What's hot in Machine Learning?

Anima Anandkumar

U.C. Irvine

What's hot in ML: Representation Learning

Label	Features			
0	2.1	5.2	0	0
1	0	0	2	1
1	1.1	0	0	0
0	0	0	7	0

Feature Engineering

- Learn good features/representations for classification tasks, e.g. image and speech recognition.
- Sparse representations, low dimensional hidden structures.

What's hot in ML: Optimization Methods

Convex Optimization

- Fast convergence for non-smooth (and not strongly convex) functions.
- Online learning: variance reduction for stochastic gradient methods.

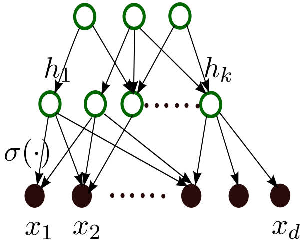
Non-convex problems

- When can we hope to reach global optimum?
- What problem structures make this possible?
- Can we have fast convergence?

Challenges in Feature Learning

In practice

- Deep learning has provided impressive gains.
- Parameter training challenging and not stable.

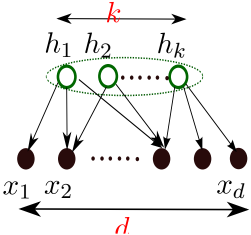


Theory

- Representational power of networks.
- Guaranteed learning of probabilistic models with latent variables?
- Maximum likelihood is NP-hard.
- Practice: EM, Variational Bayes have no consistency guarantees.
- Efficient computational and sample complexities?

Outline

Linear Neural Networks



- Observed sample $x = Ah$.
- h is hidden variable and A is dictionary.
- $x \in \mathbb{R}^d$, $h \in \mathbb{R}^k$ and $A \in \mathbb{R}^{d \times k}$.

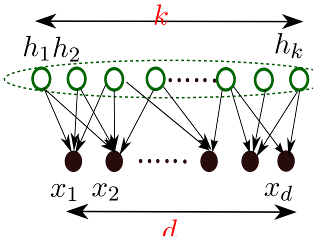
Learning through SVD

- Pairwise moments: $M_2 = \mathbb{E}[xx^\top] = A\mathbb{E}[hh^\top]A^\top$.
- SVD: $M_2 = U\Lambda U^\top$: a valid linear representation.

- Learning through SVD: cannot learn overcomplete representations. $(k > d)$ learnable?
- SVD cannot enforce sparsity, non-negativity etc.

Linear Neural Networks

- Observed sample $x = Ah$.
- h is hidden variable and A is dictionary.
- $x \in \mathbb{R}^d$, $h \in \mathbb{R}^k$ and $A \in \mathbb{R}^{d \times k}$.

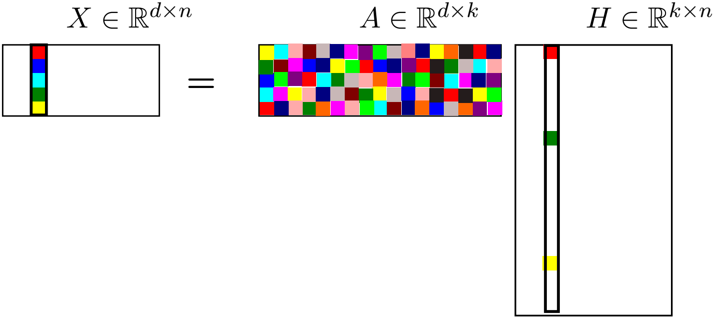


Learning through SVD

- Pairwise moments: $M_2 = \mathbb{E}[xx^\top] = A\mathbb{E}[hh^\top]A^\top$.
- SVD: $M_2 = U\Lambda U^\top$: a valid linear representation.

- Learning through SVD: cannot learn overcomplete representations. $(k > d)$ learnable?
- SVD cannot enforce sparsity, non-negativity etc.

Learning Overcomplete Dictionaries



- Linear model: $X = AH$, both A, H unknown.
- Sparse H: each column is randomly s-sparse
- Overcomplete dictionary $A \in \mathbb{R}^{d\times k}$: $k \geq d$.
- Incoherence: $\max_{i\neq j} |\langle a_i, a_j\rangle| \approx 0$. (satisfied by random vectors)

"Learning Sparsely Used Overcomplete Dictionaries" by A. Agarwal, A., P. Jain, P. Netrapalli, R. Tandon. COLT 2014.

Intuitions: how incoherence helps

- Each sample is a combination of dictionary atoms: $x_i = \sum_j h_{i,j} a_j$.
- Consider x_i and x_j s.t. they have no common dictionary atoms.
- What about $|\langle x_i, x_j \rangle|$?

