Unsupervised Learning from Dyadic Data

Thomas Hofmann and Jan Puzicha

Presented by: Ajay Joshi

September 18, 2007

Outline

- Problem description
- Contributions in a nutshell
- Models for dyadic data
- Parameter learning with EM
- Relationships between different models
- Results in one domain
- Concluding remarks

Dyadic Data

- Two finite sets of objects X and Y.
- Observations made for pairs consisting of one element from each set: dyads (x, y).
- In some cases, a third observation $w(x, y)$ might be available indicating similarity or association values between x and y.
- This paper restricts itself to purely dyadic observations: (x, y).

Problem Description

- The objective is to accomplish two tasks:
  - Learning a joint or conditional probability distribution over $X \times Y$.
  - Discovering structure: Learning clusters or data hierarchies.
- Looks very similar to the usual learning setting.
- Why is this a hard problem ?
  - **Metric distances are not known !** In the standard setting, features are represented as vectors in a Euclidean space.
  - **Data can be extremely sparse**, zero frequencies are common.

To summarize the problems

- Consider two sets $X = \{x_1, x_2, x_3\}$ and $Y = \{y_1, y_2, y_3, y_4\}$. Suppose our observations are $(x_1, y_2), (x_1, y_3), (x_2, y_2), (x_3, y_1), (x_1, y_2)$. Note that we can have $|X| * |Y|$ different dyads. Hence many pairs usually have zero frequencies in observations.

- How to handle sparse data ? What do we do when some pairs are never sampled (the zero-frequency problem)?

- How to handle the lack of a distance metric ?

- Sparseness problem also occurs in the standard setting, however, **knowing a distance metric can help us generalize to unseen examples**.

Applications

- Computer Vision: In image segmentation, data is usually obtained as pixel values (or processed using some filters) for each image location. The task is to group similar regions together.

- Information retrieval: One set is the collection of documents, the other represents vocabulary. A dyad is the occurrence of an object from the vocabulary in the corresponding document.

- Consumer preference analysis: One set represents consumers and the other is objects they can consume.

- Note that in some of these examples, there does not exist an obvious "distance" metric between two objects of a set.

Basic modeling principles

- Modeling is done using latent variables by specifying a joint probability distribution for latent and observable variables.

- Marginalization (summing over latent variables) gives a model over observable variables.

- Bayes' rule is used to obtain posterior probabilities over latent variables with respect to observed variables - used for structure discovery. Useful where data groups and/or hierarchies are to be found.

Summary of Contributions

- The paper proposes a family of latent class models to deal with the data sparseness problem.

- Flat as well as hierarchical models are described and evaluated. A close relationship between aspect models and clustering models is shown.

- Previously studied models like n-gram models, distributional clustering, aggregate Markov models arise as special cases of this framework.

- The authors study EM algorithms for application to these models. Some promising results in the previously mentioned application domains are provided.

Aspect models

- Consider an observation sequence $S = (x^n, y^n)_{1 \leq n \leq N}$ as a realization of an underlying sequence of random variables $(X^n, Y^n)_{1 \leq n \leq N}$.

- We can introduce a latent class for each dyad observed. (x^n, y^n) is associated with A^n over a finite set $A = \{a_1, a_2, \ldots, a_K\}$.

- Notice that aspect models partition the *observations*. Therefore, identical dyads can be associated with *different* latent classes.

