Simple Classification using Binary Data

Deanna Needell
Mathematics
UCLA

Thanks:

- NSF CAREER DMS#1348721
- NSF BIGDATA DMS#1740325
- MSRI NSF DMS#1440140
- Alfred P. Sloan Fdn

Joint work with

Rayan Saab (UCSD)

Tina Woolf (CGU)

So much data…

The Economist
Obama the warrior
Misgoverning Argentina
The economic shift from West to East
Genetically modified crops blossom
The right to eat cats and dogs
The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT

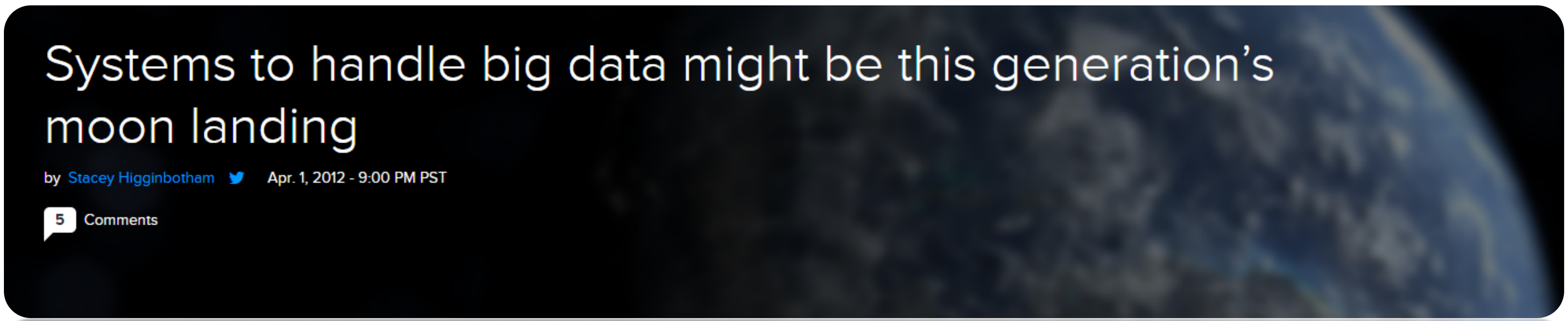
So much data…

COMPUTER
WHY BIG DATA IS A BIG DEAL
nature
The Economist
The data delu
POPULAR SCIENCE
THE FUTURE NOW
THE CONTROL CENTERS
OFFICER ALGORITHM
NEW WAYS OF SEEING
PLUS
SPECIAL ISSUE
DATA IS POWER
HOW INFORMATION IS DRIVING THE FUTURE

So much data…


METU SIAM Student Chapter Presents
16-24 May 2012
WHAT WOULD YOU DO WITH ALL THIS DATA?
Pictures
Gerhard Wilhelm Weber
Estimation of Dynamics under Uncertainty
May 16, Wednesday, 11:40-12:30 *
Annette Hohenberger
Fractals in Cognitive Science
May 17, Thursday, 11:40-12:30 *
Özlem İlk
Introduction to R and GGobi
May 17, Thursday, 14:00-17:30 **
Aybar Acar
MapReduce and Hadoop: Mining Big Data in the Cloud
May 23, Wednesday, 14:00-16:30 *
Cem İyigün
Introduction to Clustering
May 23, Wednesday, 11:40-12:30 *
Fatma Yerlikaya-Özkurt
Modeling with MARS and CMARS
May 24, Thursday, 14:00-15:30 *
* Institute of Applied Math, S 209
** Department of Mathematics, Computer Lab (M 202)
For more info visit http://siam.metu.edu.tr
Mathematics, Statistics, and the Data Deluge
MATHEMATICS AWARENESS MONTH

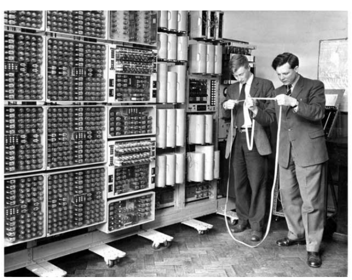
So much data…


Systems to handle big data might be this generation's moon landing
by Stacey Higginbotham Apr. 1, 2012 - 9:00 PM PST
5 Comments

How can we handle all this data?

Option 1 : Build bigger computing systems

- We need the resources
- Fundamental limitations
- Wasteful (resources, energy, cost, …)



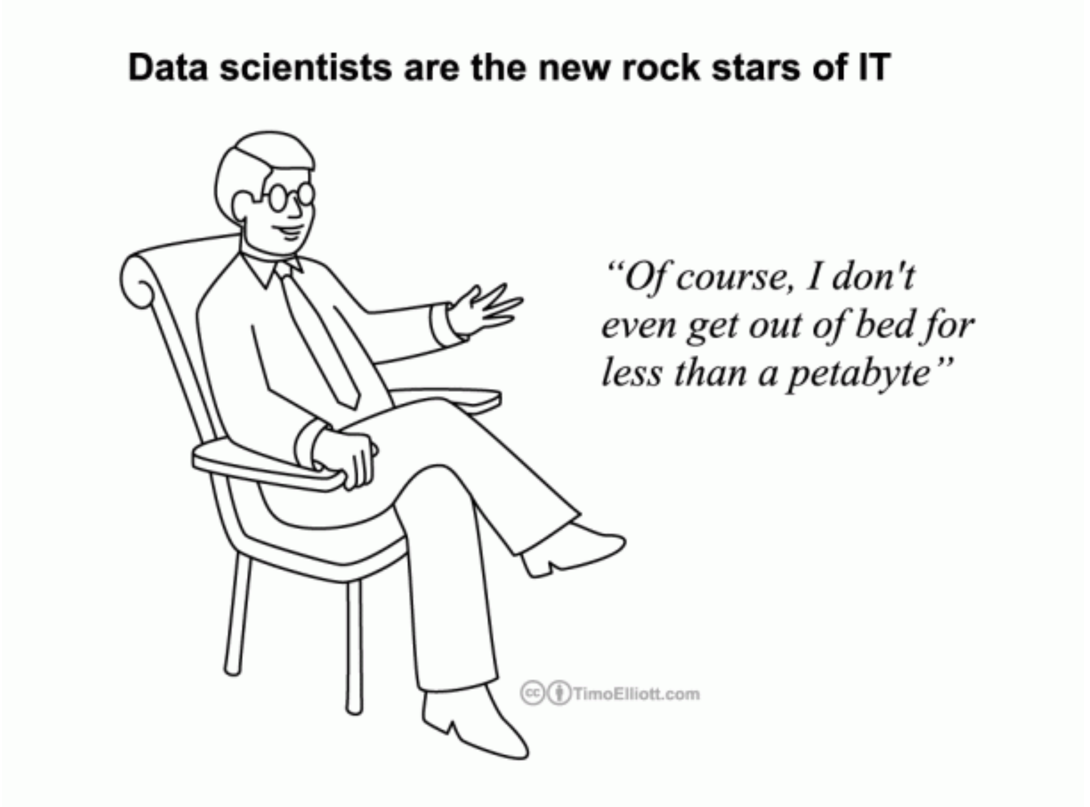
How can we handle all this data?

3 MB of internet data transfer = boiling one cup of water

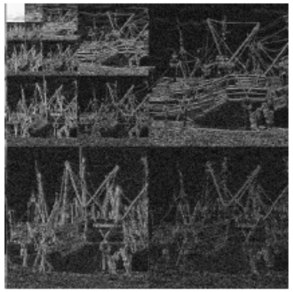
(https://www.katescomment.com/energy-of-downloads/)

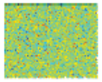
How can we handle all this data?