Intuitions: how incoherence helps

- Each sample is a combination of dictionary atoms: $x_i = \sum_j h_{i,j} a_j$.
- Consider x_i and x_j s.t. they have no common dictionary atoms.
- What about $|\langle x_i, x_j \rangle|$?
- Under incoherence: $|\langle x_i, x_j \rangle| \approx 0$.

Intuitions: how incoherence helps

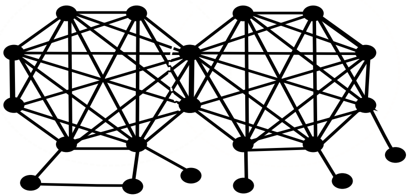
- Each sample is a combination of dictionary atoms: $x_i = \sum_j h_{i,j} a_j$.
- Consider x_i and x_j s.t. they have no common dictionary atoms.
- What about $|\langle x_i, x_j \rangle|$?
- Under incoherence: $|\langle x_i, x_j \rangle| \approx 0$.

Construction of Correlation Graph

- Nodes: Samples $x_1, \ldots, x_n$.
- Edges: $|\langle x_i, x_j \rangle| > \tau$ for some threshold τ.

How does the correlation graph help in dictionary learning?

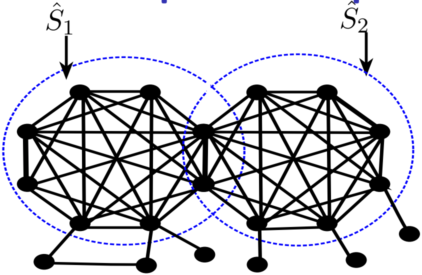
Correlation Graph and Clique Finding



Main Insight

- (x_i, x_j): edge in correlation graph $\Rightarrow$ x_i and x_j have at least one dictionary element in common.

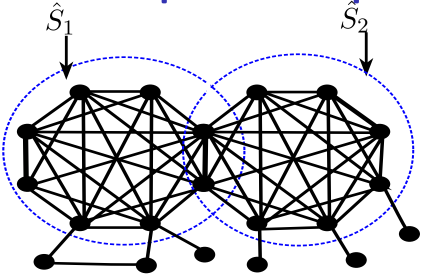
Correlation Graph and Clique Finding



Main Insight

- (x_i, x_j): edge in correlation graph $\Rightarrow$ x_i and x_j have at least one dictionary element in common.

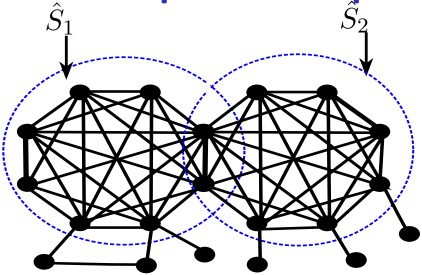
Correlation Graph and Clique Finding



Main Insight

- (x_i, x_j): edge in correlation graph $\Rightarrow x_i$ and x_j have at least one dictionary element in common.
- Consider a large clique: a large fraction of pairs have exactly one element in common.

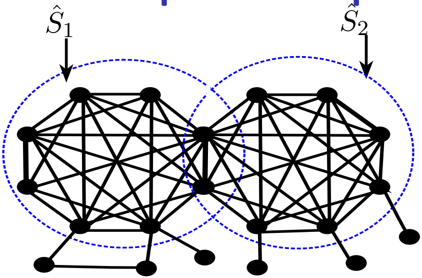
Correlation Graph and Clique Finding



Main Insight

- (x_i, x_j): edge in correlation graph $\Rightarrow$ x_i and x_j have at least one dictionary element in common.
- Consider a large clique: a large fraction of pairs have exactly one element in common.
- How to find such a large clique efficiently?

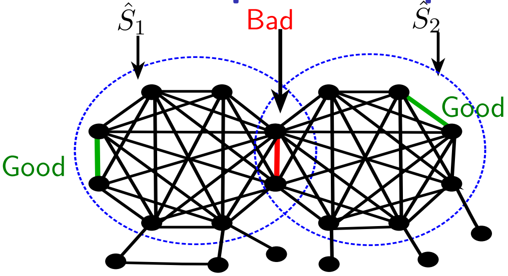
Correlation Graph and Clique Finding



Main Insight

- (x_i, x_j): edge in correlation graph $\Rightarrow$ x_i and x_j have at least one dictionary element in common.
- Consider a large clique: a large fraction of pairs have exactly one element in common.
- How to find such a large clique efficiently? Start with a random edge.