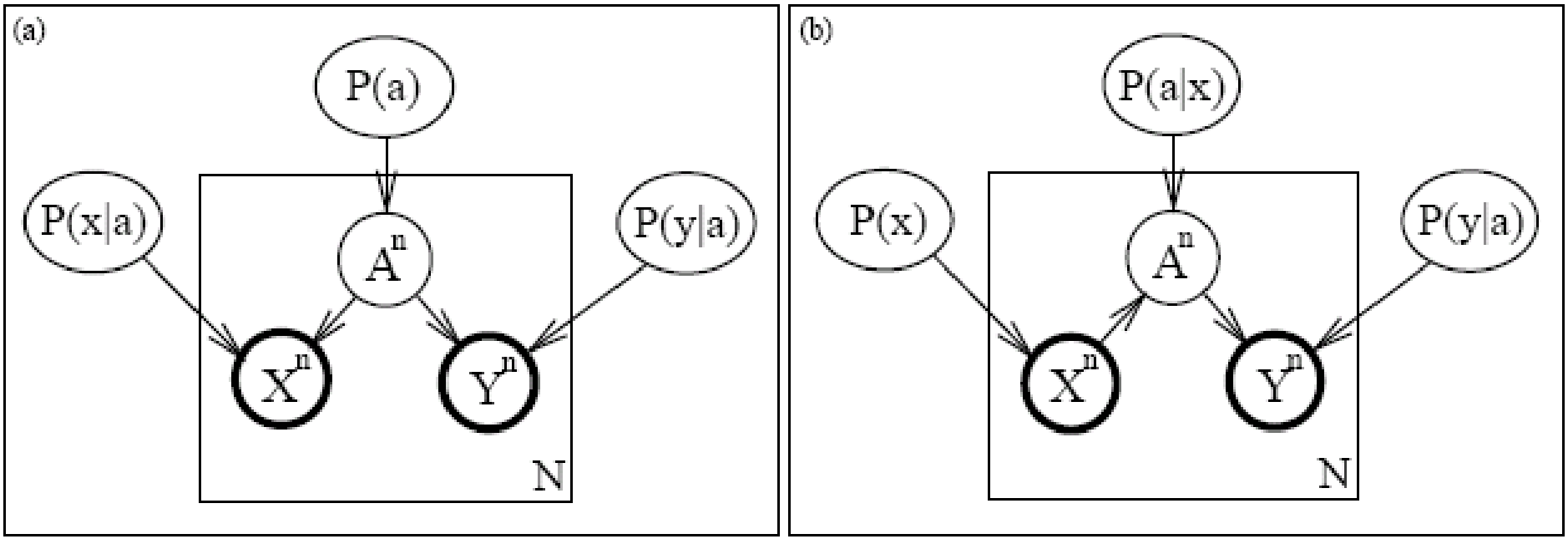
- The set of observations can be thought of as generated by a (latent) finite mixture model. The choice of the latent class from A corresponds to choosing the component of a mixture. In this view, the observed variables (x^n, y^n) are those generated by this component a^n of the mixture.

Model description

- Assumption: The observations are i.i.d. and pairs of random variables X^n and Y^n are conditionally independent given the latent class A^n.

- Data generation process:
  - Choose an aspect a with $P(a)$.
  - Choose an object x from the set X with $P(x|a)$.
  - Choose an object y from the set Y with $P(y|a)$.

Aspect model



Graphical representation of an aspect model.
(a) In the symmetric parameterization
(b) In the asymmetric parameterization.

Model description

- As described earlier, the complete data probability is first obtained and then marginalized.

$$P(S, a) = \prod_{n=1}^{N} P(x^n, y^n, a^n), \text{ where}$$

$$P(x^n, y^n, a^n) = P(a^n)P(x^n|a^n)P(y^n|a^n).$$

- Decomposition into products follows from the i.i.d. assumption, simplification of chain rule follows from the conditional independence assumption.

Model description

- Marginalizing w.r.t a gives $P(x,y) = \sum_{a \in A} P(a)P(x|a)P(y|a)$.

- Grouping identical dyads together, we get

$$P(S) = \prod_{x \in X} \prod_{y \in Y} P(x,y)^{n(x,y)}$$

.

- $n(x,y)$ represents the empirical co-occurrence frequencies. This number is expected to be zero for most (x,y) pairs as we have sparse data.

