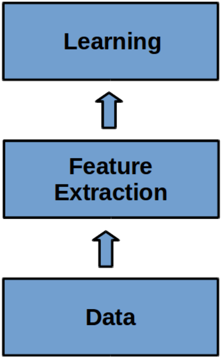
Learning from Complex and Heterogeneous Data

Piyush Rai

IIT Kanpur

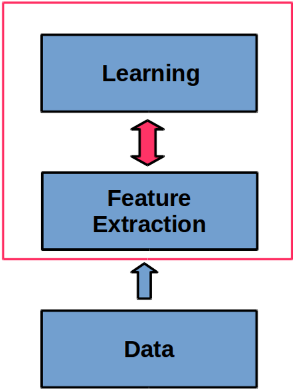
December 3, 2015

Learning from Data: The Traditional Way



A two-stage process. Stage 1 often hand-crafted.

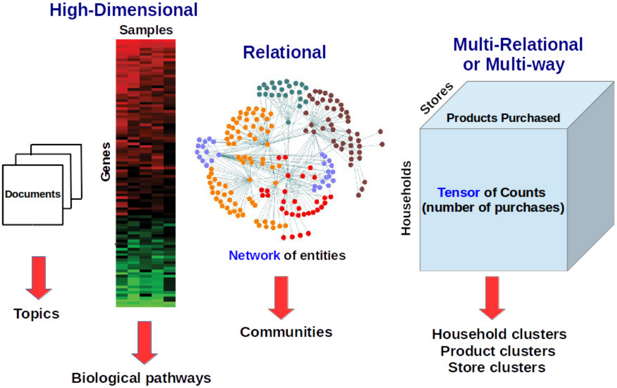
Learning via "Feature Learning"



Learning features tuned for specific tasks. Lot of recent buzz (e.g., deep learning).

Feature Learning for Complex Data

Flexible and scalable probabilistic models for learning feature representations and latent structures to unravel and understand massive and complex data

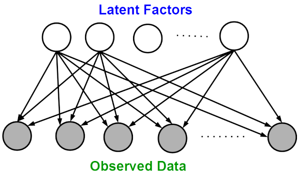


Features and structures that are expressive and interpretable with good predictive power; learned using models that adapt in size as data warrants.

Feature Learning via Latent Factor Models

Infer **latent factors** that compactly represent and explain data

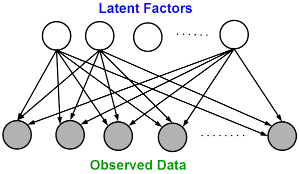
Also commonly known as Factor Analysis



Feature Learning via Latent Factor Models

Infer latent factors that compactly represent and explain data

Also commonly known as Factor Analysis



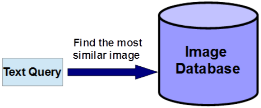
Some examples:

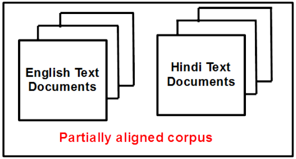
- Observed: Gene-expression data, latent: biological pathways
- Observed: Words in documents, latent: topics
- Observed: Images, latent: basic images (or "dictionary")
- Observed: Edges in a network, latent: community of each node
- Observed: Execution traces of programs, latent: bugs in the source codes

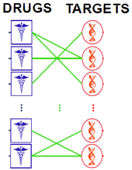
Can be made "deep" by stacking multiple layers of FAs

Learning from Multi-modal Data

- How to (efficiently) compare/align objects across modalities?
    - Text queries vs images
    - English documents vs Hindi documents
    - Drugs vs Targets





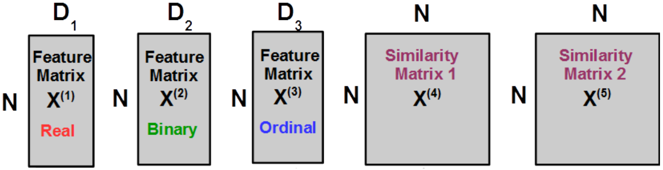


Partially known associations

Learning from Multi-modal Data

- How to reconcile heterogeneity and integrate data from multiple modalities?

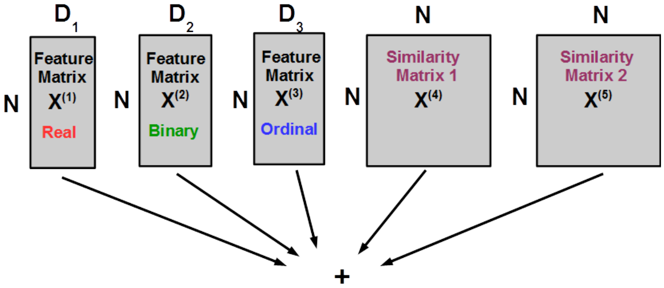
Given: N objects in multiple feature-based
and/or similarity-based "views"



Learning from Multi-modal Data

- How to reconcile heterogeneity and integrate data from multiple modalities?

