

Large-Scale Machine Learning

Sanjiv Kumar, Google Research, NY
EECS-6898, Columbia University - Fall, 2010

Brief Intro

About Me

- PhD (Robotics Institute, School of Computer Science, CMU).
 - Discriminative Graphical Models for image understanding
- Research Scientist at Google NY – 15th St & 8th Ave
- Large scale machine learning and computer vision
- Very large scale matrix decompositions and nearest neighbor search

About You

- Background in ML
- Machine learning application area (e.g., vision/speech/biology/finance)

Machine Learning

Given a few examples (training data), make a machine learn how to predict on new samples, or discover patterns in data

Statistics + Optimization + Computer Science

Past

- Significant advances in last several decades but focus on relatively small-scale applications – academic publications on UCI datasets!
- Large-scale kernel methods book (1998) has largest experiment with ~50K training points!

Current

- Landscape of machine learning applications has changed dramatically in last decade – Web, Finance, Biology,...
- “**Throw more machines at it**” is not always a solution
- Need to revisit traditional algorithms and models
- **A strong model with approximate solution Vs a weaker model with exact solution?**

Machine Learning

Traditional CS view: Polynomial time algorithm, Wow!

Large-scale learning: Sometimes even $O(n)$ is bad!

Need a shift in the way we think about developing or applying machine learning techniques

Simple example: Matrix multiplication

$$\begin{array}{c} n \\ \square \\ n \end{array} \times \begin{array}{c} \square \\ \square \\ \square \end{array} = \begin{array}{c} \square \\ \square \\ \square \end{array} \quad O(n^3)!$$

Our Goal: Add another dimension in the study of machine learning techniques: scalability

What is "large scale" ?

Many definitions

- If data cannot fit in RAM
- **Algorithm dependent** – If a technique cannot be applied on a single machine in 'reasonable' time (e.g., for linear classifiers $O(100K)$ is medium size but for kernel classifiers, it is large)

The size of datasets can be so huge that even linear methods may not be enough

- Finding nearest neighbors from a database of $O(B+)$ items (for retrieval, kNN classification, density estimation, graph construction ...), is very expensive

The desired models (i.e., classes) and parameters may be huge

- Image categorization with $O(10M)$ classes

Data may be arriving on-line as streams at a rate higher than the learning rate of an algorithm

Large-Scale Data

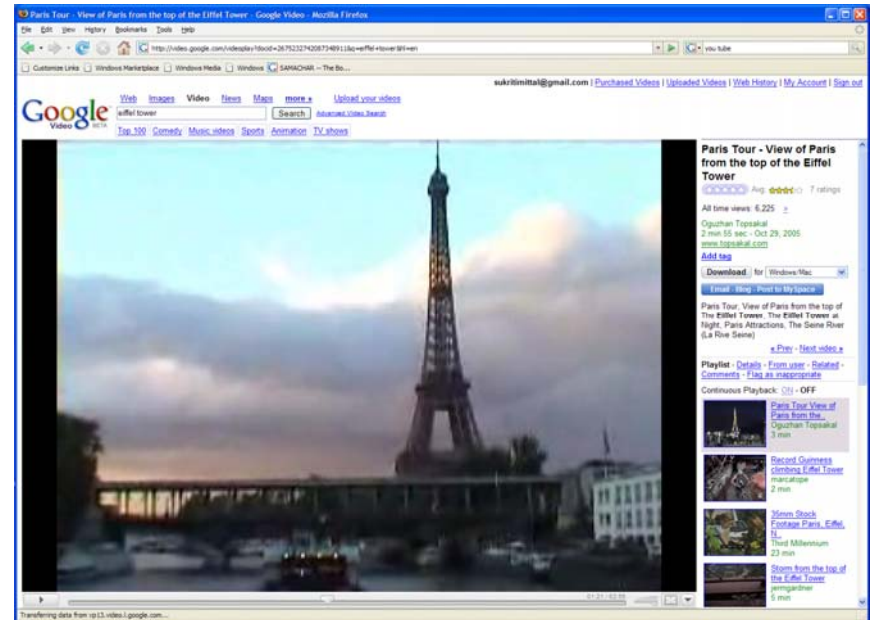
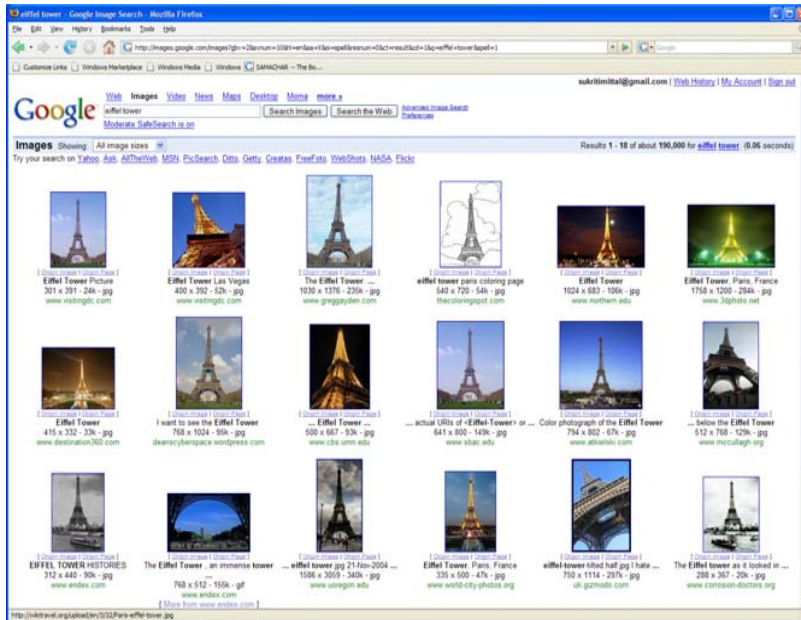
Web applications have practically evolved as the largest consumers of machine learning

- Documents and blogs
- User reviews
- Social networks
- Images and videos
- Imaging every street / Creating Maps
- Scanned books