Expectation Maximization

- Parameter estimation is to be done using Maximum Likelihood. However, there exist problems since we have a log of a summation. The EM algorithm is therefore applied.

- Expectation step: Estimating the posterior probabilities of the unobserved mapping $P(a|S,\theta')$. θ' is the current parameter estimate.

- Maximization step: Maximization of the expected complete data log-likelihood $L(\theta|\theta') = \sum_a P(a|S,\theta')\log P(S,a,\theta)$ with respect to θ.

An equivalent asymmetric parameterization

- From the asymmetric graphical model shown earlier,

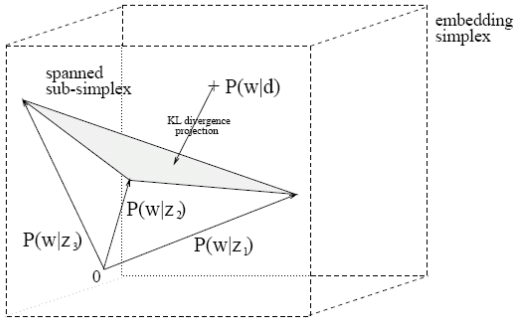
$$P(x, y) = P(x)P(y|x) = P(x) \sum_{a \in A} P(a|x)P(y|a). \tag{1}$$

- For a fixed x, all conditional distributions $P(y|x)$ are obtained by convex combinations of $P(y|a)$.

- Also, $P(x)$ can be estimated independently as $n(x)/N$ (the occurrence frequency). Thus, maximizing the joint likelihood $P(x, y)$ and the conditional likelihood $P(y|x)$ are equivalent.

Word generation example from documents

- Pick a document d with probability $P(d)$.
- Pick a latent class z with probability $P(z|d)$.
- Generate a word w with probability $P(w|z)$.
- From before, we have $P(w|d) = \sum_{z \in Z} P(w|z)P(z|d)$.
- Document specific word distributions are obtained by a convex combination of the aspects $P(w|z)$.
- Documents are not assigned to clusters, they are characterized by a specific mixture of factors with weights $P(z|d)$. These mixing weights offer modeling flexibility.



- Assume a vocabulary of size M. Then, we have $M-1$ dimensional vectors $P(\cdot|z)$ over the vocabulary. There are K such points assuming K latent classes.

- Because of the convex combination of described earlier, $P(\cdot|d)$ can lie in a $K-1$ dimensional sub-simplex in the $M-1$ dimensional simplex as shown in the figure. The mixing weights $P(z|d)$ correspond to the coordinates of a document in that sub-simplex.

- Dimensionality of the sub-simplex is $K-1$ as opposed to $M-1$ for the complete probability simplex. This can be seen as dimensionality reduction.

Cross Entropy Minimization

- Say the empirical co-occurrence frequencies are $\hat{P}(x,y) = n(x,y)/N$.

- The complete data likelihood is

$$\prod_{n=1}^{N} P(x^n, y^n, a^n) \;=\; \prod_{n=1}^{N} P(x^n) \sum_{a \in A} P(a^n|x^n) P(y^n|a^n)$$

  . The log-likelihood is

$$\begin{aligned} L(S,\theta) &= \log[\prod_{x \in X} \prod_{y \in Y} \left(P(x) \sum_{a \in A} P(a|x) P(y|a) \right)^{n(x,y)}] \\ &= \sum_{x,y} n(x,y) \cdot \log[P(x) \sum_{a} P(a|x) P(y|a)] \end{aligned}$$

  . Hence

$$\frac{1}{N} L(S,\theta) \;=\; \sum_{x,y} \hat{P}(x,y) \log[P(x) \sum_{a} P(a|x) P(y|a)]$$

  .

Cross Entropy Minimization

- Separating out terms gives

$$\frac{1}{N}L(S,\theta) = \sum_x \hat{P}(x)\left[\log P(x) + \sum_y \hat{P}(y|x)\log\sum_a P(a|x)P(y|a)\right].$$

- The estimation of $P(x)$ can be done independently.