Option 2 : Design more efficient data analysis methods

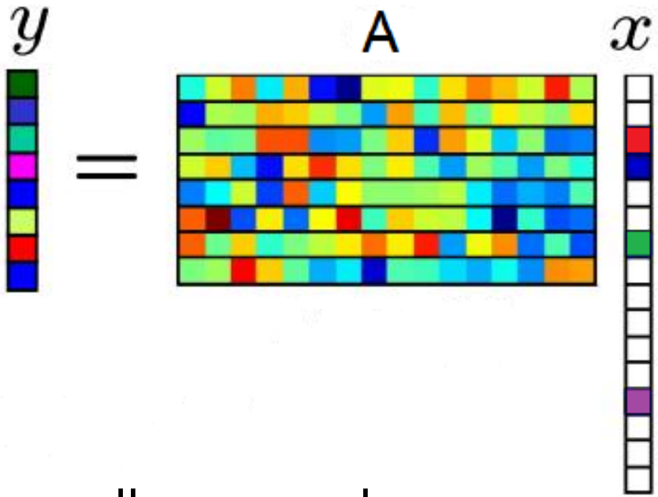


Compressed sensing





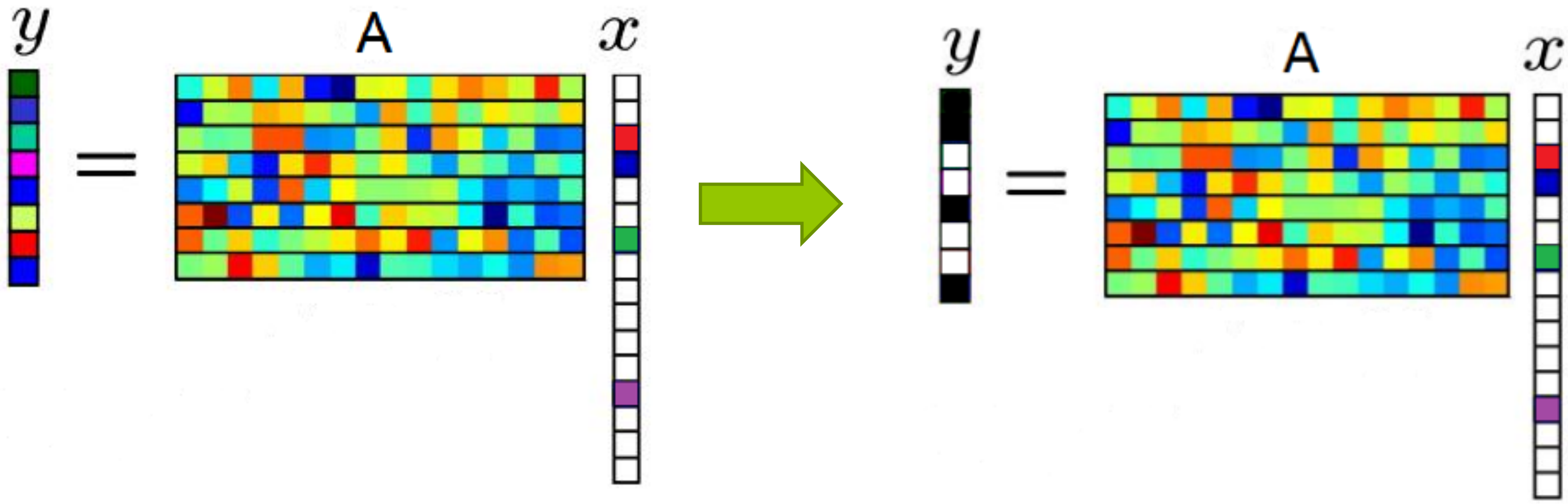
- ❖ Need to solve highly underdetermined linear system



- ❖ A has a null-space!

1-bit compressed sensing

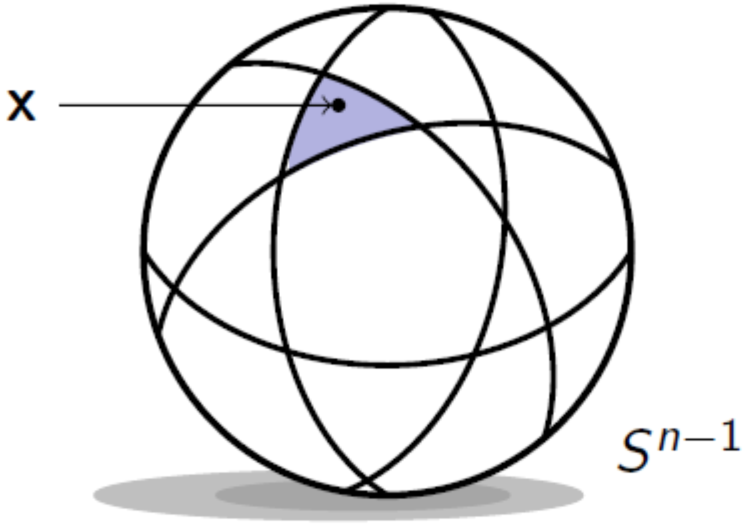
- Store only the first bit – the SIGN of each measurement



1-bit compressed sensing

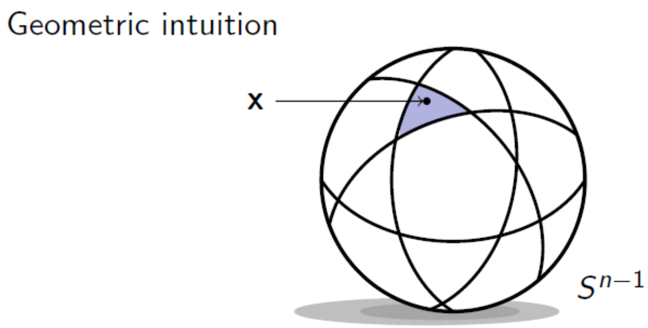
- Store only the first bit – the SIGN of each measurement

Geometric intuition



1-bit compressed sensing

- Store only the first bit – the SIGN of each measurement

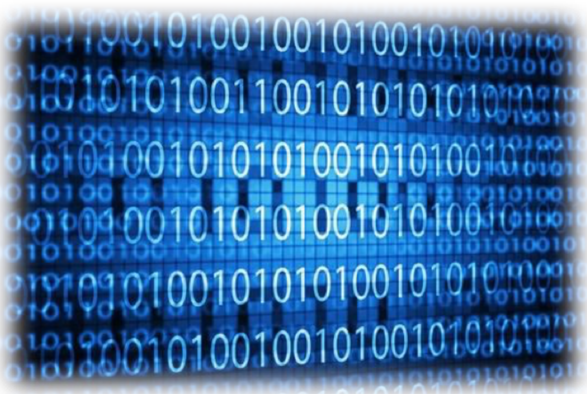


- Remedy: Use "dithers" to estimate the norm of x

$$y_i = \mathrm{sign}(\langle \mathbf{a}_i, \mathbf{x} \rangle - \tau_i)$$

Moral:

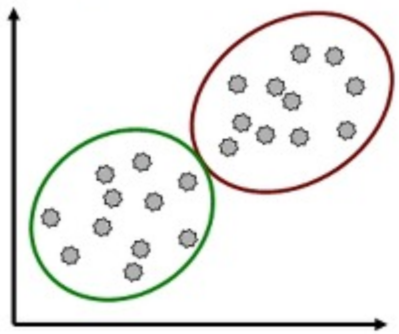
- One-bit (binary) data is as efficient as it gets
- It may still contain enough "information" about the signal to perform inference tasks



Problem: classification

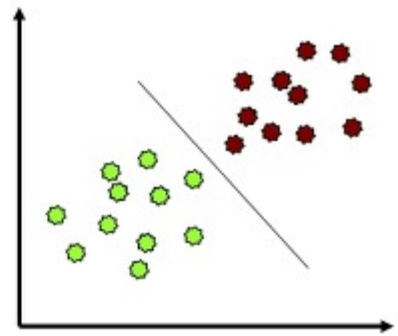
CLUSTERING

- Data is not labeled
- Group points that are "close" to each other
- Identify structure or patterns in data
- Unsupervised learning

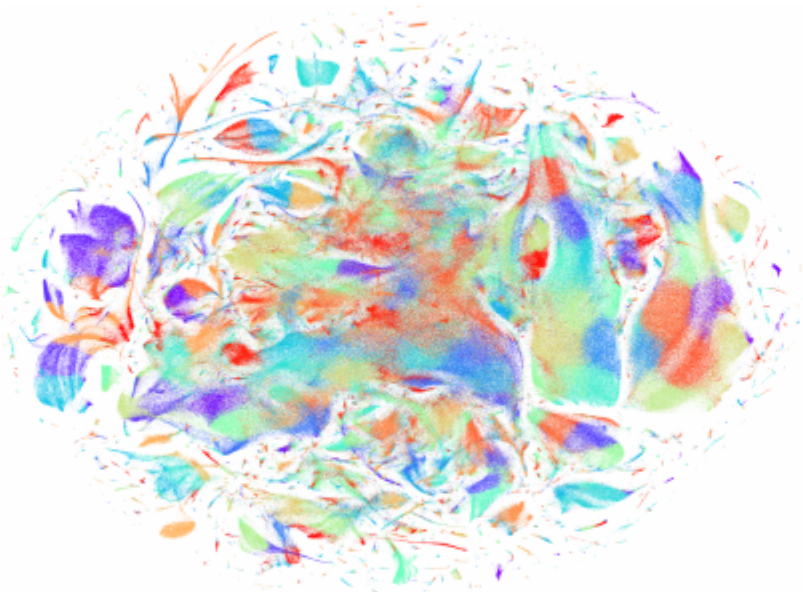


CLASSIFICATION

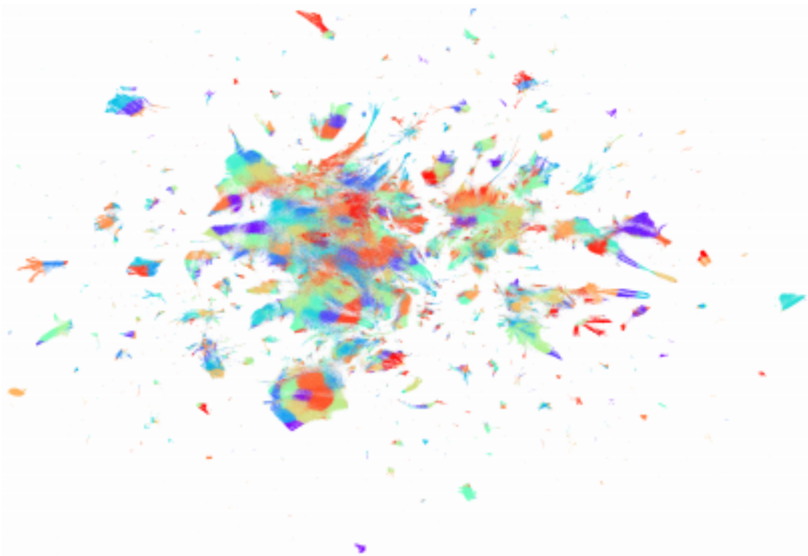
- Labeled data points
- Want a "rule" that assigns labels to new points
- Supervised learning