Other Sources

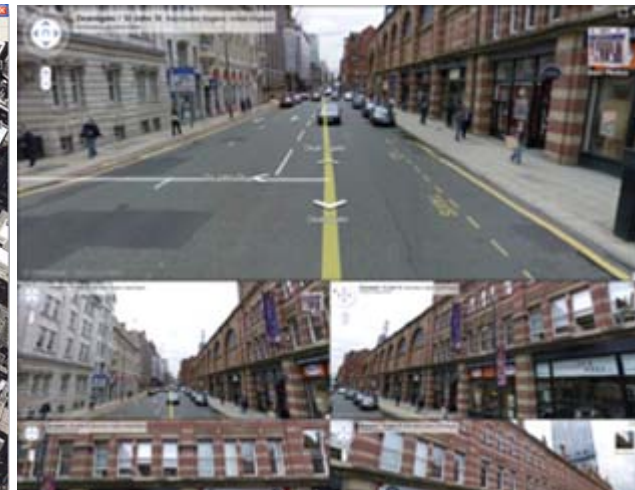
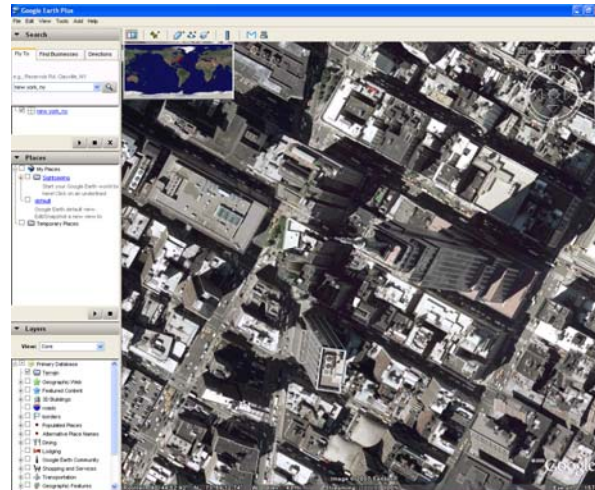
- Biology (Genes data)
- Finance (fraud-detection, credit approval, time-series data)
- Astronomy
- Smart digital homes (continuous health records of a person)

Image/Video Search



- $O(B^+)$ images on the web and photo sharing sites (picasa, flicker...)
- Large scale indexing and retrieval (meta-data/content)
- Content-based automatic annotation: $O(10M)$ classes ?
- Object/event/location recognition
- YouTube - more than 20 hours of video uploaded every minute
 - Even real-time algorithms are not sufficient!

Earth / Maps / StreetView



Make a single gigantic seamless image of 149 million Sq Km land area !

Map all the streets in the world !

- Understanding roads, buildings, etc. in the scene
- Estimate green cover in a city
- Face/Car detection in street images
- Identify buildings, businesses,...

Digital Library/Book Search



Want to scan tens of millions of books !

- Document layout understanding
- Text-image-table separation
- Optical Character Recognition in hundreds of languages
- Translation

Shopping/Product-Search

amazon.com Hello, [Sign in](#) to get personalized recommendations. New customer? [Start here](#).
Your Amazon.com | [Today's Deals](#) | [Gifts & Wish Lists](#) | [Gift Cards](#)

Shop All Departments Search Watches

Watches What's New Men's Watches Women's Watches



Click for larger image and other views



[View and share related images](#)

Citizen Men's AT0200-05E Eco-Drive Chronograph Watch
by [Citizen](#)
★★★★☆ (161 customer reviews)

List Price: ~~\$215.00~~

Price: **\$129.00** & this item ships for **FREE with S**
You Save: **\$86.00 (40%)**

In Stock.

Ships from and sold by **Amazon.com**. Gift-wrap available.

Want it delivered Wednesday, September 8? Order it in **Shipping** at checkout. [Details](#)

6 new from \$129.00

Gift with Purchase

For a limited time, get a free [sizing tool](#) (a \$15.00 value) with purchase. Add both a [sizing tool](#) and a [qualifying watch](#) to your cart. A [qualifying](#) sizing tool will be applied at checkout. Items may ship separately.

[See more product promotions](#)

Customers Who Viewed This Item Also Viewed



[Citizen Men's BM8180-03E Eco-Drive Canvas Strap Watch](#)

★★★★☆ (247)

\$81.00



[Citizen Men's AT0870-37A Eco-Drive Exclusive Chronograph Canvas Strap Watch](#)

★★★★☆ (5)



[Citizen Men's BM8475-00X Eco-Drive Military Black Plated Steel Watch](#)

★★★★☆ (10)

\$123.75

- $O(M)$ products, $O(B)$ reviews
- Sentiment Analysis
- Product recommendation for $O(100M)$ users ...

Sequence Data

Google translate

Machine Translation

From: English - detected To: Chinese (Traditional) Translate

English to Chinese (Traditional) translation

如何解決大規模學習問題

Listen Read phonetically

Rúhé jiějué dà guīmó xuéxí wèntí

How to solve large scale learning problems

Listen



Speech Recognition



Stock Prediction

Sequence Data



From: English - detected To: Chinese (Traditional) Translate

English to Chinese (Traditional) translation

如何解決大規模學習問題

How to solve large scale learning problems

Read phonetically
dà guīmó xuéxí wèntí

Listen



- Hundreds of languages
- Millions of words
- Huge number of models

Graphical Data - Social Networks

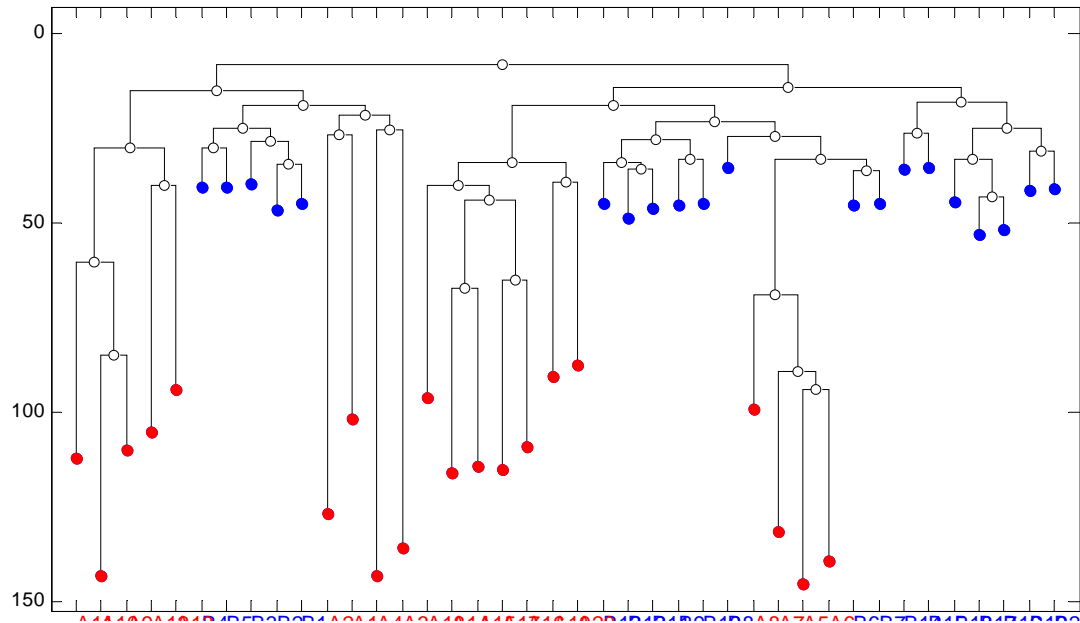
facebook

Facebook helps you connect and share with the people in your life.



- About 500M users
- Information propagation
- Recommendation systems
- Effect of hubs

Human Cell Lineage Tree



- History of body cells since conception
- Leaves are current body cells
- $O(100T)$ leaves and $O(100T)$ branches for one person
- About 100K times bigger than Human Genome

[Shapiro E.]

