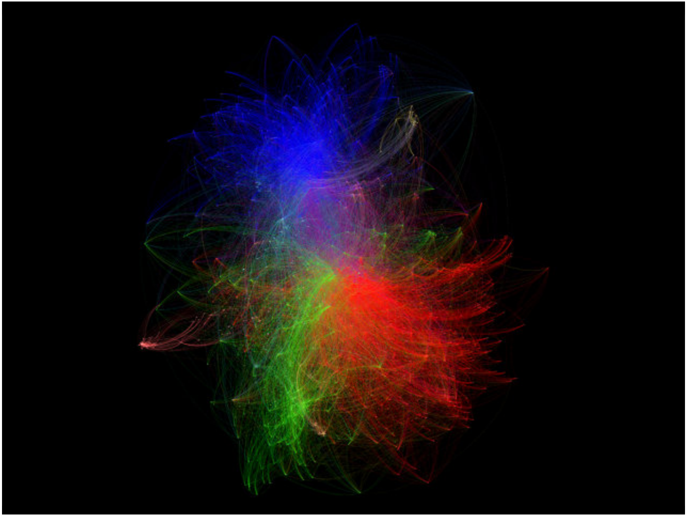
Data Deluge

- Each time you use your credit card: who purchased what, where and when
- Netflix, Hulu, smart TV: what do different groups of people like to watch
- Social networks like Facebook, Twitter, . . . : who is friends with who, what do these people post or tweet about
- Millions of photos and videos, many tagged
- Wikipedia, all the news websites: pretty much most of human knowledge

Guess?



Social Network of Marvel Comic Characters!



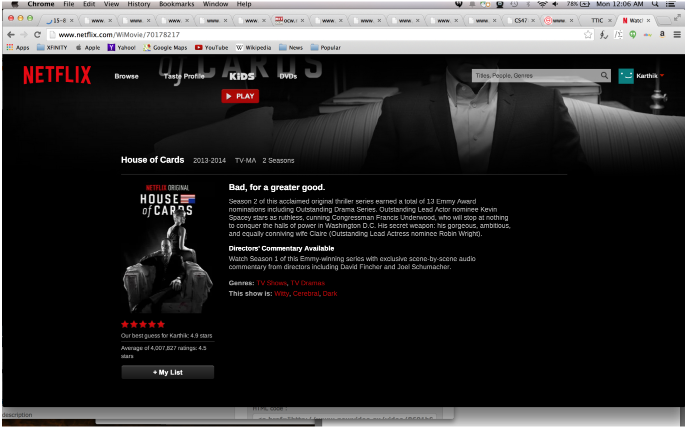
What can we learn from all this data?

What is Machine Learning?

Use data to automatically learn to perform tasks better.

WHERE IS IT USED ?

Movie Rating Prediction


Chrome File Edit View History Bookmarks Window Help
Mon 12:06 AM
www.netflix.com/WiMovie/70178217
Apps XFINITY Apple Yahoo! Google Maps YouTube Wikipedia News Popular
NETFLIX
Browse
Taste Profile
KIDS
DVDs
Titles, People, Genres
Karthik
PLAY

House of Cards 2013-2014 TV-MA 2 Seasons

Bad, for a greater good.

Season 2 of this acclaimed original thriller series earned a total of 13 Emmy Award nominations including Outstanding Drama Series. Outstanding Lead Actor nominee Kevin Spacey stars as ruthless, cunning Congressman Francis Underwood, who will stop at nothing to conquer the halls of power in Washington D.C. His secret weapon: his gorgeous, ambitious, and equally conniving wife Claire (Outstanding Lead Actress nominee Robin Wright).

Directors' Commentary Available

Watch Season 1 of this Emmy-winning series with exclusive scene-by-scene audio commentary from directors including David Fincher and Joel Schumacher.

Genres: TV Shows, TV Dramas

This show is: Witty, Cerebral, Dark

★★★★★

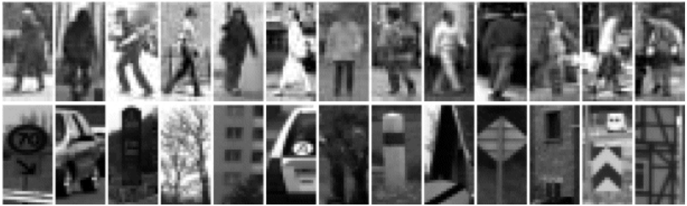
Our best guess for Karthik: 4.9 stars

Average of 4,007,827 ratings: 4.5 stars

+ My List

Where is it used ?