Problem: reality



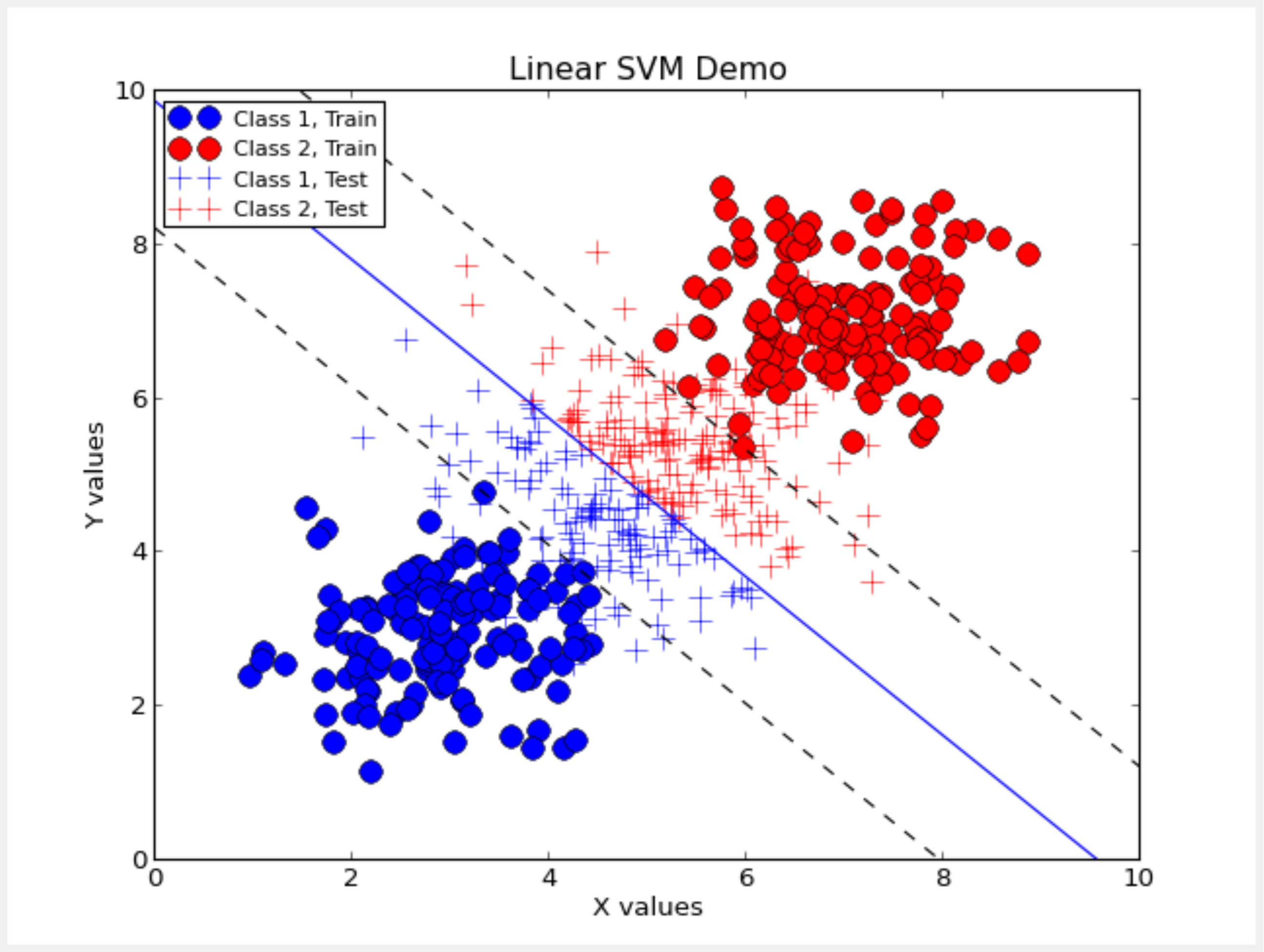
(c) WikiDoc (t-SNE)



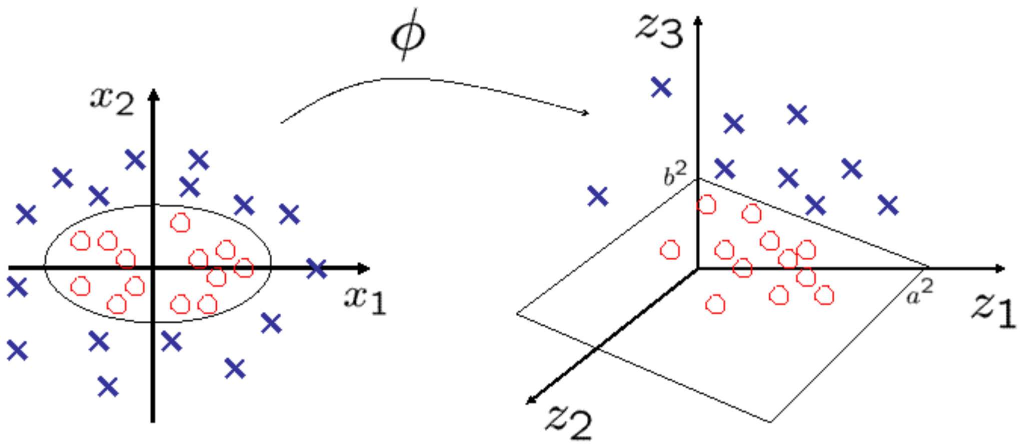
(d) WikiDoc (LargeVis)

(Tang etal. 2016)

Background: SVM



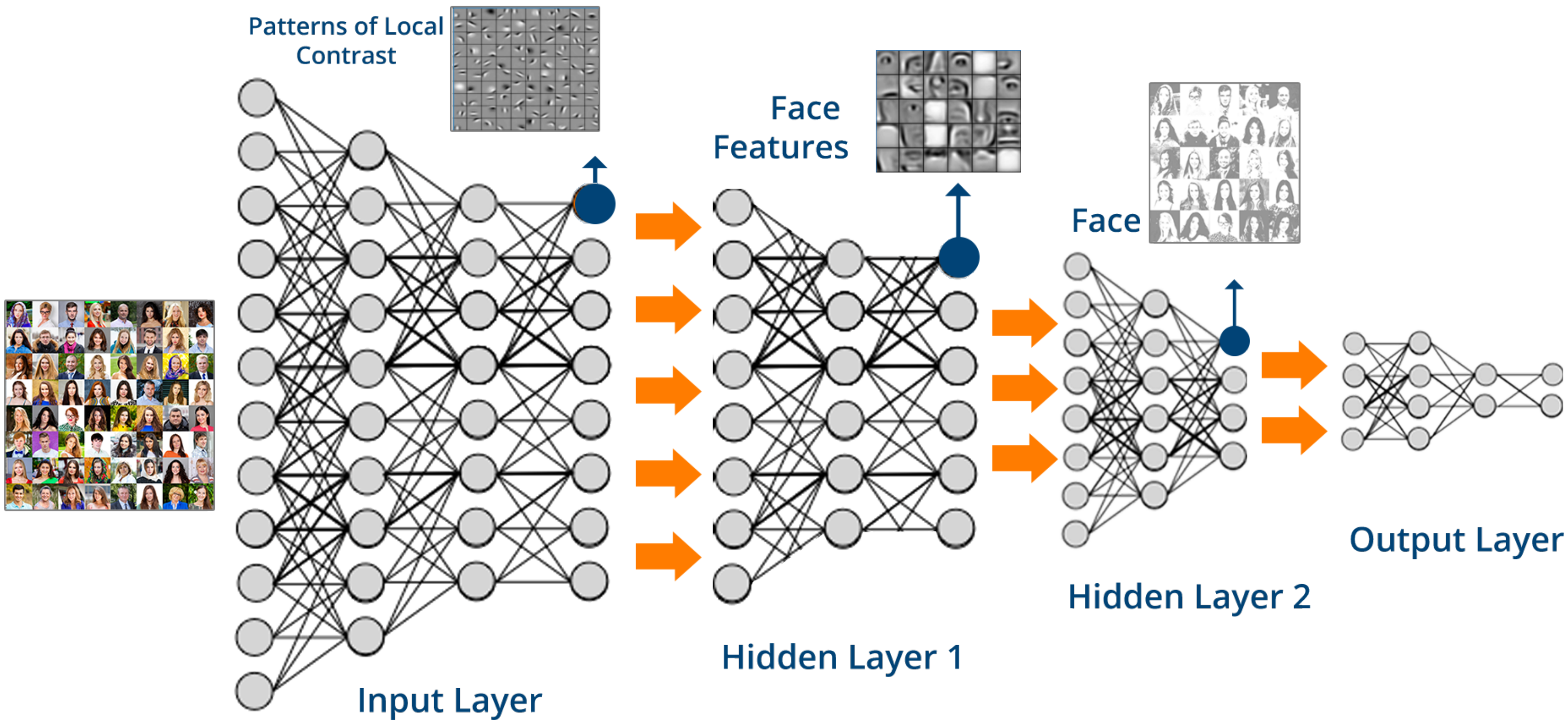
Background: SVM



$$\phi : (x_1, x_2) \longrightarrow (x_1^2, \sqrt{2}x_1x_2, x_2^2)$$

$$\left(\frac{x_1}{a}\right)^2 + \left(\frac{x_2}{b}\right)^2 = 1 \longrightarrow \frac{z_1}{a^2} + \frac{z_3}{b^2} = 1$$