Streaming Data

The screenshot shows the Twitter homepage. At the top, there is a search bar with the text "Search for a keyword or phrase..." and a "Search" button. Below the search bar, the text "Discover what's happening right now, anywhere in the world" is displayed. A horizontal bar contains trending topics: "Rock2", "Kissed Henry", "TRENDING TOPICS", "Year Without Rain", "College Gameday", "Silvio Santos morreu", and "Fathers". On the left side, there is a "See who's here" section with a grid of profile pictures and a list of names including "HBR", "malaria", "Coupon Moms", and "BieberTeamUSA". The main content area is titled "Top Tweets" and features three tweets. The first tweet is from "FoxNewsSunday" about Sen McCain. The second is from "Radio1Playlist" about McFly. The third is from "karunchandhok" about Tomizawa. The fourth tweet is from "BieberTeamUSA" with the text "#happybirthdaygoogle thanks for always helping".

The screenshot shows the Blogger homepage. At the top right, there is a language dropdown menu set to "English". Below it, there is a sign-in section titled "Sign in to use Blogger with your Google Account". It includes fields for "Username (Email): sanjivk@google.com" and "Password: (?)", a "SIGN IN" button, and a link "Use a different account". The main content area is titled "Create a blog. It's free." and features three sections: "Beautiful templates. Customize your layout, fonts, colors and more..." with a link to "Try the template designer"; "Your blog. Share your thoughts, photos, and more with your friends and the world."; and "Easy to use. It's easy to post text, photos, and videos from the web or your mobile phone." On the right side, there is a "CREATE A BLOG" button and a section titled "Learn more:" with links to "Take a quick tour", "Watch a video tutorial", "Discover more features", and "Read Blogger Buzz". Below this, there is a "Blogs of Note" section with a link to "Brideet Farmer".

- News/Blogs/Social Networks
- Data arrives at a rate faster than we can learn from it
- Data distribution is usually not fixed !

Massive Data - Blessing or Liability?

Blessing

- Traditional theoretical analysis for many algorithms in statistics and machine learning suggest better learning as the training size increases (e.g., histogram is a very good approximation of true data density)
- Even simple techniques such as kNN can yield optimal results
- Stronger contextual features can be used without worrying about sparsity (e.g., n -grams with bigger n for sequence data)

Liability

- How to handle huge amounts of data for training?
- After a certain limit, does more data matter?
- Much of the data is unlabeled?
- Usually very noisy data (especially from the Web, e.g., image search)

Things to worry about

Training time

- Usually offline but incremental learning for fast arriving data

Memory requirement

- 10 M books * 300 pages * 10 MB/page = 30,000 TB
- Loading time may overwhelm training
- Need to be creative about data storage – lossless compression

Test time

- Yes, that can depend on the training data size! (kernel methods)

Trivial Solutions

- Subsample the data, run standard methods **Good baseline !**
 - accuracy may suffer, does not address large-scale modeling issues (e.g., many classes)
- Simply run through multiple machines
 - Map-reduce style or message-passing style but algorithms may not parallelizable

Hardware based solutions - MapReduce

Distributed computing architecture on a cluster of machines

- Data is chunked and passed to multiple machines

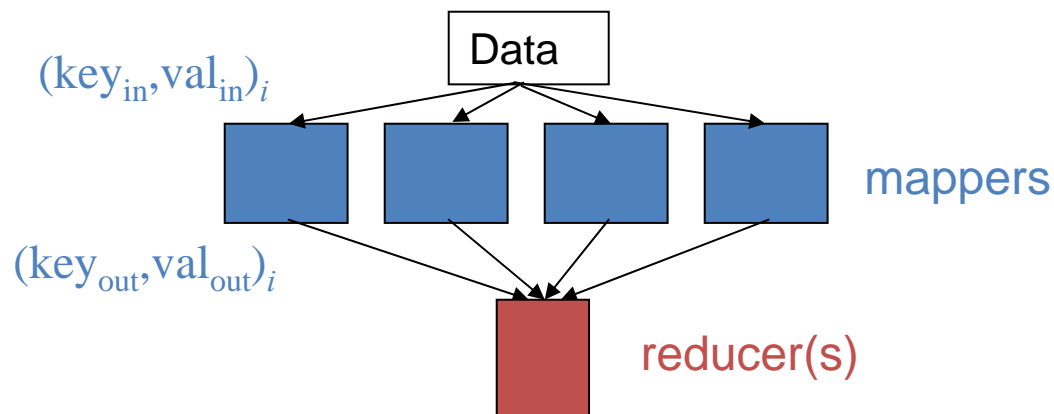
Advantages

- Massively parallel \rightarrow thousands of nodes with generic hardware
- Fault-tolerant

Average color histogram from a billion images

$(key_{in}, val_{in})_i = (\text{image-id}, \text{pixel-val})$

$(key_{out}, val_{out})_i = ('0', \text{3D-hist})$



Laws of diminishing returns

- High network costs
- Probability of at least one machine failing becomes high

Hardware based solutions - GPUs

- Based on CUDA parallel computing architecture from Nvidia
- Emphasis on executing many concurrent threads slowly instead of one thread fast as in CPUs

Advantages

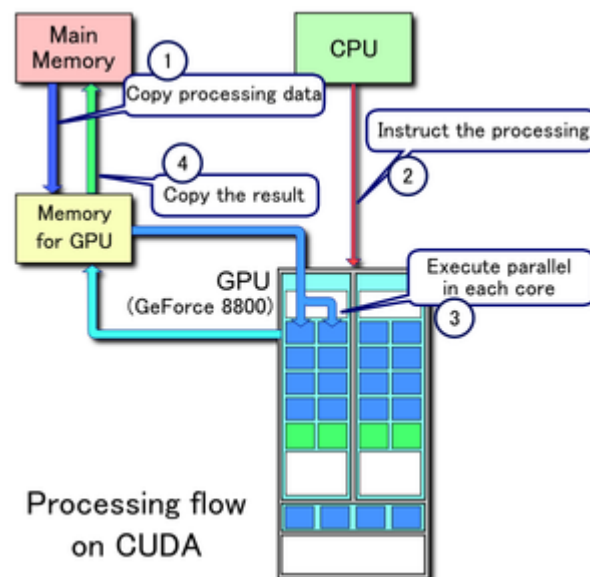
- Massively parallel
- Hundreds of cores, millions of threads
- High throughput

Limitations

- May not be applicable for all tasks
- Generic hardware (CPUs) closing the gap

Applications

- Used successfully for many ML methods, e.g., Deep Networks (both Neural as well as Belief).
- Training cut down from weeks to hours



Data labeling Tools

Mechanical Turk is a marketplace for work.

We give businesses and developers access to an on-demand, scalable workforce. Workers select from thousands of tasks and work whenever it's convenient.