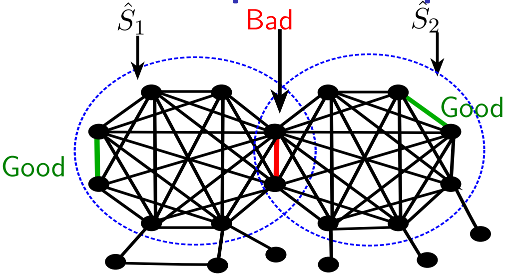
Correlation Graph and Clique Finding



Main Insight

- (x_i, x_j): edge in correlation graph $\Rightarrow$ x_i and x_j have at least one dictionary element in common.
- Consider a large clique: a large fraction of pairs have exactly one element in common.
- How to find such a large clique efficiently? Start with a random edge.

Correlation Graph and Clique Finding

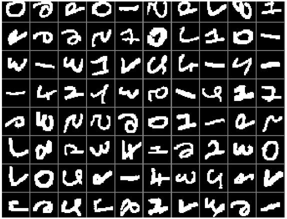


Main Insight

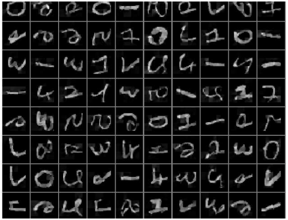
- (x_i, x_j): edge in correlation graph $\Rightarrow$ x_i and x_j have at least one dictionary element in common.
- Consider a large clique: a large fraction of pairs have exactly one element in common.
- How to find such a large clique efficiently? Start with a random edge.
- Refinement through alternating minimization.

Experiments on MNIST

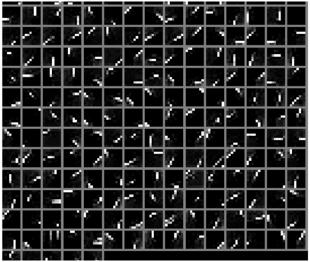
Original



Reconstruction

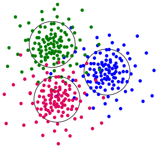


Learnt Representation



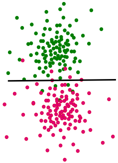
Outline

Warm-up: PCA on Gaussian Mixtures



- Mixture of Spherical Gaussians.
- PCA on pairwise moments: span of mean vectors.

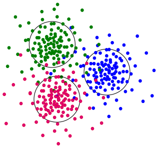
Learning Mean Vectors through Spectral Clustering



- Project samples on to span of mean vectors.
- Distance-based clustering (e.g. k-means).

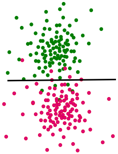
Warm-up: PCA on Gaussian Mixtures

- Mixture of Spherical Gaussians.
- PCA on pairwise moments: span of mean vectors.

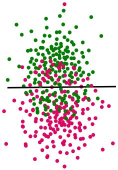


Learning Mean Vectors through Spectral Clustering

- Project samples on to span of mean vectors.
- Distance-based clustering (e.g. k-means).



Failure to cluster under large variance.



Learning Gaussian Mixtures Without Separation Constraints?

Beyond PCA: Spectral Methods on Tensors

- How to learn the component means (not just its span) without separation constraints?
- PCA is a spectral method on (covariance) matrices.
  - Are higher order moments helpful?

Beyond PCA: Spectral Methods on Tensors

- How to learn the component means (not just its span) without separation constraints?
- PCA is a spectral method on (covariance) matrices.
  - Are higher order moments helpful?
- What if number of components is greater than observed dimensionality $k > d$?

Beyond PCA: Spectral Methods on Tensors

- How to learn the component means (not just its span) without separation constraints?
- PCA is a spectral method on (covariance) matrices.
  - Are higher order moments helpful?
- What if number of components is greater than observed dimensionality $k > d$?
  - Do higher order moments help to learn overcomplete models?

Beyond PCA: Spectral Methods on Tensors

- How to learn the component means (not just its span) without separation constraints?
- PCA is a spectral method on (covariance) matrices.
  - Are higher order moments helpful?
- What if number of components is greater than observed dimensionality $k > d$?
  - Do higher order moments help to learn overcomplete models?
- What if the data is not Gaussian?

Beyond PCA: Spectral Methods on Tensors

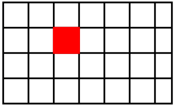
- How to learn the component means (not just its span) without separation constraints?
- PCA is a spectral method on (covariance) matrices.
  - Are higher order moments helpful?
- What if number of components is greater than observed dimensionality $k > d$?
  - Do higher order moments help to learn overcomplete models?
- What if the data is not Gaussian?
  - Moment-based Estimation of probabilistic latent variable models?