Background: Deep Learning



Our goals

Design classification scheme that:

- ❖ Uses binary data
- ❖ Is simple and efficient
- ❖ Uses layers in an interpretable way

Notation

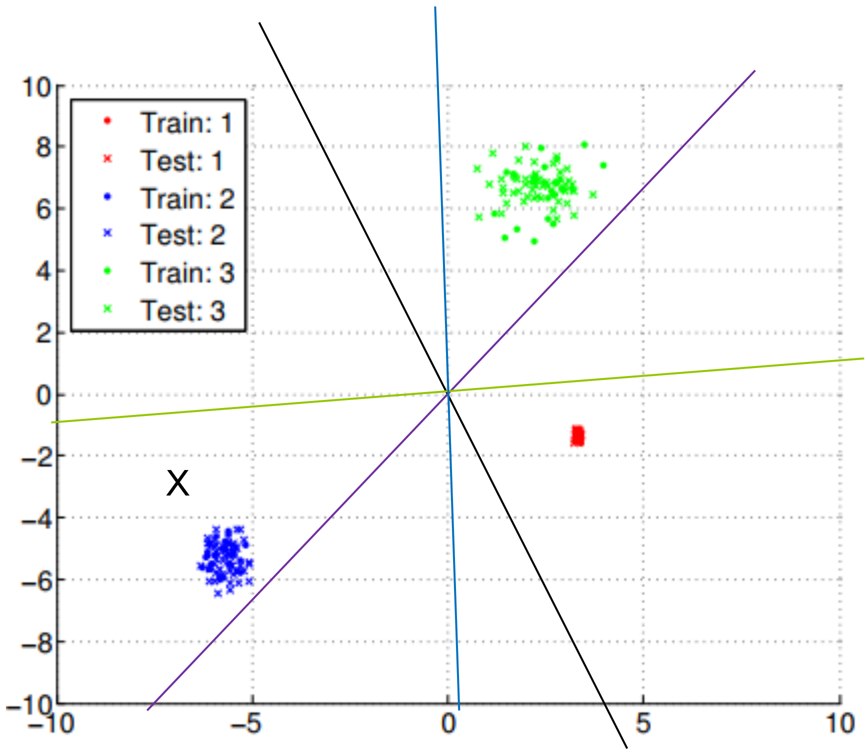
- X : nxp training data matrix (data in columns)
- A : mxn (random) matrix
- Q = sign(AX) : mxp binary training data
- G : # of classes
- b : p training labels (1-G)
- L : # of layers in our design

Main idea

- Each row of A corresponds to a hyperplane
- Each binary measurement Q_{ij} indicates on which side of the i^{th} hyperplane data point X_j lies
- If we gather enough of this info for all the training data, we can use it to predict the class label for a new test point x

Single layer

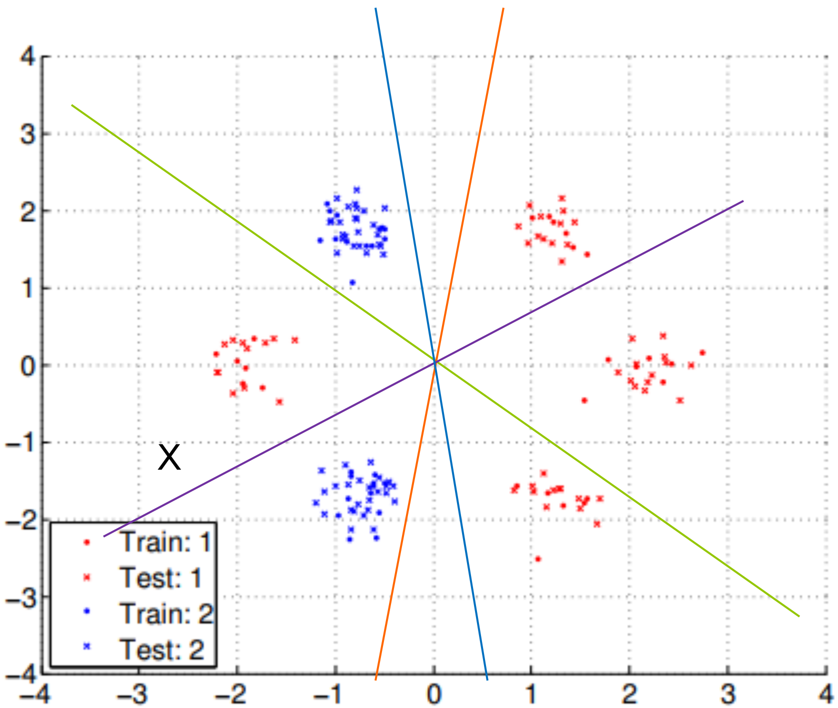
❖ All the hyperplane information in Q is enough



For a new point x, simply compare its sign pattern with those of the training points and choose the label it matches the most often

Multiple layers

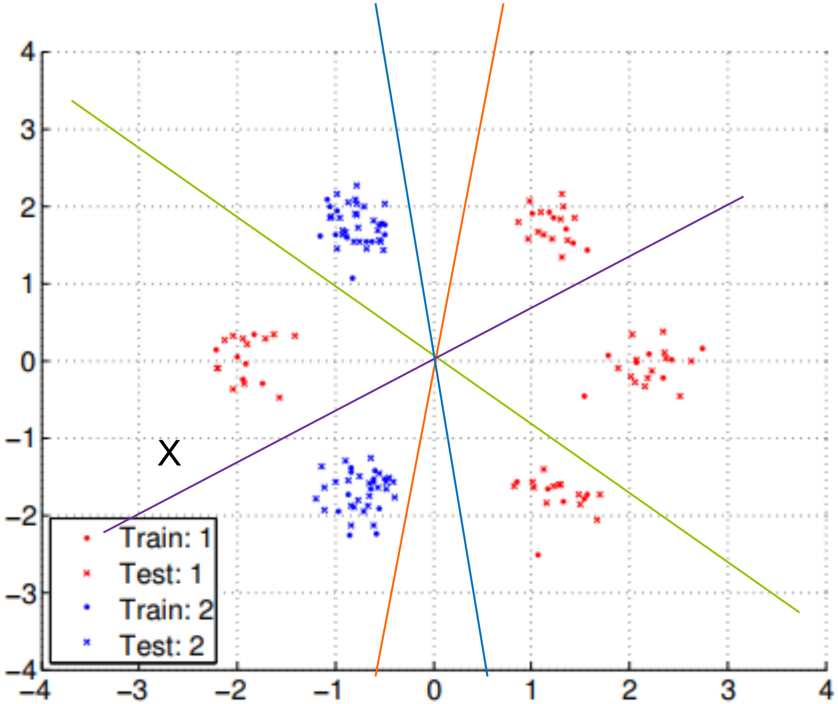
❖ What about?



For ANY hyperplane, there are both red and blue on at least one side of it

Multiple layers

- Now PAIRS of hyperplanes do the trick



For a new point x, simply compare its sign pattern for hyperplane PAIRS with those of the training points and choose the label it matches the most often

Multiple layers

What about higher dimensions?

- We continue this strategy to build layers, the l^{th} layer corresponding to l–tuples of hyperplanes
- For simplicity (and computation), we consider m l–tuples at each layer, selected randomly from all $\binom{m}{\ell}$ possible
- For a new test point x, we use the sign patterns across all layers for classification

Using all layers

How to integrate the info from all layers?