- Therefore, maximum likelihood estimation of the above is equivalent to minimizing the sum over cross entropies between empirical conditional distributions and model distributions weighted by occurrence frequencies.

- This is the same as minimizing the Kullback-Leibler divergence between the two distributions.

One-Sided Clustering Model

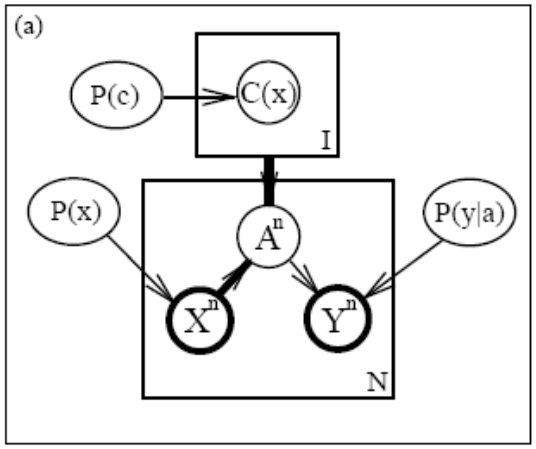
- In this model, latent classes are introduced for *objects* in one of the spaces (X or Y).

- Consider latent variables $C(x)$ over $C = \{c_1, c_2, \ldots, c_K\}$. A realization of these latent variables partitions the space X.

- In the aspect model, identical *observations* can have different latent classes. In contrast, in this clustering model, we can have multiple *objects* belonging to the same latent class. Notice that the partition is on *objects* and not *observations*.

- The authors show that the clustering model can be derived as a constrained aspect model.

Constraints on the aspect model



A one-sided clustering model

- Introduce additional latent clustering variables where latent variable states for clusters and aspects are identified, $c_k \cong a_k$. Consistency constraints on the aspect variables are

$$P(a|x,c) \equiv P\{A^n = a|X^n = x, C(x) = c\} = \delta_{ac}.$$

- $P(a|x,c)$ are not free parameters because of the above constraints. They are therefore not shown in the above graphical model.

Complete data likelihood

- The complete data likelihood for the one-sided model is

$$P(S,c) = \prod_{x \in X} P(C(x)) \prod_{y \in Y} [P(x)P(y|c(x))]^{n(x,y)}.$$

- Marginalizing with respect to the clustering variables gives

$$P(S) = \prod_{x \in X} P(S_x) \quad \text{where,}$$
$$P(S_x) = \sum_{c \in C} P(c) \prod_{y \in Y} [P(x)P(y|c)]^{n(x,y)}.$$

- Co-occurrences in S_x are not independent for given parameters, but are coupled by the latent variable $C(x)$.

One-sided model corresponds to Naive Bayes'

- The E-step in the EM algorithm for update of posterior probabilities is

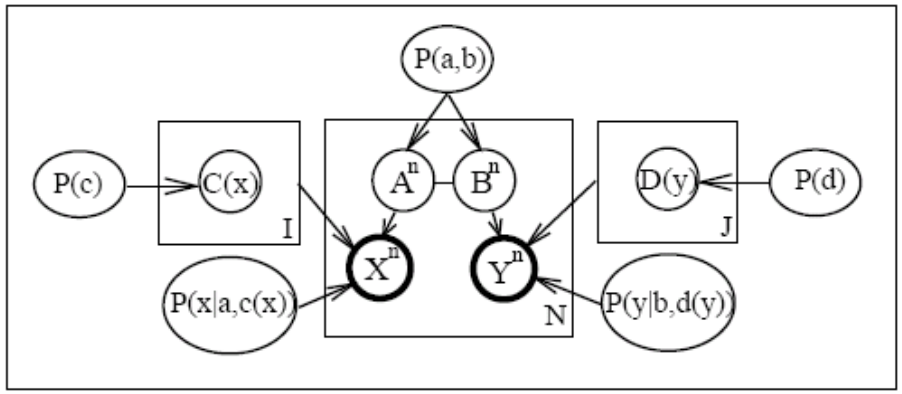
$$P\{C(x) = c|S_x, \theta\} = \frac{P(c)\prod_{y\in Y} P(y|c)^{n(x,y)}}{\sum_{c'} P(c')\prod_{y\in Y} P(y|c')^{n(x,y)}}.$$