50,952 HITs available. [View them now.](#)

Make Money by working on HITs

HITs - *Human Intelligence Tasks* - are individual tasks that you work on. [Find HITs now.](#)

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work



or [learn more about being a Worker](#)

Get Results from Mechanical Turk Workers

Ask workers to complete HITs - *Human Intelligence Tasks* - and get results using Mechanical Turk. [Register Now](#)

As a Mechanical Turk Requester you:

- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results



Inexpensive, fast, can control quality!

Major Learning Paradigms

Supervised Learning

- Regression (to predict a continuous output)
- Classification (to predict a class or category)
- Ranking (to predict rank ordering)

Unsupervised Learning

- Clustering
- Density Estimation
- Dimensionality Reduction
- Reconstruction - Sparse Coding

Semi-supervised Learning

- Graph propagation
- Manifold regularization

Next: Explore popular techniques from these paradigms to check their scalability and what tools are needed to make them scalable.

Regression



The image shows a screenshot of a YouTube video player. At the top, the YouTube logo is on the left, followed by a search bar and links for 'Search', 'Browse', and 'Upload'. Below this, the video title 'bear desert island' is displayed, along with the channel name 'gogovid', a video count of '73 videos', and a 'Subscribe' button. The video player itself shows a 3D animated polar bear looking to the right against a yellow background with a bokeh effect. The title 'THE DESERT ISLAND' is written in a stylized, pink, outlined font. Below the video frame is a control bar with play/pause, volume, and progress indicators (0:00 / 3:00), along with resolution (360p) and other settings icons. At the bottom, the channel name 'gogovid' and upload date 'April 03, 2007' are shown on the left, and the view count '2,373,878 views' is shown on the right.

Expected age of the most likely viewer of this video?

Regression - Linear Regression

Given: A labeled training set, $\{x_i, y_i\}_{i=1\dots n}$ $x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$
 $n \sim O(100M), d \sim O(100K)$

Goal: Learn a predictive function $f(x; w) = w^T x + w_0$

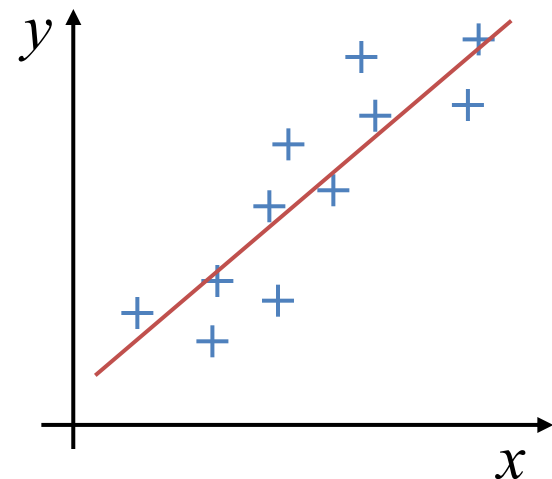
Absorbing w_0 in w and augmenting x_i 's with extra 1

$$L(w) = \sum_i \underbrace{(w^T x_i - y_i)^2}_{\text{squared loss}} + \underbrace{\lambda w^T w}_{\text{regularizer}}$$

$y_i \sim N(w^T x_i, \sigma^2 I)$ $w \sim N(0, \lambda^{-1} I)$

$$= (y - X^T w)^T (y - X^T w) + \lambda w^T w$$

$$\hat{w} = (XX^T + \lambda I)^{-1} Xy$$



Regression - Linear Regression

Given: A labeled training set, $\{x_i, y_i\}_{i=1\dots n}$ $x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$

$$n \sim O(100M), d \sim O(100K)$$

Goal: Learn a predictive function $f(x; w) = w^T x + w_0$

$$\hat{w} = (XX^T + \lambda I)^{-1} Xy$$

$$O(nd^2)$$

$$O(d^3)$$

$$n \sim O(100M), d \sim O(100K)$$

Matrix multiplication is slower than inversion!

Matrix inversion is intractable!

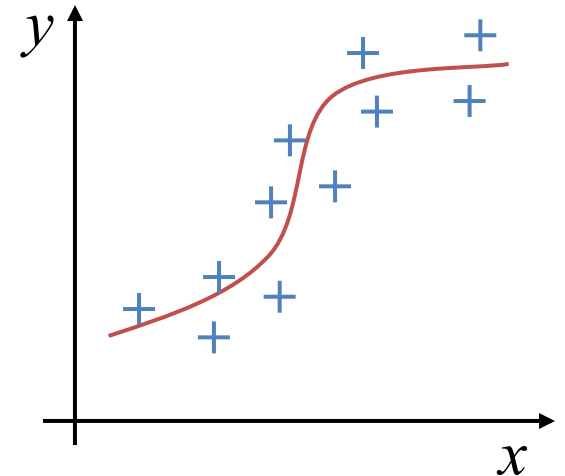
First-order linear solvers $Aw = b$?

But this is just linear !

Regression - Kernel Ridge Regression

Given: A labeled training set, $\{x_i, y_i\}_{i=1\dots n}$ $x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$

Goal: Learn a predictive function $f(x; \alpha) = \sum_{i=1}^n \alpha_i k(x, x_i)$ Ignore bias for now



Regression - Kernel Ridge Regression

Given: A labeled training set, $\{x_i, y_i\}_{i=1\dots n}$ $x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$

Goal: Learn a predictive function $f(x; \alpha) = \sum_{i=1}^n \alpha_i k(x, x_i)$

$k(x, y)$ is a measure of similarity between any two points. For mercer kernels:

$$k(x, y) = \Phi(x) \cdot \Phi(y)$$

Gaussian kernel: $k(x, y) = \exp(-\|x - y\|^2 / \sigma^2)$

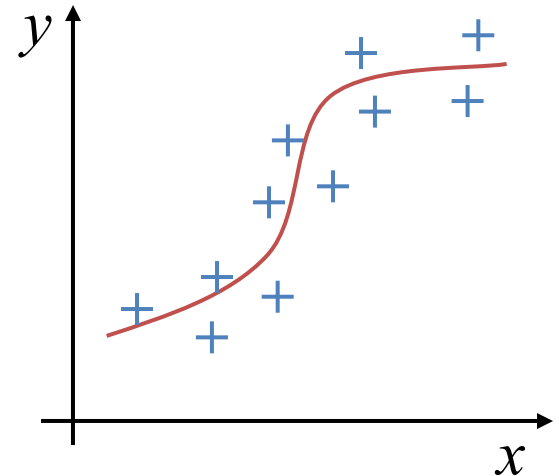
polynomial kernel: $k(x, y) = (x \cdot y + b)^{d_0}$

$$\hat{\alpha} = (K + \lambda I)^{-1} y$$

$$O(n^2 d)$$

$$O(n^3)$$

Number of parameters
same as number of points!



$$n \sim O(100M), d \sim O(100K)$$

K ~ 40,000 TB!

Building K and its inversion is intractable!

Low-rank approximations

Regression - Kernel Ridge Regression

Given: A labeled training set, $\{x_i, y_i\}_{i=1\dots n}$ $x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$

Goal: Learn a predictive function $f(x; \alpha) = \sum_{i=1}^n \alpha_i k(x, x_i)$

Training

$$\hat{\alpha} = (K + \lambda I)^{-1} y$$

Number of parameters
same as number of points!