- l : layer
- i : index from 1 to m
- $\Lambda_{\ell,i}$: the set of l indices indicating which hyperplanes are selected in the ith l-tuple
- t : a possible sign pattern of l +/- 1s
- g : a class index (from 1 to G)
- $P_{g|t}$: the # of training points from the gth class having sign pattern t from the hyperplanes in $\Lambda_{\ell,i}$

Using all layers

- $P_{g|t}$: the # of training points from the g^th class having sign pattern t from the hyperplanes in $\Lambda_{\ell,i}$

$$r(\ell, i, t, g) = \frac{P_{g|t}}{\sum_{j=1}^{G} P_{j|t}} \frac{\sum_{j=1}^{G} |P_{g|t} - P_{j|t}|}{\sum_{j=1}^{G} P_{j|t}}$$

Using all layers

- $P_{g|t}$: the # of training points from the g[th] class having sign pattern t from the hyperplanes in $\Lambda_{\ell,i}$

$$r(\ell, i, t, g) = \frac{P_{g|t}}{\sum_{j=1}^{G} P_{j|t}} \frac{\sum_{j=1}^{G} |P_{g|t} - P_{j|t}|}{\sum_{j=1}^{G} P_{j|t}}$$

fraction of training points in class g out of all points with pattern t

Using all layers

- $P_{g|t}$: the # of training points from the g[th] class having sign pattern t from the hyperplanes in $\Lambda_{\ell,i}$

$$r(\ell, i, t, g) = \frac{P_{g|t}}{\sum_{j=1}^{G} P_{j|t}} \frac{\sum_{j=1}^{G} |P_{g|t} - P_{j|t}|}{\sum_{j=1}^{G} P_{j|t}}$$

fraction of training points in class g out of all points with pattern t

gives more weight to group g when its size is much different than others with same sign pattern

Our method: training

Algorithm 1 Training

input: binary training data Q, training labels b, number of classes G, number of layers L

for ℓ from 1 to L, i from 1 to m **do**
  select: Randomly select $\Lambda_{\ell,i} \subset [m]$, $|\Lambda_{\ell,i}| = \ell$
  determine: Determine the $T_{\ell,i} \in \mathbb{N}$ unique column patterns in $Q^{\Lambda_{\ell,i}}$
  for t from 1 to $T_{\ell,i}$, g from 1 to G **do**
    compute: Compute $r(\ell, i, t, g)$ by (1)
  end for
end for

- "For each layer, pick the l–tuples and then compute all values of r(l,i,t,g)"

Our method: testing

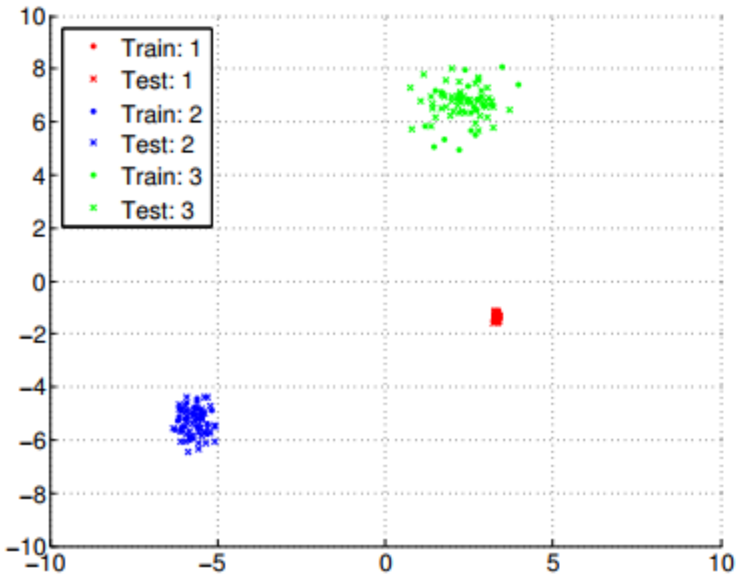
Algorithm 2 Classification

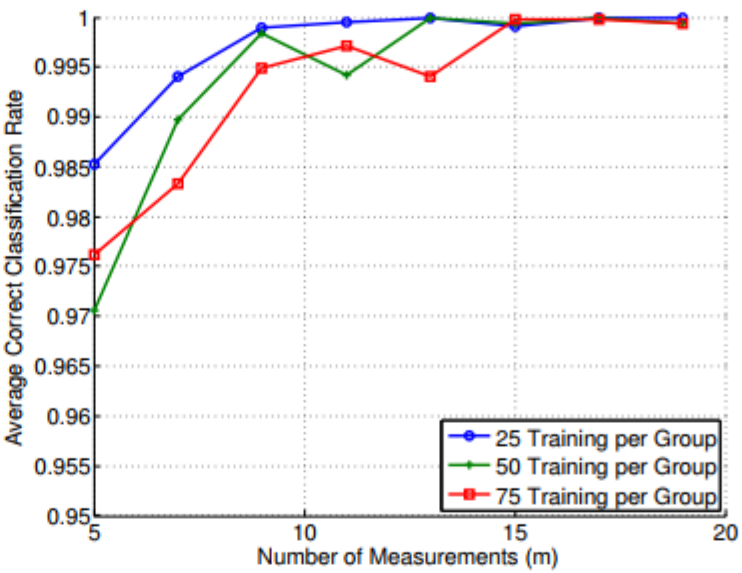
input: binary data q, number of classes G, number of layers L, learned parameters $r(\ell, i, t, g)$, $T_{\ell,i}$, and $\Lambda_{\ell,i}$ from Algorithm 1

initialize: $\tilde{r}(g) = 0$ for $g = 1, \ldots, G$.
for ℓ from 1 to L, i from 1 to m **do**
  identify: Identify the pattern $t^\star \in [T_{\ell,i}]$ to which $q^{\Lambda_{\ell,i}}$ corresponds
  for g from 1 to G **do**
    update: $\tilde{r}(g) = \tilde{r}(g) + r(\ell, i, t^\star, g)$
  end for
end for
scale: Set $\tilde{r}(g) = \frac{\tilde{r}(g)}{Lm}$ for $g = 1, \ldots, G$
classify: $\widehat{b}_x = \text{argmax}_{g \in \{1,\ldots,G\}} \tilde{r}(g)$

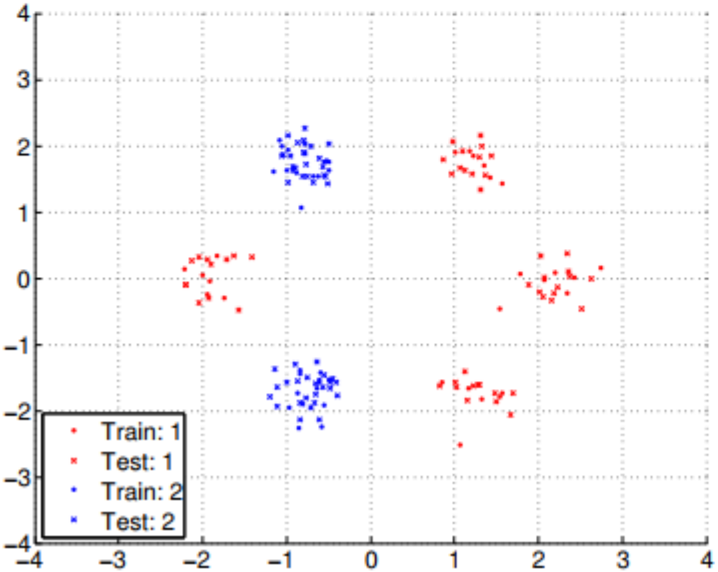
- "For a new point x with sign pattern t*, compute the sum of all r(l,i,t*,g) for each class g and then assign the label g which has the largest sum."

Results (L=1)

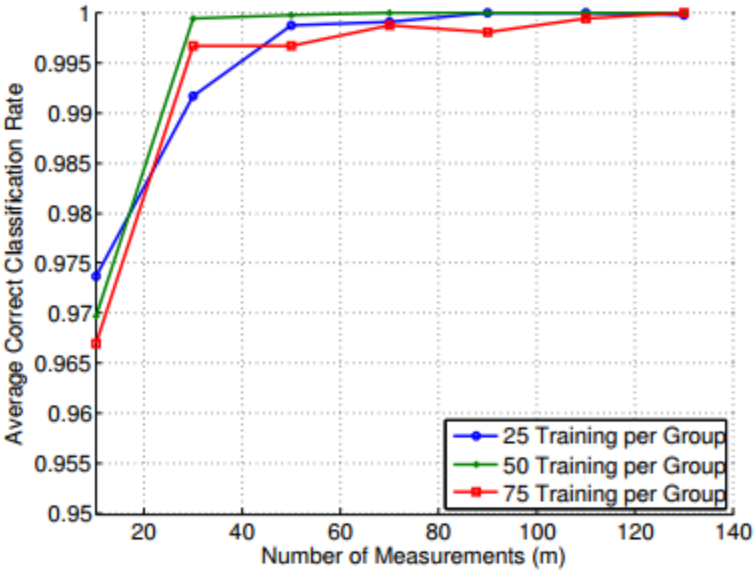

Train: 1
Test: 1
Train: 2
Test: 2
Train: 3
Test: 3


Average Correct Classification Rate
Number of Measurements (m)
25 Training per Group
50 Training per Group
75 Training per Group

Results (L=4)