Pedestrian Detection

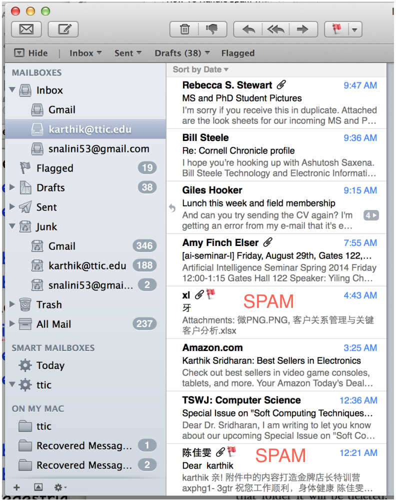


Where is it used ?

Market Predictions

WHERE IS IT USED ?

Spam Classification



Rebecca S. Stewart 9:47 AM
MS and PhD Student Pictures
I'm sorry if you receive this in duplicate. Attached are the look sheets for our incoming MS and P…

Bill Steele 9:36 AM
Re: Cornell Chronicle profile
I hope you're hooking up with Ashutosh Saxena. Bill Steele Technology and Electronic Informati…

Giles Hooker 9:15 AM
Lunch this week and field membership
And can you try sending the CV again? I'm getting an error from my e-mail that it's e… 4

Amy Finch Elser 7:55 AM
[ai-seminar-l] Friday, August 29th, Gates 122,…
Artificial Intelligence Seminar Spring 2014 Friday 12:00-1:15 Gates Hall 122 Speaker: Yiling Ch…

xl SPAM 4:43 AM
牙
Attachments: 微PNG.PNG, 客户关系管理与关键客户分析.xlsx

Amazon.com 3:25 AM
Karthik Sridharan: Best Sellers in Electronics
Check out best sellers in video game consoles, tablets, and more. Your Amazon Today's Deal…

TSWJ: Computer Science 12:36 AM
Special Issue on "Soft Computing Techniques…
Dear Dr. Sridharan, I am writing to let you know about our upcoming Special Issue on "Soft Co…

陈佳雯 SPAM 12:21 AM
Dear karthik
karthik 亲! 附件中的内容打造金牌店长特训营 axphg1- 3gtr 祝您工作顺利，身体健康 陈佳雯…

More Applications

- Each time you use your search engine
- Autocomplete: Blame machine learning for bad spellings
- Biometrics: reason you shouldn't smile
- Recommendation systems: what you may like to buy based on what your friends and their friends buy
- Computer vision: self driving cars, automatically tagging photos
- Topic modeling: Automatically categorizing documents/emails by topics or music by genre
- …

Topics we will cover

1. Dimensionality Reduction:
   Principal Component Analysis (PCA), Canonical Component Analysis (CCA), Random projections, Compressed Sensing (CS), …

2. Clustering and Mixture models:
   k-means clustering, gaussian mixture models, single-link clustering, spectral clustering, …

3. Probabilistic Modeling & Graphical Models:
   Probabilistic modeling, MLE Vs MAP Vs Bayesian approaches, inference and learning in graphical models, Latent Dirichlet Allocation (LDA), Hidden Markov Models (HMM), …

Unsupervised Learning

Given (unlabeled) data, find useful information, pattern or structure

- Dimensionality reduction/compression : compress data set by removing redundancy and retaining only useful information
- Clustering: Find meaningful groupings in data
- Topic modeling: discover topics/groups with which we can tag data points

Dimensionality Reduction

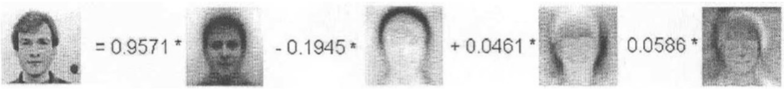
- You are provided with n data points each in $\mathbb{R}^d$

- Goal: Compress data into n, points in $\mathbb{R}^K$ where $K << d$
  - Retain as much information about the original data set
  - Retain desired properties of the original data set

- Eg. PCA, compressed sensing, …

Principal Component Analysis (PCA)

Eigen Face:



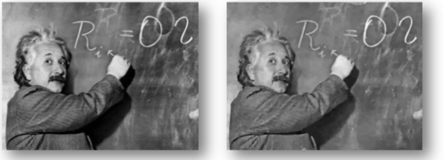
- Write down each data point as a linear combination of small number of basis vectors

- Data specific compression scheme

- One of the early successes: in face recognition: classification based on nearest neighbor in the reduced dimension space

Compressed Sensing

From Compressive Sensing Camera

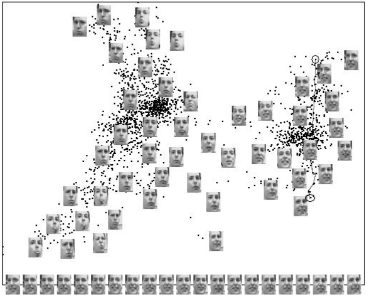


Original Target InView SWIR Reproduction

- Can we compress directly while receiving the input?
- We now have cameras that directly sense/record compressed information …and very fast!
- Time spent only for reconstructing the compressed information
- Especially useful for capturing high resolution MRI's