$O(n^3)$

Optimal choice of hyperparameters (λ)?

Testing

$$f(x) = \sum_{i=1}^n \alpha_i k(x, x_i)$$

Grows linearly with n

Too slow for most practical purposes

Need to induce sparsity in α - L_1 prior **Sparse methods**

Optimization not as clean as for L_2 but not very difficult

Can we approximate the kernel problem with a linear one?

Classification

ASYLUM CAMP ABIDJAN
From: Markson Juliet/Kamara
ADDRESS: MOTHER THREASA CAMP ABIDJAN COTE D'IVOIRE
EMAIL ADDRESS: (kamara35@latinmail.com)

Dear Sir,

I am markson Kamara the only son of late former Director of finance, Chief Vincent R. Kamara of Sierra-Leone diamond and mining corporation. I must confess my agitation is real, and my words are my bond in this proposal. My late father diverted this fund acquired from the over influencing of price of sales/purchasing of raw materials., now he deposited the money with a BANK IN ABIDJAN BY FIXED DEPOSIT FORM, where I am residing under political asylum with my younger sister Juliet who is 17 years old. Now the war in my country is overwith the help of ECOMOG soldiers, the present government of Sierra Leone has revoked the passport of all officers who served under the former regime and now ask countries to expel such person at the same time, freeze their account and confiscate their assets, it is on this note that I am contacting you, all I needed from you is to furnish me with your bank particulars:

- 1) Bank name
- 2) Account name
- 3) Account number
- 4) Bank address, telephone and fax numbers to enable me transfer this money in your private bank account, the said amount is twenty Six Million United States Dollars (US\$26.000.000.00).

I am compensating you with 20% of the total money amount, now all my hope is banked on you and I really want to invest this money in your country, where there is stability of Government, political and economic welfare. Honestly I want you to believe that this transaction is real and never a joke. My late father Chief Kamara gave me the photocopy of the deposit certificate issued to him by the BANK IN ABIDJAN on the date of deposit, and he called me closer to his bed side before his call to glory (R.I.P) that I should pray to God first, before contacting any foreigner and he warned me strictly not to deal with a greedy and evil minded people since this is the only legacy we are inheriting from him.

Please, for you to be clarified because, I do not expose myself to anybody I see, I believe that you are able to keep this transaction secret for me because this money is the hope of my life, it is important. Please contact me immediately after you must have gone through my message, and feel free to make it urgent. That is the reason why I offered you 20 % of the total amount, and in case of any other necessary expenses you might incur during this transaction including your telephone calls.

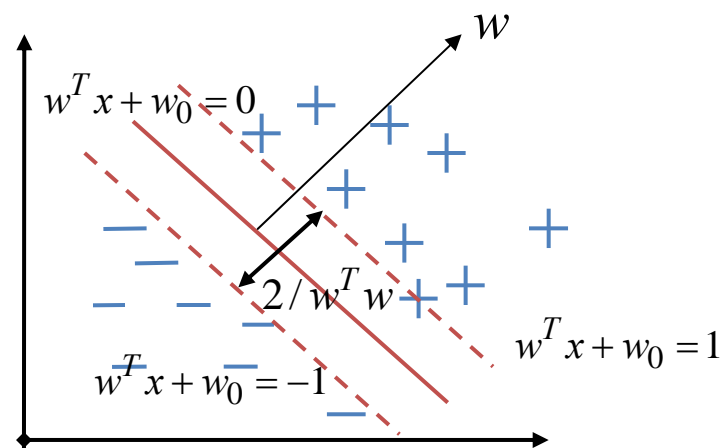
Is this email spam/fraud?

Classification - Support Vector Machine

Given a labeled training set, $\{x_i, y_i\}_{i=1\dots n}$ $x_i \in \mathbb{R}^d, y_i \in \{-1, 1\}$

Want to learn $f(x; w) = \text{sgn}(w^T x + w_0)$

$$\begin{aligned} \min \quad & w^T w \\ \text{s.t.} \quad & y_i(w^T x_i + w_0) \geq 1 \quad \forall i \end{aligned}$$

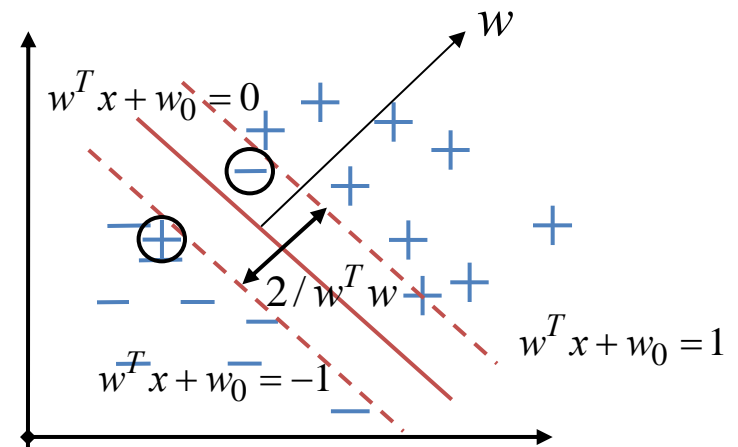


Classification - Support Vector Machine

Given a labeled training set, $\{x_i, y_i\}_{i=1\dots n}$ $x_i \in \mathbb{R}^d, y_i \in \{-1, 1\}$

Want to learn $f(x; w) = \text{sgn}(w^T x + w_0)$

$$\begin{aligned} \min \quad & w^T w + C \sum_i \xi_i \\ \text{s.t.} \quad & y_i(w^T x_i + w_0) \geq 1 - \xi_i \quad \forall i \\ & \xi_i \geq 0 \end{aligned}$$



Classification - Support Vector Machine

Given a labeled training set, $\{x_i, y_i\}_{i=1\dots n}$ $x_i \in \mathbb{R}^d, y_i \in \{-1, 1\}$

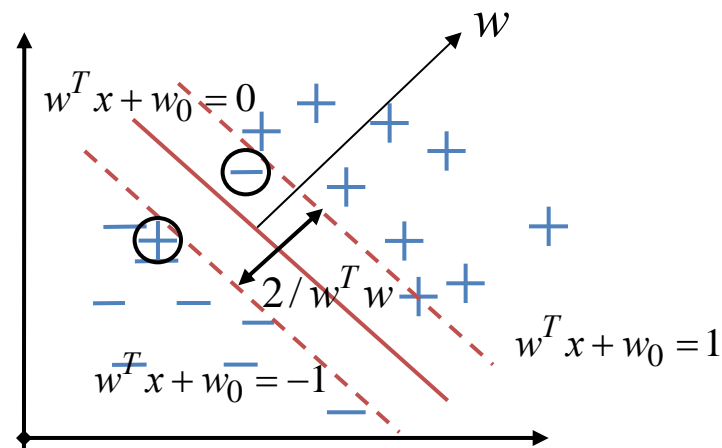
Want to learn $f(x; w) = \text{sgn}(w^T x + w_0)$

$$\begin{aligned} \text{Primal} \quad & \min w^T w + C \sum_i \xi_i \\ \text{s.t.} \quad & y_i(w^T x_i + w_0) \geq 1 - \xi_i \quad \forall i \\ & \xi_i \geq 0 \end{aligned}$$