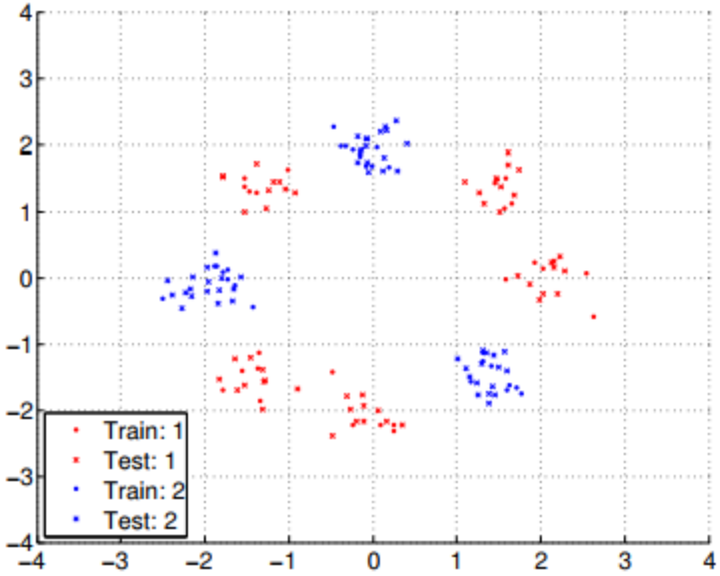


Train: 1
Test: 1
Train: 2
Test: 2

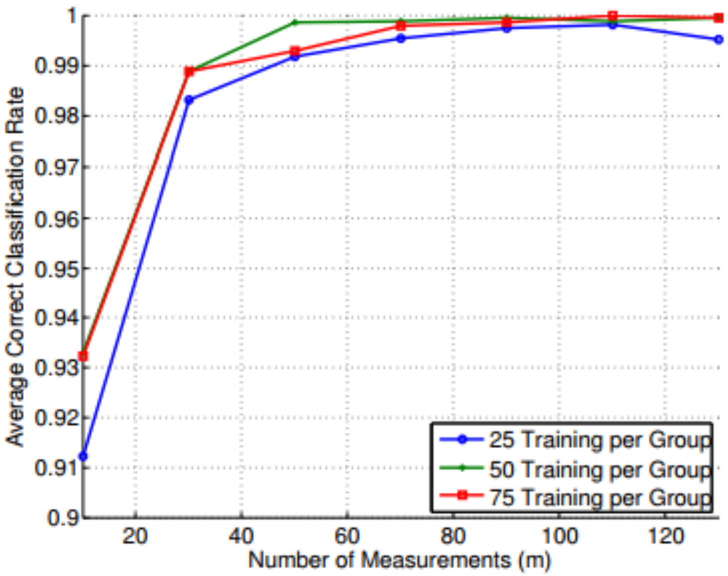


Average Correct Classification Rate
Number of Measurements (m)
25 Training per Group
50 Training per Group
75 Training per Group

Results (L=5)

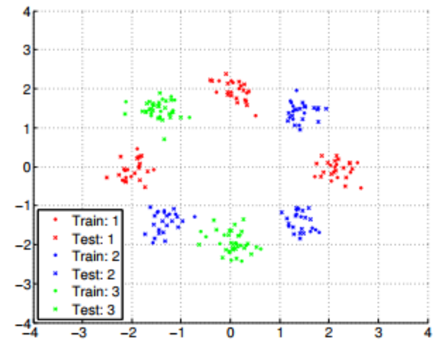


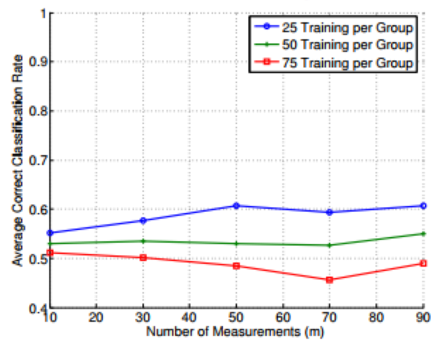
Train: 1
Test: 1
Train: 2
Test: 2



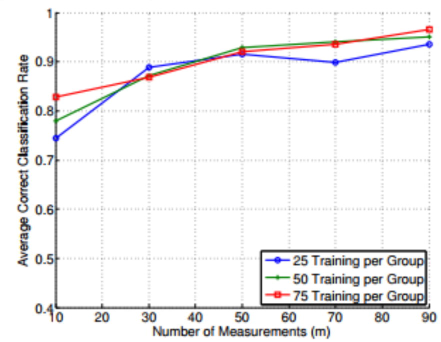
Average Correct Classification Rate
Number of Measurements (m)
25 Training per Group
50 Training per Group
75 Training per Group

Results

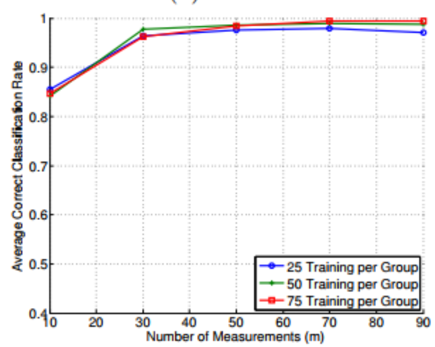




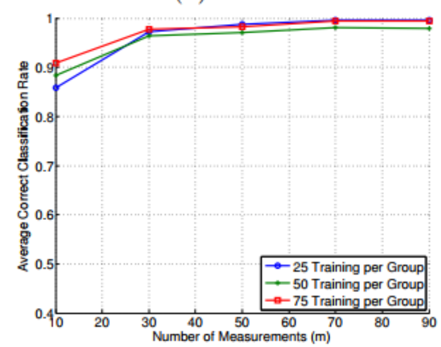
(a) $L = 1$



(b) $L = 2$

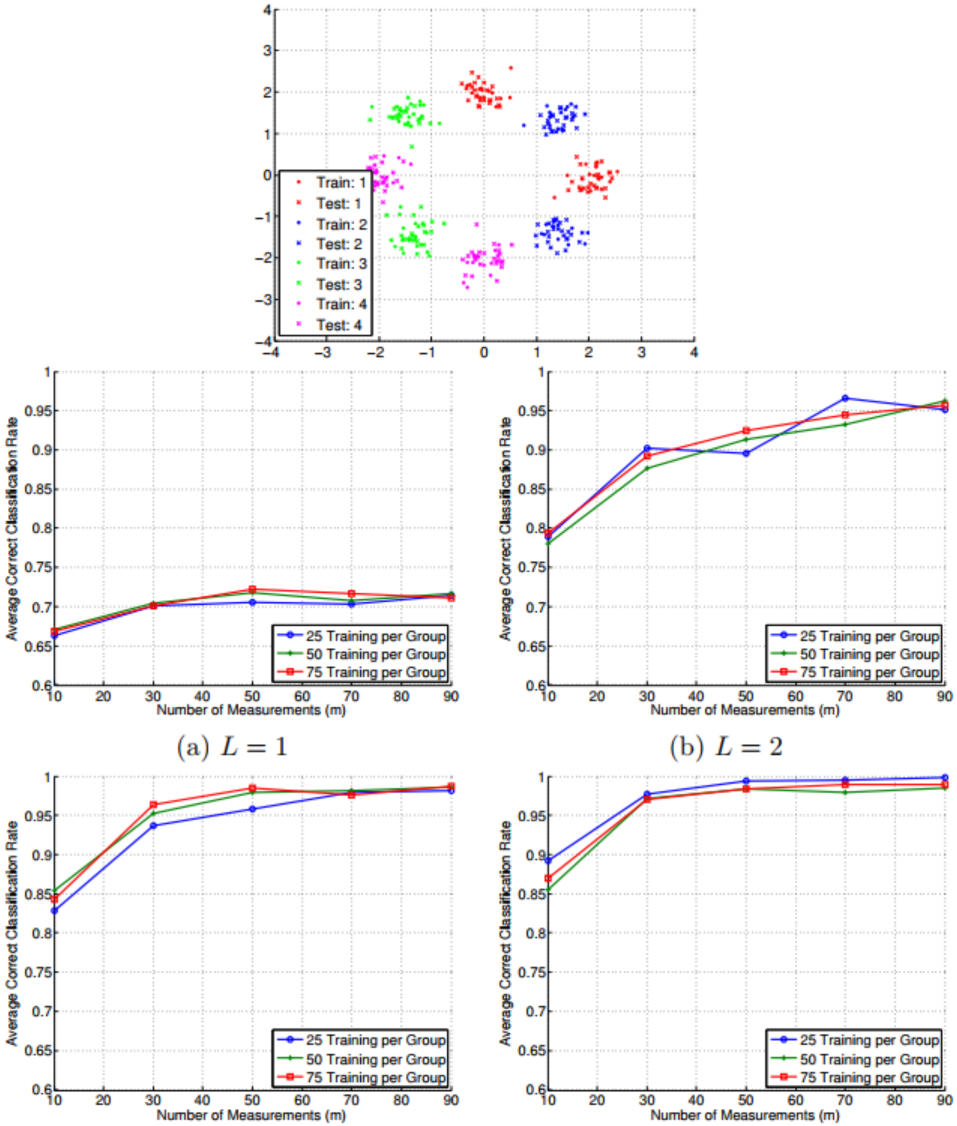


(c) $L = 3$



(d) $L = 4$

Results

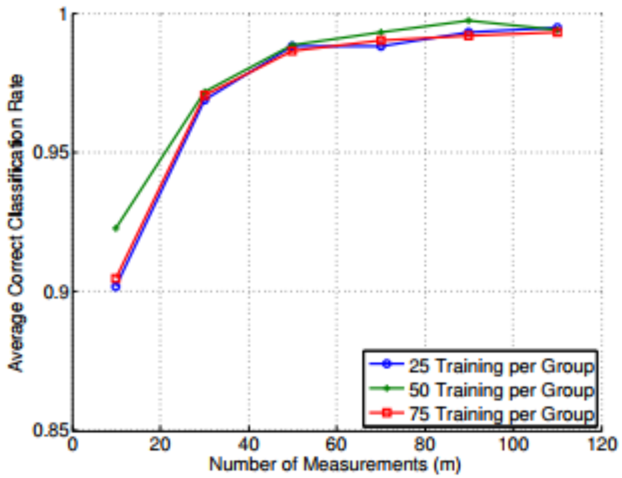


(a) $L = 1$ (b) $L = 2$

(c) $L = 3$ (d) $L = 4$

Results (L=1)