Tensor Notation for Higher Order Moments

- Multi-variate higher order moments form tensors.
- Are there spectral operations on tensors akin to PCA?

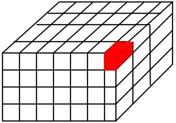
Matrix

- $\mathbb{E}[x \otimes x] \in \mathbb{R}^{d \times d}$ is a second order tensor.
- $\mathbb{E}[x \otimes x]_{i_1,i_2} = \mathbb{E}[x_{i_1} x_{i_2}]$.
- For matrices: $\mathbb{E}[x \otimes x] = \mathbb{E}[xx^\top]$.



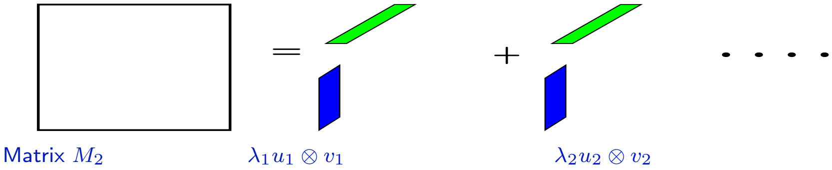
Tensor

- $\mathbb{E}[x \otimes x \otimes x] \in \mathbb{R}^{d \times d \times d}$ is a third order tensor.
- $\mathbb{E}[x \otimes x \otimes x]_{i_1,i_2,i_3} = \mathbb{E}[x_{i_1} x_{i_2} x_{i_3}]$.



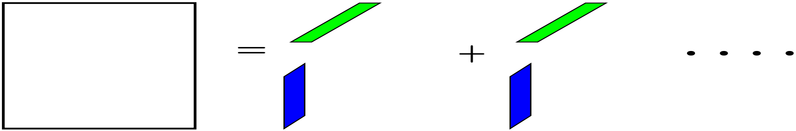
Matrices vs. Tensors

$$M_2 \approx \sum_i \lambda_i u_i \otimes v_i$$



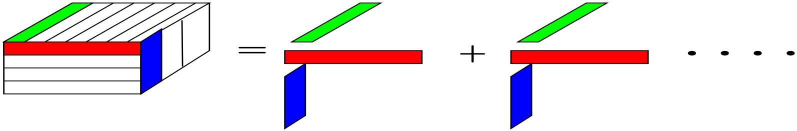
Matrices vs. Tensors

$$M_2 \approx \sum_i \lambda_i u_i \otimes v_i$$



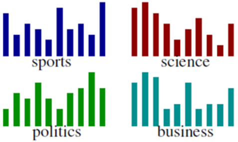
Matrix M_2 $\lambda_1 u_1 \otimes v_1$ $\lambda_2 u_2 \otimes v_2$

$$M_3 \approx \sum_i \lambda_i u_i \otimes v_i \otimes w_i$$



Tensor M_3 $\lambda_1 u_1 \otimes v_1 \otimes w_1$ $\lambda_2 u_2 \otimes v_2 \otimes w_2$

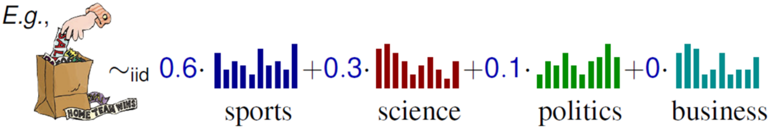
Topic Modeling



k topics (distributions over vocab words).

Each document $\leftrightarrow$ mixture of topics.

Words in document $\sim_{\text{iid}}$ mixture dist.



aardvark	0
athlete	3
	⋮
zygote	1

$\Pr_\theta[\text{“play”} \mid \text{sports}] = 0.0002$

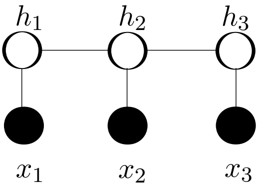
$\Pr_\theta[\text{“game”} \mid \text{sports}] = 0.0003$

$\Pr_\theta[\text{“season”} \mid \text{sports}] = 0.0001$

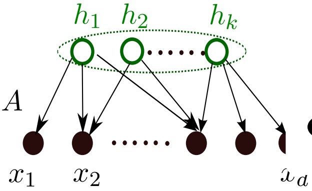
⋮

Tensor Factorizations for Other Models

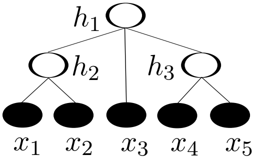
HMM



ICA



Latent Trees

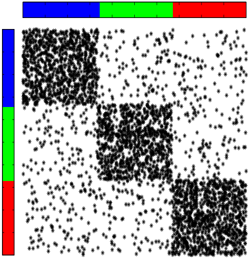


Method of Moments: Analyze moment tensors under statistical models.