- Doing away with the normalization term, this can be written as

$$P\{C(x) = c|S_x, \theta\} \propto P(c)\exp\left[-n(x)\left(-\sum_{y\in Y}\widehat{P}(y|x)\log P(y|c)\right)\right].$$

- Compare the above with a Gaussian Mixture model. If we interpret Y as a feature space for $x \in X$, then the one-sided model can be viewed as an unsupervised version of Naive Bayes' classifier.

Two-sided clustering model



A two-sided clustering model

- This model is defined by

  $P(x, y|c, d) \equiv P\{X^n = x, Y^n = y|C(x) = c, D(x) = d\} = P(x)P(y)\phi(c, d).$

- Prior probabilities are $P(c, d) = [\prod_x P(c(x))] \cdot [\prod_y P(d(y))]$.

- The two-sided clustering model can also be interpreted as a constrained aspect model as earlier.

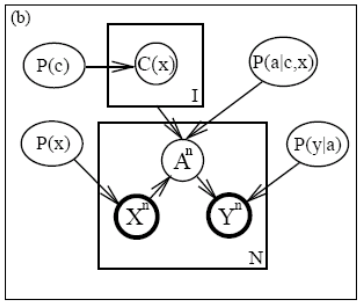
- Imposing the constraints, we (rather, the authors !) get

$$P(x, y|C(x) = c, D(y) = d) = P(x)P(y) \sum_{a,b} \delta_{ac}\delta_{bd} \frac{P(a, b)}{P(a)P(b)}.$$

- Comparing the two, we see that the cluster association parameters $\phi(c, d)$ correspond to the ratio of aspect probabilities and the product of marginal probabilities.

- $\phi(c, d)$ can also be seen as cluster association strengths. They control the probability of observing the dyad (x, y) relative to the model with an unconditional independence assumption.

Hierarchical clustering model



A hierarchical clustering model

- Hierarchical models here are defined combining aspects and clusters.

- Aspects are identified with nodes of a hierarchy, clusters are identified with terminal nodes only.

- Compatibility constraints

$$P(a|x,c) \equiv P\{A^n = a | X^n = x, C(x) = c\} = 0,$$
if a is not on the path from root to c .

One-sided model vs hierarchical model

- The one-sided model is a degenerate hierarchical with only terminal nodes.

- The structure for a hierarchical model is less constrained as compared to the one-sided model. The clustering structure does not completely define the aspect variables. Thus, there are more parameters in the hierarchical model.