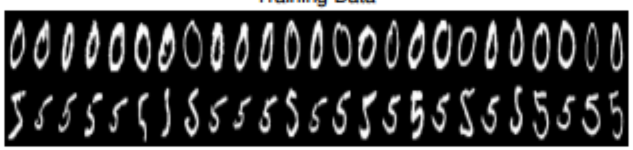
Training Data

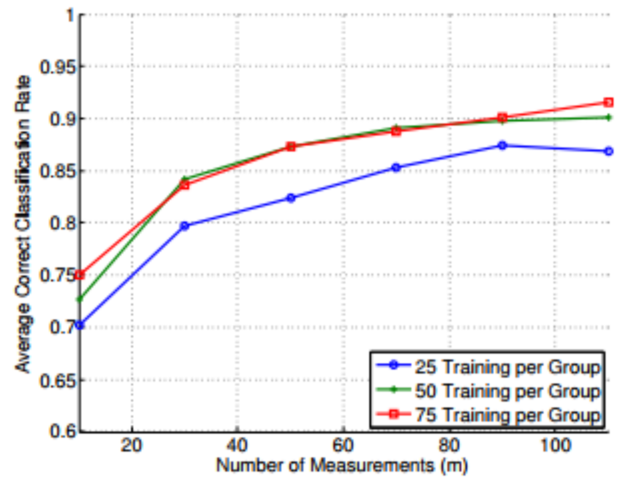


Testing Data

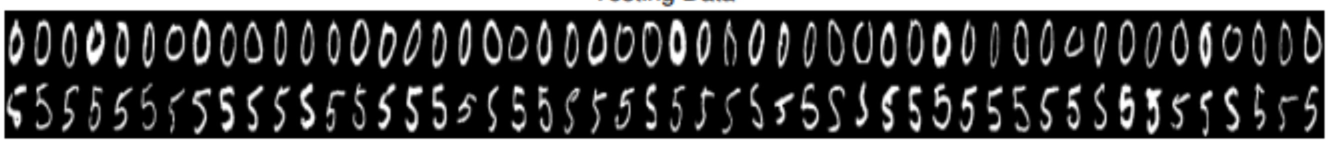
Results (L=4)

Training Data



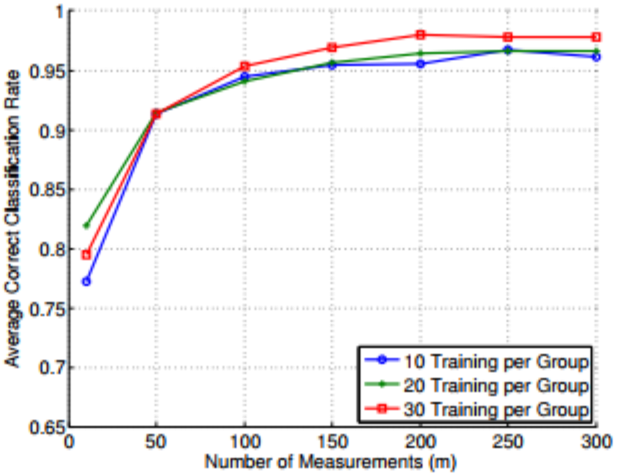


Testing Data



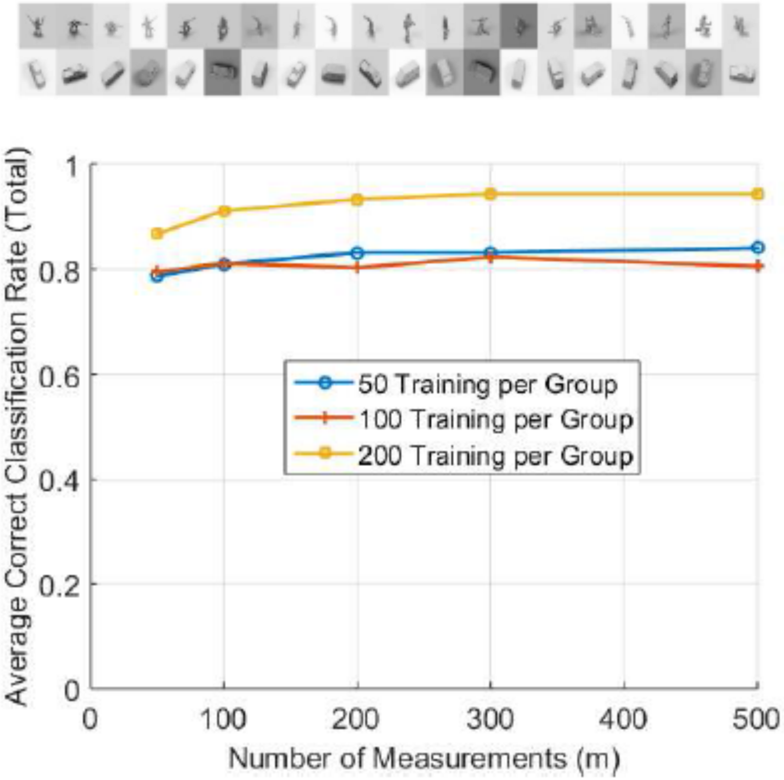
Results (L=5)

Training Data



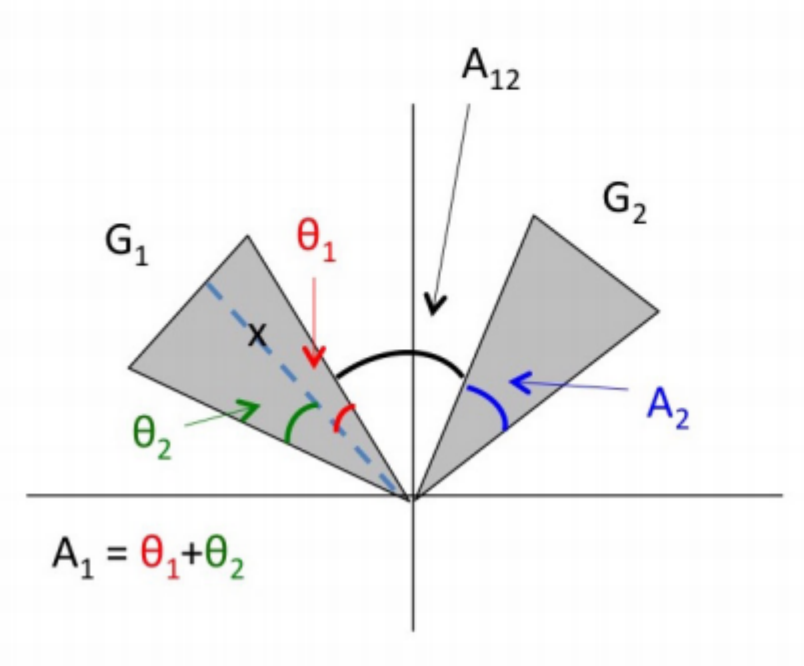
Testing Data

Results (L=5)

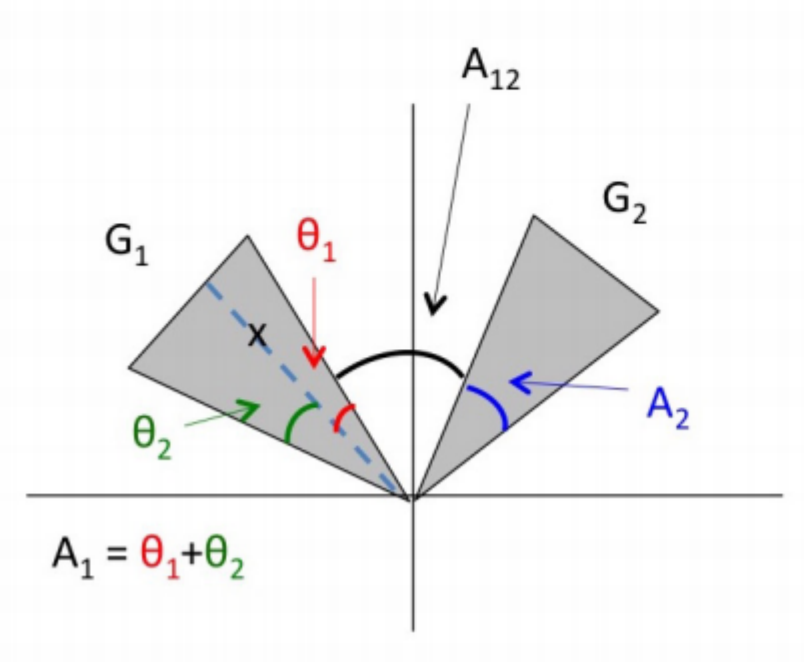


(b) $L = 5$

Theory



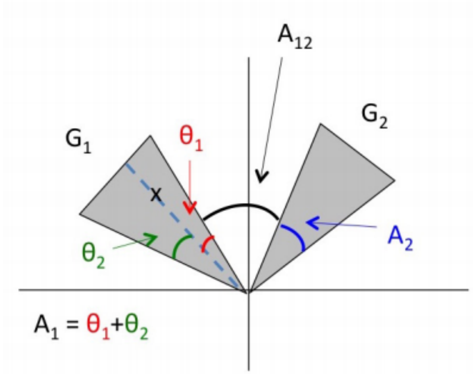
Theory



$A_{g|t}$: the angle of class g with sign pattern t for the i^{th} l – tuple in layer l

$$r(\ell, i, t, g) = \frac{A_{g|t}}{\sum_{j=1}^{G} A_{j|t}} \frac{\sum_{j=1}^{G} |A_{g|t} - A_{j|t}|}{\sum_{j=1}^{G} A_{j|t}}$$