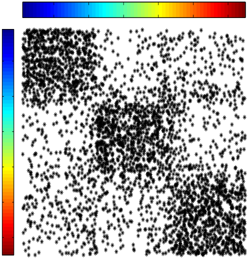
"Tensor Decompositions for Learning Latent Variable Models" by A. Anandkumar, R. Ge, D. Hsu, S.M. Kakade and M. Telgarsky. Preprint, October 2012.

Finding Hidden Communities in Networks

Pure Memberships

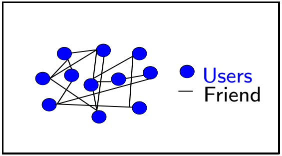


Mixed Memberships



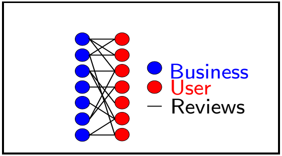
Tensor methods can find overlapping communities in networks

Experimental Results



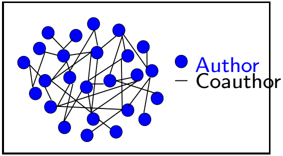
Facebook

$n \sim 20,000$



Yelp

$n \sim 40,000$



DBLP

$n \sim 1$ million

Error ($\mathcal{E}$) and Recovery ratio ($\mathcal{R}$)

Dataset	$\hat{k}$	Method	Running Time	$\mathcal{E}$	$\mathcal{R}$
Facebook(k=360)	500	ours	468	0.0175	100%
Facebook(k=360)	500	variational	86,808	0.0308	100%
Yelp(k=159)	100	ours	287	0.046	86%
Yelp(k=159)	100	variational	N.A.		
DBLP(k=6000)	100	ours	5407	0.105	95%

Huang, Niranjan, Hakeem and Anandkumar, "Fast Detection of Overlapping Communities via Online Tensor Methods," Preprint, Sept. 2013.

Experimental Results on Yelp

Lowest error business categories & largest weight businesses

Rank	Category	Business	Stars	Review Counts
1	Latin American	Salvadoreno Restaurant	4.0	36
2	Gluten Free	P.F. Chang's China Bistro	3.5	55
3	Hobby Shops	Make Meaning	4.5	14
4	Mass Media	KJZZ 91.5FM	4.0	13
5	Yoga	Sutra Midtown	4.5	31

Experimental Results on Yelp

Lowest error business categories & largest weight businesses

Rank	Category	Business	Stars	Review Counts
1	Latin American	Salvadoreno Restaurant	4.0	36
2	Gluten Free	P.F. Chang's China Bistro	3.5	55
3	Hobby Shops	Make Meaning	4.5	14
4	Mass Media	KJZZ 91.5FM	4.0	13
5	Yoga	Sutra Midtown	4.5	31

Bridgeness: Distance from vector $[1/\hat{k}, \ldots, 1/\hat{k}]^\top$

Top-5 bridging nodes (businesses)

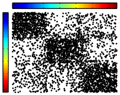
Business	Categories
Four Peaks Brewing	Restaurants, Bars, American, Nightlife, Food, Pubs, Tempe
Pizzeria Bianco	Restaurants, Pizza, Phoenix
FEZ	Restaurants, Bars, American, Nightlife, Mediterranean, Lounges, Phoenix
Matt's Big Breakfast	Restaurants, Phoenix, Breakfast& Brunch
Cornish Pasty Co	Restaurants, Bars, Nightlife, Pubs, Tempe

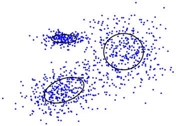
Outline

Conclusion

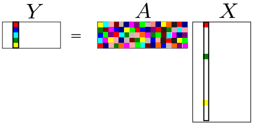
Guaranteed Learning of Latent Variable Models

- Guaranteed to recover correct model
- Efficient sample and computational complexities
- Better performance compared to EM, Variational Bayes etc.





- Tensor approach: mixed membership communities, topic models, latent trees...
- Sparsity-based approach: overcomplete models, e.g sparse coding and topic models.



---

http://newport.eecs.uci.edu/anandkumar/MLSS.html