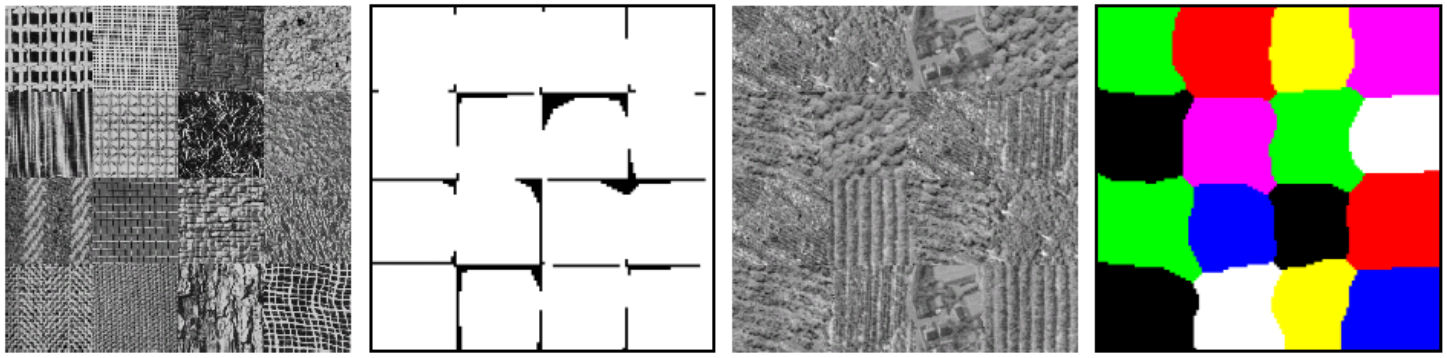
Putting it all together

Model	$P(y\|x,S)$
Aspect	$\sum_a P(a\|x)P(y\|a)$
One-sided clustering	$\sum_c P\{C(x)=c\|S\}P(y\|c)$
Hierarchical clustering	$\sum_c P\{C(x)=c\|S\}\sum_a P(a\|x,c)P(y\|a)$
Two-sided clustering	$\sum_c P\{C(x)=c\|S\}P(y)\sum_d P\{D(y)=d\|S\}\phi(c,d)$

- Aspect model is the most general (least constrained), two-sided model is the most constrained.

- $P(\cdot)$ are parameters different from the posterior probabilities $P\{\cdot\}$. $P\{C(x)=c|S\}$ asymptotically approach Boolean values while this does not hold for $P(a|x)$. Thus, the parameters in the aspect model are free and it is less constrained that other models.

Results (1/2)

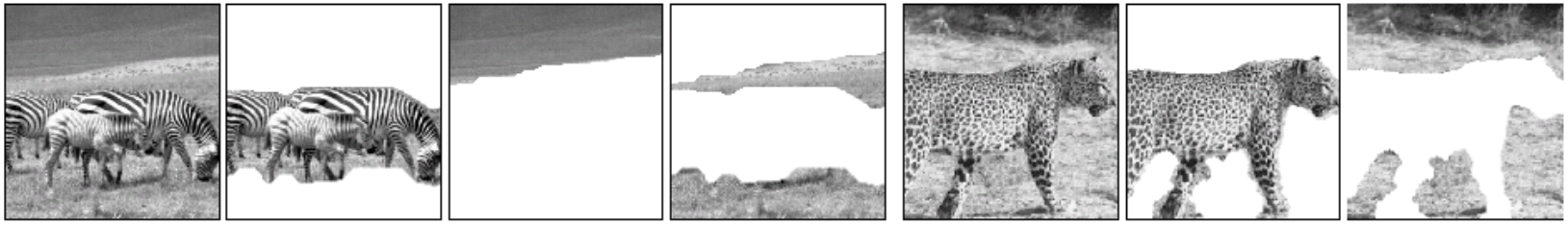


Texture segmentation results

- The segmentation is done using the one-sided clustering model described in the paper. Looks fairly good.

- The paper does not mention how the number of latent clusters are chosen. If this is input by a human, then one of the hardest problems in image segmentation is circumvented. Given the number of clusters, according to my experience, other approaches will be able to do equally well.

Results(2/2)



Segmentation results

- Segmentation using the one-sided clustering model for an outdoor scene. Again, the results are good, however, if the number of clusters are known, then it is not as useful.

Concluding Remarks

- The paper proposes statistically sound models for modeling dyadic data.
- The approach can be used for prediction or discovering latent structure in the data.
- Since the model is completely probabilistic, many powerful methods can be directly used for parameter estimation.
- It is not very clear (to me) how these models actually benefit predictive performance, or structure discovery as compared to other models.
- Specifically for the zero-frequency problem, I cannot see why models described in this paper excel as compared to others. Inputs ?

Questions?? Comments ??

Thank You!