Using Lagrange multipliers (with KKT conditions)

$$w = \sum_i \alpha_i y_i x_i$$

$$\begin{aligned} \text{Dual} \quad & \max \sum_{i=1}^n \alpha_i - \sum_{i,j} \alpha_i (y_i y_j x_i^T x_j) \alpha_j \\ & \sum_i \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C \end{aligned}$$



Fast training in $O(n)$

cutting-plane
stochastic gradient descent
quasi-Newton
coordinate descent

Testing $O(d)$

Classification - Kernel SVM

Given a labeled training set, $\{x_i, y_i\}_{i=1\dots n}$ $x_i \in \mathbb{R}^d, y_i \in \{-1, 1\}$

$$f(x; w) = \text{sgn}(w^T \Phi(x) + w_0) \quad k(x, y) = \Phi(x) \cdot \Phi(y)$$

Cannot solve in primal
since $\Phi(x)$ is unknown!

$$w = \sum_i \alpha_i y_i \Phi(x_i)$$

Classification - Kernel SVM

Given a labeled training set, $\{x_i, y_i\}_{i=1\dots n}$ $x_i \in \mathbb{R}^d, y_i \in \{-1, 1\}$

$$f(x; w) = \text{sgn}(w^T \Phi(x) + w_0) \quad k(x, y) = \Phi(x) \cdot \Phi(y)$$

Cannot solve in primal
since $\Phi(x)$ is unknown!

$$w = \sum_i \alpha_i y_i \Phi(x_i)$$

$$\text{Dual} \quad \max \alpha^T 1 - \alpha^T K \alpha$$

$$\alpha^T y = 0$$

$$0 \leq \alpha_i \leq C$$

Training $O(n^2) \sim O(n^3)$

Testing $O(\#_{sv}) \approx O(n)$

$$f(x; \alpha) = \text{sgn}\left(\sum_i \alpha_i y_i k(x, x_i) + \alpha_0\right)$$

Low-rank approximation of K

Linear approximation of K

Ranking

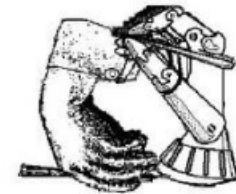
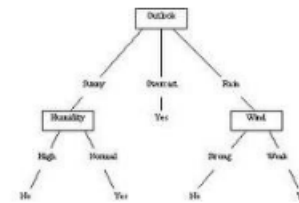
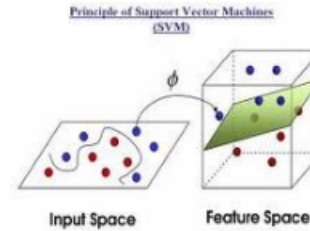
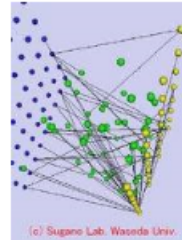
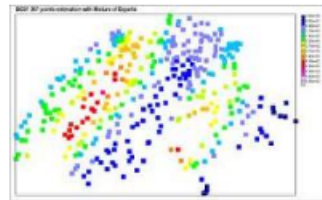
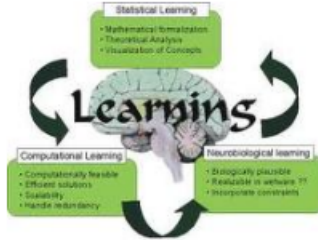
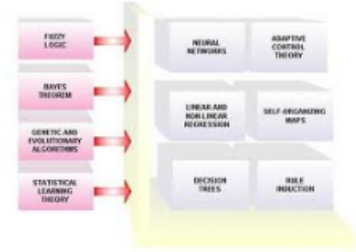
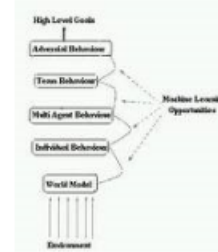
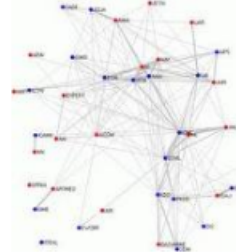
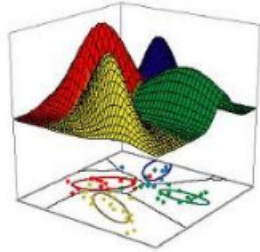
machine learning

Search

SafeSearch moderate ▾

About 11,900,000 results (0.52 seconds)

Advanced search



Ranking

A simple margin-based formulation

Given a query q and a database X

$$f(q, x) = q^T W x \quad q, x \in \mathbb{R}^d$$

Given many (query, relevant, irrelevant) triplets

$$\min \sum_{(q, x_+, x_-)} \max(0, 1 - q^T W x_+ + q^T W x_-) + \lambda \|W\|_F^2$$

Ranking


A simple margin-based formulation

Given a query q and a database X

$$f(q, x) = q^T W x \quad q, x \in \mathbb{R}^d$$

Given many (query, relevant, irrelevant) triplets

$$\min \sum_{(q, x_+, x_-)} \max(0, 1 - q^T W x_+ + q^T W x_-) + \lambda \|W\|_F^2$$

 $d \times d$

$O(n^3)$ triplets for dataset of size n

$O(d^2)$ complexity – large d in many applications

Low-rank approximation or sparse W

First order methods very successful

Distance Metric Learning

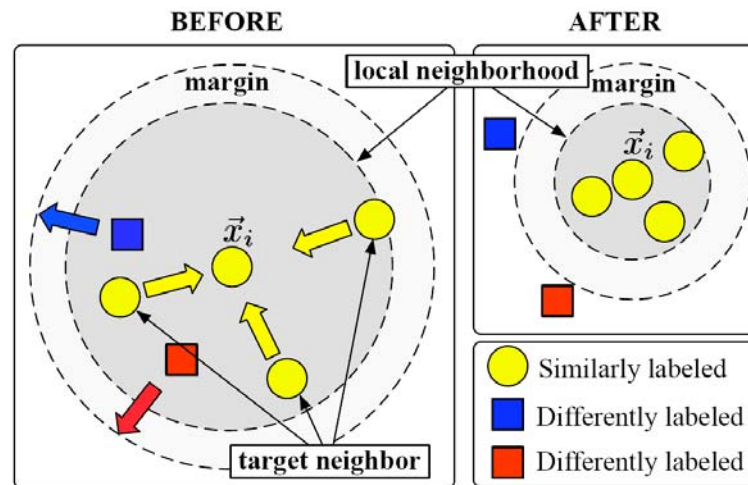


Given a large collection of items and associated features, find a similarity metric!