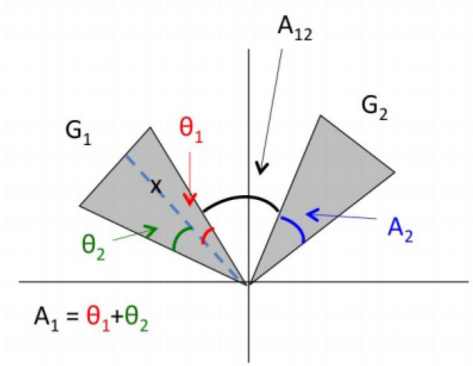
Theory



Theorem 1. *Let the classes G_1 and G_2 be two cones in $\mathbb{R}^2$ defined by angular measures A_1 and A_2, respectively, and suppose regions of the same angular measure have the same density of training points. Suppose $A_1 = A_2$, $\theta_1 = \theta_2$, and $A_{12} + A_1 + A_2 \leq \pi$. Then, the probability that a data point $x \in G_1$ gets classified in class G_1 by Algorithms 1 and 2 using a single layer and a measurement matrix $A \in \mathbb{R}^{m \times 2}$ with independent standard Gaussian entries is bounded as follows,*

$$\mathbb{P}[\widehat{b}_x = 1] \geq 1 - \sum_{\substack{j=0 \\ j+k_{1,\theta_1}+k_{1,\theta_2}+k_2+k=m,\ k_{1,\theta_2} \geq 9(j+k_{1,\theta_1})}}^{m} \sum_{k_{1,\theta_1}=0}^{m} \sum_{k_{1,\theta_2}=0}^{m} \sum_{k_2=0}^{m} \sum_{k=0}^{m} \binom{m}{j, k_{1,\theta_1}, k_{1,\theta_2}, k_2, k} \left(\frac{A_{12}}{\pi}\right)^j \left(\frac{A_1}{2\pi}\right)^{k_{1,\theta_1}+k_{1,\theta_2}} \times \left(\frac{A_1}{\pi}\right)^{k_2} \left(\frac{\pi - 2A_1 - A_{12}}{\pi}\right)^k . \tag{3}$$

Theory

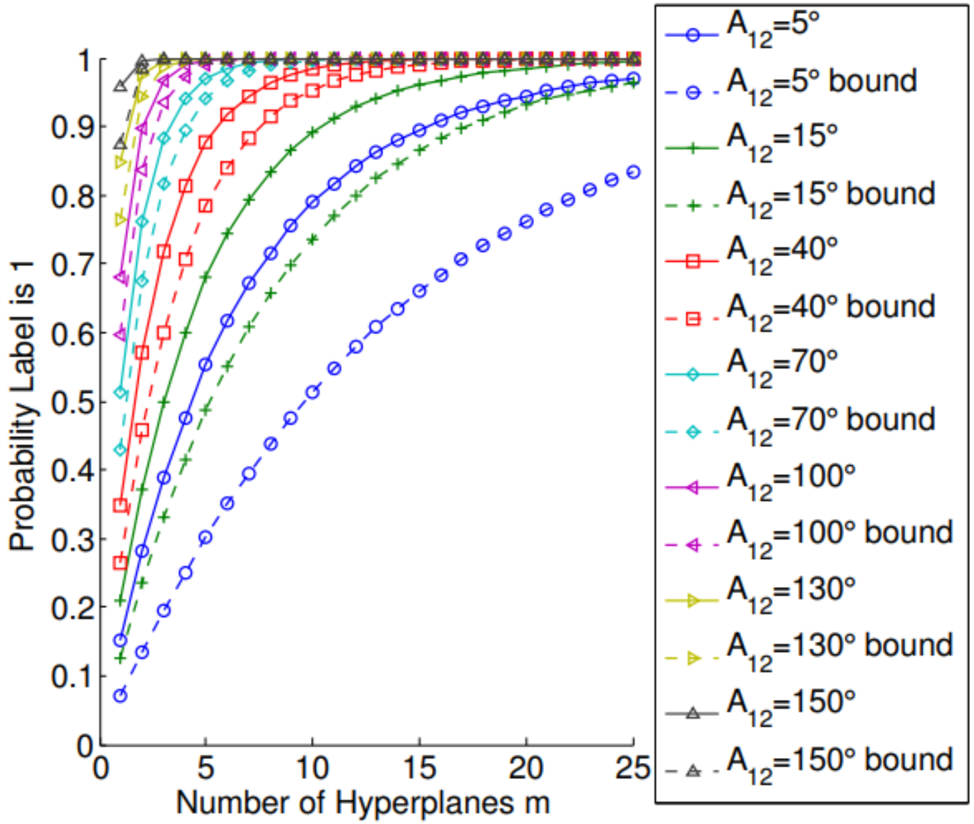


Theorem 1. *Let the classes G_1 and G_2 be two cones in $\mathbb{R}^2$ defined by angular measures A_1 and A_2, respectively, and suppose regions of the same angular measure have the same density of training points. Suppose $A_1 = A_2$, $\theta_1 = \theta_2$, and $A_{12} + A_1 + A_2 \leq \pi$. Then, the probability that a data point $x \in G_1$ gets classified in class G_1 by Algorithms 1 and 2 using a single layer and a measurement matrix $A \in \mathbb{R}^{m\times 2}$ with independent standard Gaussian entries is bounded as follows,*

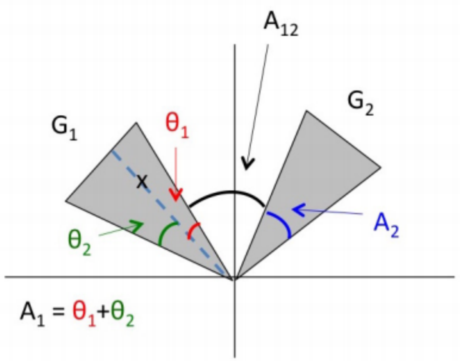
$$\mathbb{P}[\widehat{b}_x = 1] \geq 1 - \sum_{\substack{j=0 \\ j+k_{1,\theta_1}+k_{1,\theta_2}+k_2+k=m,\ k_{1,\theta_2}\geq 9(j+k_{1,\theta_1})}}^{m} \sum_{k_{1,\theta_1}=0}^{m} \sum_{k_{1,\theta_2}=0}^{m} \sum_{k_2=0}^{m} \sum_{k=0}^{m} \binom{m}{j, k_{1,\theta_1}, k_{1,\theta_2}, k_2, k} \left(\frac{A_{12}}{\pi}\right)^j \left(\frac{A_1}{2\pi}\right)^{k_{1,\theta_1}+k_{1,\theta_2}} \times \left(\frac{A_1}{\pi}\right)^{k_2} \left(\frac{\pi - 2A_1 - A_{12}}{\pi}\right)^k. \quad (3)$$

❖ As m → ∞, P(correct label) → 1

Theory



Probability Label is 1
Number of Hyperplanes m
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
5
10
15
20
25
A_{12}=5°
A_{12}=5° bound
A_{12}=15°
A_{12}=15° bound
A_{12}=40°
A_{12}=40° bound
A_{12}=70°
A_{12}=70° bound
A_{12}=100°
A_{12}=100° bound
A_{12}=130°
A_{12}=130° bound
A_{12}=150°
A_{12}=150° bound



A_{12}
G_1
G_2
θ_1
x
θ_2
A_2
$A_1 = \theta_1+\theta_2$

Take-away

- Simple classification from binary data
  - Efficient storage of the data
  - Efficient and simple algorithm
  - Theoretical analysis possible
  - Already competes with state of the art

- Future work
  - Dithers to allow for more complicated geometries?
  - Theoretical analysis of the discrete case?

- deanna@math.ucla.edu
- math.ucla.edu/~deanna
- "Simple Classification using Binary Data"
  – Needell, Saab, Woolf