Data Visualization

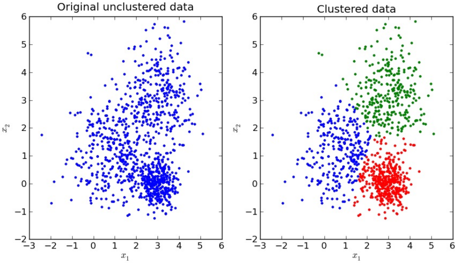
2D projection



- Help visualize data (in relation to each other)
- Preserve relative distances among data-points (at least close by ones)

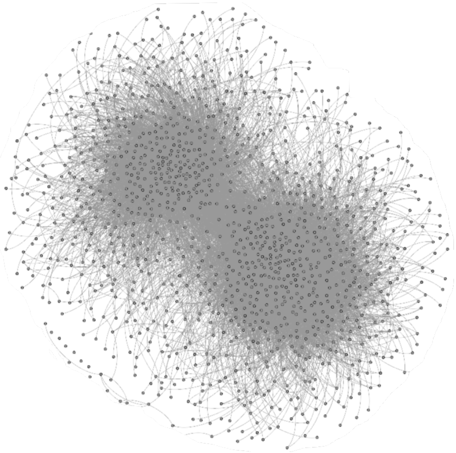
CLUSTERING

K-means clustering

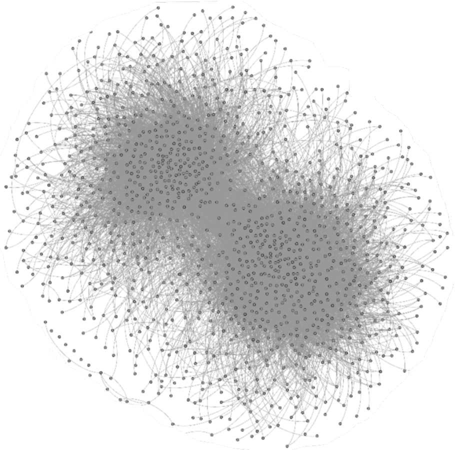


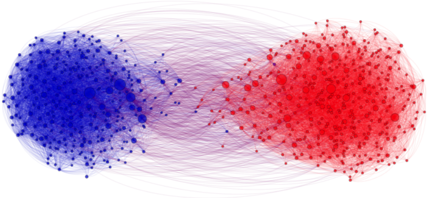
- Given just the data points group them in natural clusters
- Roughly speaking
  - Points within a cluster must be close to each other
  - Points between clusters must be separated
- Helps bin data points, but generally hard to do

Tell me who your friends are . . .



TELL ME WHO YOUR FRIENDS ARE . . .

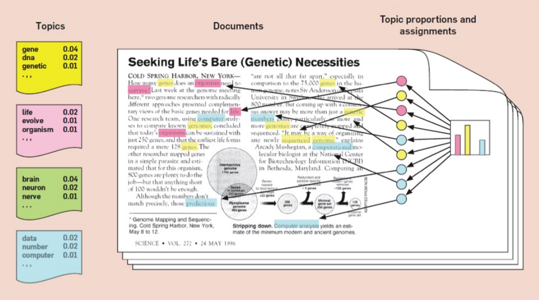




- Cluster nodes in a graph.
- Analysis of social network data.

Topic Modelling

Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.



- Probabilistic generative model for documents
- Each document has a fixed distribution over topics, each topic is has a fixed distribution over words belonging to it
- Unlike clustering, groups are non-exclusive

Supervised Learning

- Training data comes as input output pairs (x, y)
- Based on this data we learn a mapping from input to output space
- Goal: Given new input instance x, predict outcome y accurately based on given training data
- Classification, regression

What we won't cover

- Feature extraction is a problem/domain specific art, we won't cover this in class
- We won't cover optimization methods for machine learning
- Implementation tricks and details won't be covered
- There are literally thousands of methods, we will only cover a few!

What you can take home

- How to think about a learning problem and formulate it
- Well known methods and how and why they work
- Hopefully we can give you an intuition on choice of methods/approach to try out on a given problem

Dimensionality Reduction

Given data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ compress the data points in to low dimensional representation $\mathbf{y}_1, \ldots, \mathbf{y}_n \in \mathbb{R}^K$ where $K << d$

Why dimensionality reduction?

- For computational ease
  - As input to supervised learning algorithm
  - Before clustering to remove redundant information and noise
- Data visualization
- Data compression
- Noise reduction

Dimensionality Reduction

Desired properties:

1. Original data can be (approximately) reconstructed
2. Preserve distances between data points
3. “Relevant” information is preserved
4. Redundant information is removed
5. Models our prior knowledge about real world

Based on the choice of desired property and formalism we get different methods

Sneak Peek

- Linear projections
- Principal component analysis