Distance Metric Learning

Many applications

- kNN-classification, clustering, density estimation, graph construction



$$\min \sum_{(x_i, x_j) \in S} d_A(x_i - x_j)$$

$$d_A(x_i - x_k) - d_A(x_i - x_j) \geq 1 \quad \forall (i, j, k)$$

$$A \succ 0$$

Semi-Definite Programming (SDP): $\sim O(d^3)$

[Weinberger K. et al.]

Unsupervised - Kernel Density Estimation

Given an unlabeled training set, $\{x_i\}_{i=1\dots n}$ $x_i \in \mathcal{R}^d$ learn a nonparametric density function $p(x)$

- To compute data likelihood (e.g. for Bayesian prediction)
- For large datasets, nonparametric density can approximate true density very well

Explore flu trends - United States

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. [Learn more »](#)



Unsupervised - Kernel Density Estimation

Given a unlabeled training set, $\{x_i\}_{i=1\dots n}$ learn a nonparametric density function $p(x)$

$$p(x) = \frac{1}{n} \sum_{i=1}^n \underbrace{\frac{1}{h} k\left(\frac{x - x_i}{h}\right)}_{N(x_i, \sigma^2 I)}$$



- Too expensive for large n
- Many kernels have rapid (exponential) fall off
- Nearest Neighbors sufficient but expensive for large n

Large-scale Approximate NN search

- Similar arguments for training parametric models e.g., Gaussian Mixture Models with large number of components!

Dimensionality Reduction / Clustering



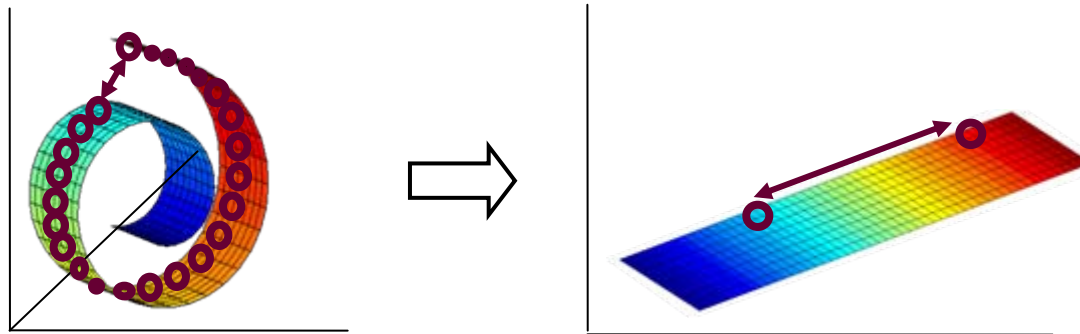
- Each face represented by 10K-dim vector (i.e., pixels)
- Faces are structured objects so dim reduction is possible
- Extract millions of face images from Web
- Can we also find meaningful clusters?

Dimensionality Reduction

Linear dimensionality reduction

- Principal Component Analysis (PCA), Multidimensional Scaling (MDS)
- Need to compute singular vectors of $d \times n$ data matrix X $O(nd^2)$
- **Randomized methods:** randomly sample a few directions from a fixed distribution for projections – with high probability preserves the pairwise distance given enough projections

Nonlinear dimensionality reduction (manifold learning)



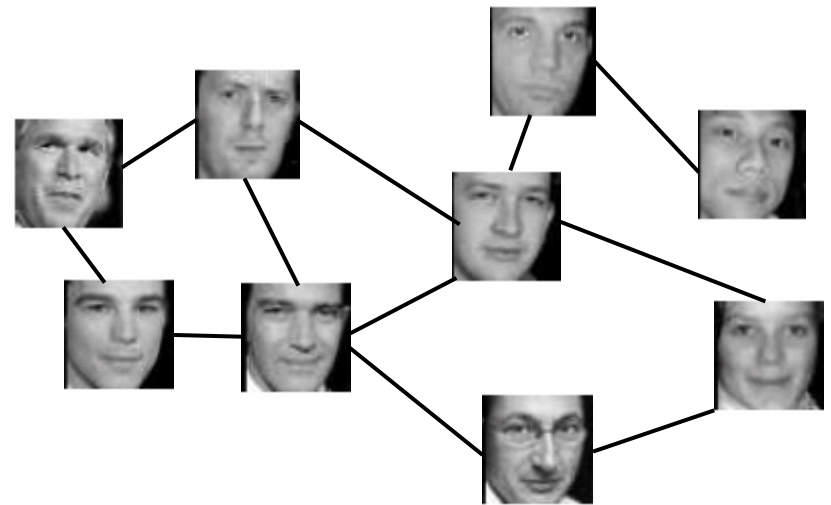
Laplacian Eigenmaps/Spectral Clustering

Minimize weighted distances between neighbors

1. Find t nearest neighbors for each image : $O(n^2)$

2. Compute weight matrix W :

$$W_{ij} = \begin{cases} \exp(-\|x_i - x_j\|^2 / \sigma^2) & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases}$$



Laplacian Eigenmaps/Spectral Clustering

Minimize weighted distances between neighbors

1. Find t nearest neighbors for each image : $O(n^2)$

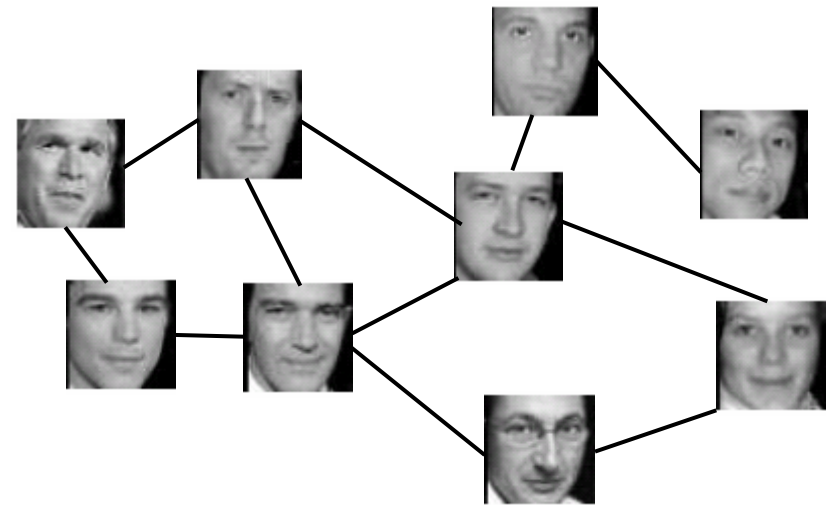
For 10M points

- 155 yrs on one machine
- 10 weeks on 1000 machines

Approximate Nearest Neighbors

2. Compute weight matrix W :

$$W_{ij} = \begin{cases} \exp(-\|x_i - x_j\|^2 / \sigma^2) & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases}$$



Laplacian Eigenmaps/Spectral Clustering

Minimize weighted distances between neighbors

1. Find t nearest neighbors for each image : $O(n^2)$

Approximate Nearest Neighbors

2. Compute weight matrix W :

$$W_{ij} = \begin{cases} \exp(-\|x_i - x_j\|^2 / \sigma^2) & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases}$$