Given: N objects in multiple feature-based
and/or similarity-based "views"



D_1 N Feature Matrix $X^{(1)}$ Real

D_2 N Feature Matrix $X^{(2)}$ Binary

D_3 N Feature Matrix $X^{(3)}$ Ordinal

N N Similarity Matrix 1 $X^{(4)}$

N N Similarity Matrix 2 $X^{(5)}$

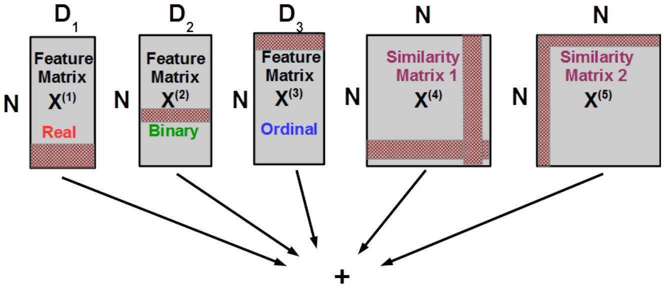
+

How to properly combine such heterogeneous data?

Learning from Multi-modal Data

- How to reconcile heterogeneity and integrate data from multiple modalities?

Given: N objects in multiple feature-based
and/or similarity-based “views”

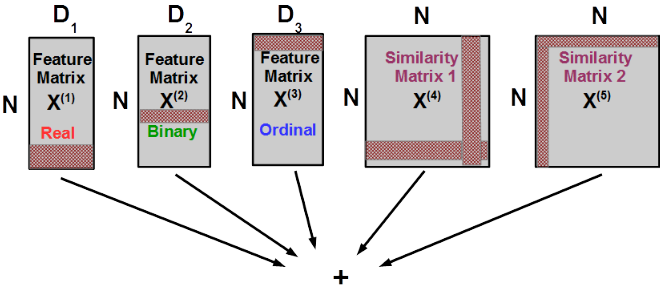


How to properly combine such heterogeneous data
(especially when each view also has missing data)?

Learning from Multi-modal Data

- How to reconcile heterogeneity and integrate data from multiple modalities?

Given: N objects in multiple feature-based
and/or similarity-based "views"



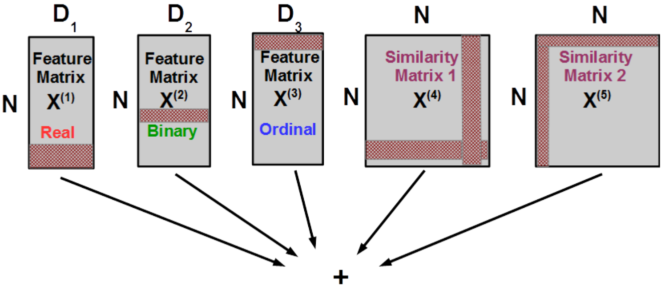
How to properly combine such heterogeneous data
(especially when each view also has missing data)?

Goal: factor analysis, matrix completion, classification,
or clustering with incomplete heterogeneous data

Learning from Multi-modal Data

- How to reconcile heterogeneity and integrate data from multiple modalities?

Given: N objects in multiple feature-based and/or similarity-based "views"

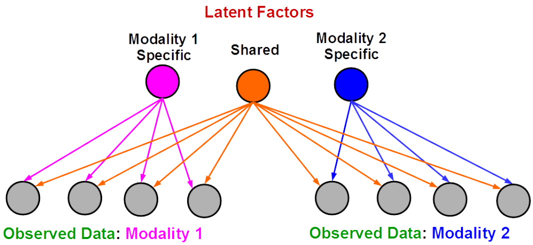


How to properly combine such heterogeneous data (especially when each view also has missing data)?

Generalizes multiview learning and multiple kernel learning

Multi-modal Latent Factor Models

Extract latent factors to compactly represent and explain **multi-modal data**

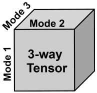


Some examples:

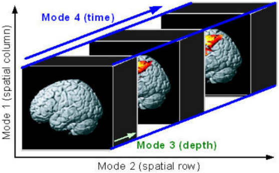
- Data: Webpages. Modality 1: Images. Modality 2: Text
- Data: Video news clips. Modality 1: Audio. Modality 2: Video
- Data: Medical images. Modality 1: fMRI Data. Modality 2: EEG Data

Learning from Multi-way/Multi-Relational Data

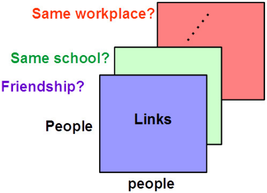
- **Multi-way arrays** with three or more "modes" / "ways"



- Found in many applications: medical imaging, computer vision, modeling knowledge bases, multi-aspect recommender systems, etc.



4D tensor (Brain imaging)

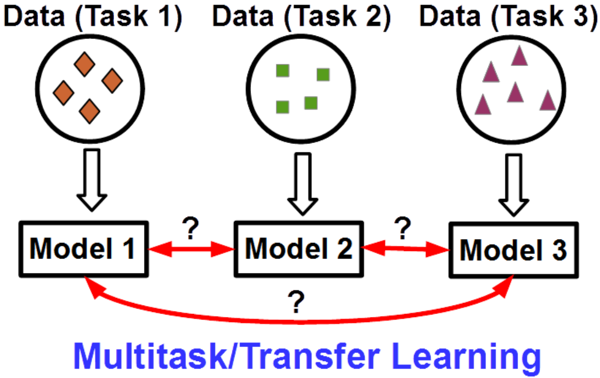


3D tensor (multi-relational data)

Key focus: Scalable probabilistic tensor factorizations for real-, binary-, and count-valued tensors. Integrating sources of side-information.

Learning Multiple (Related) Tasks

Learning the relatedness structure among multiple learning tasks to share data/information across tasks



Key focus: Avoiding negative transfer and extending this to "life-long learning".

Thanks! Questions?