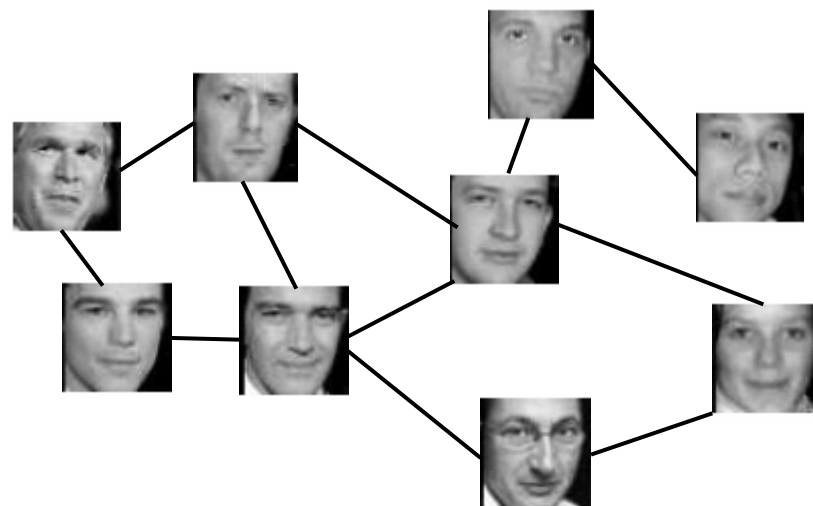
3. Compute normalized Laplacian

$$G = I - D^{-1/2} W D^{-1/2}$$

$$D_{ii} = \sum_j W_{ij}$$

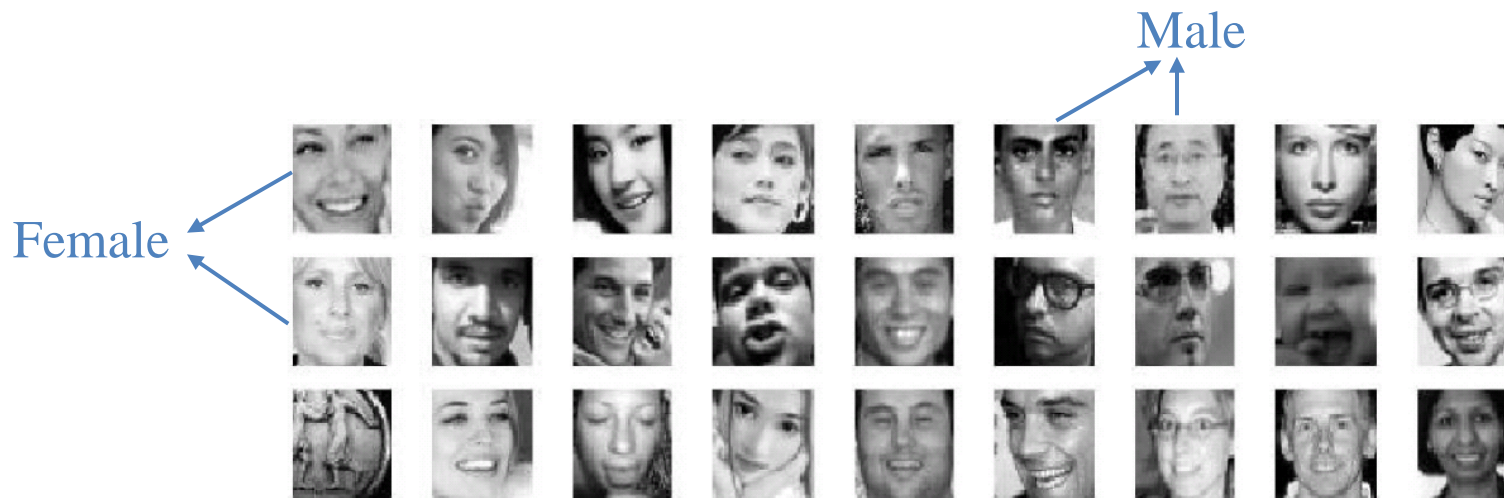
4. Optimal k reduced dims: U_k

Bottom eigenvectors of G , ignoring last



Fast matrix-vector product based methods for sparse matrix

Semisupervised



- Goal: **Gender Classification**
- A few labeled faces are available but **many** unlabeled

Semisupervised - Graph Propagation

Given a labeled set, $\{(x_i, y_i)\}_{i=1\dots l}$ $x_i \in \mathbb{R}^d$, $y_i \in \{1, \dots, C\}$
and an unlabeled set $\{x_i\}_{i=l+1, \dots, l+u}$

Want to find the labels for unlabeled points $\{y_i\}_{i=l+1, \dots, l+u}$

Intuition: Neighboring points have similar labels

Steps:

1. Build a sparse neighborhood graph using all the points
2. Construct the symmetric weight matrix W
3. Compute a stochastic transition matrix T by normalizing columns of W

$$Y_u = (I - T_{uu})^{-1} T_{ul} Y_l$$

Semisupervised - Graph Propagation

Given a labeled set, $\{(x_i, y_i)\}_{i=1\dots l}$ $x_i \in \mathbb{R}^d$, $y_i \in \{1, \dots, C\}$
and an unlabeled set $\{x_i\}_{i=l+1, \dots, l+u}$

Want to find the labels for unlabeled points $\{y_i\}_{i=l+1, \dots, l+u}$

Intuition: Neighboring points have similar labels

Steps:

1. Build a sparse neighborhood graph using all the points
2. Construct the symmetric weight matrix W
3. Compute a stochastic transition matrix T by normalizing columns of W

$$Y_u = \underbrace{(I - T_{uu})^{-1} T_{ul}}_{n \times n \text{ sparse}} Y_l$$

Manifold regularization: Loss over labeled pts and smoothness of labels over all the points \rightarrow Complexity similar to manifold learning

Topics

Randomized Algorithms

Matrix Approximations I (low-rank approximation, decomposition)

Matrix Approximations II (sparse matrices, matrix completion)

Approximate Nearest Neighbor Search I (trees)

Approximate Nearest Neighbor Search II (hashes)

Fast Optimization (first-order methods)

Kernel Methods I (fast training)

Kernel Methods II (fast testing)

Dimensionality Reduction (linear and nonlinear methods)

Sparse Methods/Streaming (sparse coding...)

Course Overview

13 classes, Tuesdays 12:35 – 2:25 pm

TA: Jun Wang / Junfeng He

Office hours: Tuesdays 2:25 pm – 3:25 pm (email: sanjivk@google.com)

Course Website: www.sanjivk.com/EECS6898

Evaluation

- Three assignments, 20% x 3, No midterm
- One final project (~6 weeks long), 40%
- Final based on project presentation and report

Focus on algorithms and tools that make large-scale learning possible!

- Can be applied to many existing machine learning algorithms
- Practical performance and utility